

UNIVERSIDAD  
**NACIONAL**  
DE COLOMBIA

# Ética de IA

**Andrés Felipe Ramírez Martínez**

Universidad Nacional de Colombia  
Facultad de Ciencias Humanas, Departamento de Filosofía  
Bogotá, Colombia  
2023



# Ética de IA

**Andrés Felipe Ramírez Martínez**

Tesis o trabajo de investigación presentada(o) como requisito parcial para optar al título de:

**Magister en Filosofía**

Director (a):

Doctor en Filosofía Gonzalo Serrano

Línea de Investigación:

Ética de Inteligencia Artificial

Universidad Nacional de Colombia

Facultad de Ciencias Humanas, Departamento de Filosofía

Bogotá, Colombia

2023



*A mis padres*

# Declaración de obra original

Yo declaro lo siguiente:

He leído el Acuerdo 035 de 2003 del Consejo Académico de la Universidad Nacional. «Reglamento sobre propiedad intelectual» y la Normatividad Nacional relacionada al respeto de los derechos de autor. Esta disertación representa mi trabajo original, excepto donde he reconocido las ideas, las palabras, o materiales de otros autores.

Cuando se han presentado ideas o palabras de otros autores en esta disertación, he realizado su respectivo reconocimiento aplicando correctamente los esquemas de citas y referencias bibliográficas en el estilo requerido.

He obtenido el permiso del autor o editor para incluir cualquier material con derechos de autor (por ejemplo, tablas, figuras, instrumentos de encuesta o grandes porciones de texto).

Por último, he sometido esta disertación a la herramienta de integridad académica, definida por la universidad.

Andrés Felipe Ramírez Martínez

Fecha 03/08/2023



# Resumen

## Ética de IA

La inteligencia artificial ha empezado a moldear nuestra sociedad. Tenemos el objetivo como filósofos de pensar los nuevos desafíos que pueda traer en el campo epistemológico, ontológico y ético.

**Palabras clave: Inteligencia Artificial, Ética, Ética de IA**

# Abstract

## **AI Ethics**

Artificial intelligence has begun to shape our society. We have the objective as philosophers of thinking about the new challenges that it may bring in the epistemological, ontological and ethical fields.

**Keywords:** Artificial intelligence, Ethics, Ai Ethics

Resumen .....	8
Introducción.....	12
1. Contexto IA .....	13
1.1 Origen.....	13
1.2 Desarrollo.....	15
1.3 Singularidad .....	21
2. Motivaciones Éticas .....	22
2.1 Accidentes.....	23
2.2 Maleficencia.....	23
2.3 Desigualdad.....	24
2.4 Sustitución.....	25
2.5 ¿Tenemos Motivos?.....	26
3. Ética de Inteligencia Artificial.....	28
3.1 Ética de Máquinas estrechas .....	29
3.1.1 Algunos Dilemas.....	29
3.1.2 Principios.....	30
3.1.3 Construcción de Maquinas Inteligentes .....	32
3.1.4 Contexto Investigativo .....	34
3.2 Ética de Máquinas Generales.....	44
4. Algunos Dilemas de la Actualidad.....	47
4.1 Chat GPT.....	47
4.2 Armamento Militar .....	62
4.3 MegaSyn.....	67
5. Implicaciones Filosóficas.....	69
5.1 Motivaciones.....	69
5.1.1 Accidentes.....	69
5.1.2 Maleficencia.....	70
5.1.3 Desigualdad.....	71
5.1.4 Sesgos.....	72
5.1.5 Sustitución.....	73
5.2 Implicaciones .....	74
5.2.1 Justicia.....	74
5.2.2 Beneficencia.....	76
5.2.3 Privacidad.....	79
5.2.4 Transparencia.....	80
5.2.5 Responsabilidad .....	81

5.2.6 Leyes .....	83
5.2.7 Implicaciones filosóficas .....	84
Bibliografía .....	94
Bibliografía Secundaria .....	99

# Introducción

Desde la década de los 50 hemos sido testigos de los esfuerzos de investigadores y científicos por crear máquinas capaces de exhibir inteligencia similar a la humana. Estas tecnologías han revolucionado gradualmente la forma en que trabajamos, nos comunicamos y vivimos. Resulta difícil determinar una fecha exacta para su inicio importantes avances han contribuido a esta revolución. A mediados de los 2000 las computadoras adquirieron más poder de procesamiento, lo que, junto con grandes conjuntos de datos, inició el aprendizaje automático y el aprendizaje profundo, sentando así las bases para implementar la IA (Inteligencia Artificial) en diversas industrias. En las últimas décadas hemos presenciado la expansión de estas tecnologías mediante la aparición de asistentes virtuales como Siri, Alexa y Google Assistant que han cambiado nuestra forma de comunicarnos y acceder a la información. Los chats y los sistemas de automatización de tareas han mejorado la eficiencia y productividad en diversos campos laborales. Actualmente, la IA está presente en aplicaciones cotidianas que nos afectan de distintas maneras y estas aplicaciones de la inteligencia moldean cómo vivimos y experimentamos el mundo.

En el ámbito de la medicina, la inteligencia artificial ha superado nuestras capacidades en el diagnóstico y tratamiento de enfermedades, permitiendo a los médicos obtener diagnósticos más precisos y diseñar terapias personalizadas. En la industria automotriz, los vehículos autónomos que promueven la seguridad vial y están transformando gradualmente la movilidad, ya circulan por las calles de algunas ciudades del primer mundo. En el sector financiero, la inteligencia artificial se utiliza para el análisis de datos y la detección de fraudes mejorando la efectividad y seguridad de las transacciones en general. Sin embargo, a medida que la inteligencia artificial avanza y las máquinas inteligentes se vuelven más sofisticadas surgen diversas tensiones éticas en torno a su uso y desarrollo. Las implicaciones éticas de las máquinas, beneficiosas y perjudiciales, se convierten en una cuestión importante para nuestra sociedad. Asimismo, el impacto socioeconómico de la inteligencia artificial adquiere relevancia, ya que está transformando el panorama laboral y contribuyendo al desempleo y la desigualdad. Por último, la privacidad y la protección de datos personales son temas centrales en el desarrollo de la inteligencia artificial. Su capacidad de recopilar, analizar y utilizar grandes cantidades de datos plantea interrogantes sobre el control, la seguridad y la confidencialidad de la información.

Examinar y comprender estas tensiones éticas que están surgiendo en diversos campos es crucial para garantizar el bienestar y la seguridad de la sociedad. Es por esto que se han propuesto y explorado diversos principios morales que pueden servir como directrices éticas y normativas, asegurando así el uso responsable de la inteligencia artificial y priorizando los valores humanos. Los principios éticos más citados en este reciente campo de investigación son: la transparencia, la justicia, la beneficencia, la privacidad y la responsabilidad. En el desarrollo de esta tesis exploraré en mayor profundidad estos principios morales y su relación con las motivaciones éticas para considerar la ética de las máquinas inteligentes. También será relevante analizar cómo se han aplicado hasta ahora para evaluar si estamos protegidos o no.

A lo largo de mi investigación he recurrido a modelos como ChatGPT para mejorar y corregir mi tesis, siempre dentro de los límites éticos y legales establecidos en cuanto al plagio y los derechos de autor. El uso de estos modelos de lenguaje generativos ha sido muy útil para revisar y perfeccionar el contenido de mi trabajo, y ha sentado un precedente sobre la importancia de este tema de investigación al mostrar los dilemas que estas tecnologías

generan en contextos como el académico o el laboral. He tomado las precauciones necesarias para asegurar que mis aportes sean originales y contribuyan al conocimiento existente en el campo de la ética de las máquinas inteligentes. Me he basado en fuentes académicas confiables evaluando críticamente las ideas expuestas para respaldar mis puntos de vista y conclusiones.

Adicionalmente, analizaré el caso del armamento militar, el nuevo protagonista en los conflictos actuales que está permitiendo un desequilibrio en el poder y una resignificación del conflicto. Debemos reflexionar sobre el panorama oscuro plagado de prejuicios que esto podría generar. Analizaremos varios casos de los últimos cinco años para reafirmar la importancia y la necesidad de regular este tipo de desarrollos de manera que se priorice nuevamente el bienestar humano.

También estudiaremos la necesidad de regular el uso de la IA en la búsqueda de medicinas, que aunque puede ser una herramienta valiosa para descubrir tratamientos, conlleva el riesgo de posibilitar la creación de toxinas. Es preocupante que, actualmente, cualquier persona con un computador de escritorio pueda estar desarrollando estas toxinas lo que resalta la urgencia de abordar este tipo de posibilidades.

Con estos tres casos que generan dilemas morales espero demostrar que las tensiones son reales y actuales, y que debemos reevaluar nuestros principios morales considerando cómo están alterado nuestro paradigma social.

El objetivo de esta investigación es destacar cómo el avance de la inteligencia artificial está moldeando los principios morales y el contexto en general. Esto nos obliga a considerarla desde perspectivas éticas, políticas y filosóficas que contemplen tanto las tecnologías como el entorno en el que se desarrollan y cómo este cambia. Se pretende demostrar cómo estos desafíos éticos requieren una profunda reflexión y un marco normativo para el desarrollo e implementación de la inteligencia artificial. La rápida evolución tecnológica exige que adaptemos constantemente nuestros principios morales, lo cual hace necesaria una mayor participación política y filosófica en la sociedad para lograr desarrollos tecnológicos éticos y políticos.

El trabajo incluirá una contextualización de la evolución de la inteligencia artificial seguida de las motivaciones para abordar estos desarrollos desde una perspectiva ética. Se cuestionará y contextualizará lo poco que se ha investigado en este campo. Finalmente, se desarrollarán brevemente los tres dilemas morales mencionados previamente, lo que nos llevará a considerar las implicaciones filosóficas que estos puedan acarrear.

## **1. Contexto IA**

### **1.1 Origen**

En el año 1956, durante una conferencia de investigación de verano celebrada en la Universidad de Dartmouth en Hanover, Nuevo Hampshire, un pequeño grupo de científicos acuñó por primera vez el término 'inteligencia artificial'. Aunque eran solo diez académicos, el grupo lo conformaban expertos en diversas áreas como teoría de la complejidad, simulación del lenguaje, redes neuronales, abstracción de contenido a partir de entradas sensoriales, aleatoriedad, pensamiento creativo y máquinas de aprendizaje. Liderados por John McCarthy, profesor de matemáticas de la universidad en aquel entonces, la conferencia tenía como objetivo específico proceder en la base conjetural a partir de la cual “todo aspecto

del aprendizaje o incluso cualquier otro rasgo de inteligencia puede en principio ser descrito de forma tan precisa que una máquina pueda simularlo” (McCarthy et al., 2006). Este fue el primer intento formal de descubrir cómo las máquinas pueden utilizar el lenguaje, formar abstracciones y conceptos, resolver problemas que antes eran exclusivos de la inteligencia humana e incluso mejorar por sí mismas.

Estas son las motivaciones iniciales que marcaron el surgimiento de un concepto que originó la investigación hoy conocida como inteligencia artificial. Sin embargo, aún no tenemos una idea delimitada sobre qué es realmente la inteligencia artificial, especialmente hoy cuando cada vez estamos más permeados por ella. Casi treinta años después de la mencionada conferencia (1985), el filósofo James H. Moor, quien también fue profesor en Dartmouth y llegó a dirigir la conferencia en su 50 aniversario, predijo este aspecto del presente en el que vivimos y habló sobre cómo la tecnología y las computadoras se convertirían en parte integral y esencial de nuestras vidas en un artículo de investigación. Moor planteó que la revolución de las computadoras se desarrollaría en dos etapas diferentes. La primera etapa ocurrió desde la década de 1940<sup>1</sup>, la cual consistió en la creación y el refinamiento gradual y exponencial de las computadoras electrónicas. La segunda etapa, que según el filósofo ya estaba ocurriendo en ese momento, era la etapa de la permeación. Moor afirmaba que la tecnología computacional se convertiría en una parte integral de nuestras instituciones en toda la sociedad, y que las instituciones sociales y las actividades humanas se transformarían para siempre debido a este tipo de tecnologías (p.270).

Moor, sin embargo, no era un adivino; simplemente se percató de lo que estaba detrás del intento de McCarthy por investigar la inteligencia humana y crear máquinas más competentes y conscientes. Él reconoció que el verdadero potencial residía en la naturaleza misma de las computadoras y su flexibilidad o maleabilidad lógica (Moor, 1985, p. 269). Estas pueden diseñarse y programarse para cualquier actividad en términos de entradas, salidas y conectores lógicos. Estos tres elementos juntos crean operaciones lógicas, que son los pasos que realiza una computadora y la guían desde una entrada hasta una salida. Dado que la lógica se aplica a todo, o, mejor dicho, prácticamente todo puede ser formalizado en un sistema lógico, Moor y muchos otros investigadores que ampliaron el campo de la IA llegaron a la conclusión de que la computadora parece tener aplicaciones ilimitadas y es lo más cercano a una herramienta universal.

Para nosotros, lectores del siglo XXI, esta idea no resulta innovadora y suena más como algo obvio. Como ya fue mencionado, nuestra realidad demuestra cada día más que la tecnología se ha vuelto fundamental para la humanidad, hasta el punto de que es casi una necesidad. Por ejemplo, el uso de teléfonos inteligentes en el planeta Tierra afecta a 6.92 billones de personas, es decir, al 86.11% de la población mundial (Turner A, 2023). Además, en el mundo existen 4.2 billones de asistentes de voz virtuales y se realizan aproximadamente 3.1 billones de búsquedas mensuales (Wardini, 2023). Con el tiempo, la etapa de permeabilidad predicha por Moor se está convirtiendo en una realidad que trae consigo una relación casi necesaria entre la humanidad y la tecnología. Parece, entonces, como argumentaba Mitcham

---

<sup>1</sup> La primera computadora electrónica programable se llamó ENIAC (Electronic Numerical Integrator and Computer) y fue creada en los Estados Unidos. Fue diseñada y construida por J. Presper Eckert y John W. Mauchly en la Universidad de Pensilvania. El ENIAC se completó en 1945 y se utilizó principalmente para cálculos balísticos durante la Segunda Guerra Mundial. Era una máquina enorme que ocupaba una gran sala y estaba compuesta por miles de tubos de vacío. Fue considerada un hito importante en el desarrollo de la informática y sentó las bases para el desarrollo de las computadoras modernas (Open AI, 2021).

(1996), que “la ontología de los artefactos no podrá divorciarse de la filosofía de la naturaleza” y que cada vez más el rasgo evolutivo que nos distingue es esta creación humana y no nuestra propia inteligencia.

Aunque el lector pueda identificarse con las cifras mencionadas, ya sea al consultar la hora en su celular o leer este artículo en su computadora, hay que destacar que no existe una única definición establecida y aceptada por los investigadores en el campo de estudio de la inteligencia artificial, y menos sobre qué investiga este campo y cómo lo hace. Por ejemplo, McCarthy (2021) define la inteligencia artificial como “la ciencia y la ingeniería de crear máquinas inteligentes.” Un estudio reciente de la Universidad de Stanford afirma que es “una ciencia y un conjunto de tecnologías computacionales inspiradas por, pero generalmente operando de manera diferente a, la forma en que los humanos utilizan su sistema nervioso y sus cuerpos para percibir, aprender, razonar y tomar acción” (Peter Stone, 2016). Torrance, un investigador centrado en la ética de las máquinas, la define como la “actividad de diseñar máquinas que realicen acciones que, cuando son realizadas por humanos, son indicativas de la posesión de inteligencia en esos agentes humanos” (Torrance, 2011). Arkoudas y Bringsjord agregaron que estas acciones deben ser realizadas por las máquinas en entornos controlados y durante períodos prolongados de tiempo (Arkoudas & Bringsjord, 2014, p. 34). Los autores del popular libro sobre IA, Stuart Russell y Peter Norving, sugieren que existen cuatro categorías distintas en este campo de investigación: i) sistemas que piensan como humanos; ii) sistemas que actúan como humanos; iii) sistemas que piensan de manera racional; y iv) sistemas que actúan de manera racional (Bundy, 2017).

Los investigadores la consideran más un campo de estudio que un término o concepto definido debido a esta amplia gama de interpretaciones. Esto ha generado numerosos subcampos de estudio, tales como el aprendizaje automático, la robótica, las redes neuronales, la visión por computadora, el procesamiento del lenguaje natural, el procesamiento del habla, entre otros. A medida que avancemos, se explorarán de manera más detallada algunos de estos subcampos. Por ahora, podemos concluir dos aspectos importantes sobre la inteligencia artificial que se desprenden de lo que hemos visto hasta ahora. En primer lugar, no debería sorprendernos la estrecha conexión que aparentemente existe entre la IA y la inteligencia humana. Gran parte de la investigación en este ámbito se ha centrado en cómo funciona la mente humana y cómo actuamos de manera inteligente para imitar esto en las máquinas. En segundo lugar, la investigación en IA abarca diversas áreas más allá de la ciencia de la computación, incluyendo la psicología, la neurociencia, la ciencia cognitiva, la filosofía, la estadística y la lógica.

## **1.2 Desarrollo**

Así pues, con un concepto tan poco definido vale la pena indagar por la manera en la que se desenvuelve este amplio campo de estudio. Hasta este punto, hemos abordado su interés por reflexionar sobre la inteligencia y cómo funciona en los seres humanos, para que máquinas inteligentes repliquen estos comportamientos. Es importante considerar qué habilidades humanas podemos implementar en el diseño de máquinas inteligentes y qué características nos distinguen definitivamente de otros animales y formas de inteligencia que podríamos desarrollar y no son imitables ya que la inteligencia humana es única en su especie (Allen et al., 2011, p. 15). La inteligencia es un tema fascinante que se analiza desde múltiples perspectivas, pero aún nos falta mucho por aprender al respecto. Sin embargo, algo que sabemos con certeza es que la inteligencia es deliberada. Esto significa que “implica algún tipo de objetivo, incluso si solo se trata de dar la respuesta correcta a una pregunta.” En otras

palabras, la inteligencia nos obliga a distinguir entre el actuar con intención o propósito y simplemente actuar (Everitt & Hutter, 2018, p. 12).

Las máquinas 'inteligentes' con las que interactuamos hoy no poseen facultades mentales comparables a las humanas. Sin embargo, la cuestión no es si es posible construir máquinas con facultades mentales humanas, es decir, con conciencia artificial. Retrocederíamos y olvidaríamos que no podemos experimentar lo mismo que un murciélago (Nagel, 2016). Sería asumir que nuestros estados mentales son exclusivos y necesarios para el acto intencional y la inteligencia, y no tenemos razones sólidas para afirmarlo de manera definitiva. La verdadera cuestión es si podemos construir máquinas que imiten este comportamiento intencional a través de modelos y algoritmos y alcancen sus objetivos. El que estudiemos la conciencia humana no quiere decir que tratemos de crear esta misma conciencia artificialmente, sino que se trataría de entender cómo funciona para imitarlo en forma adecuada para los softwares de la IA.

El estudio de la mente humana sigue siendo central, ya que es nuestro único punto de partida a posteriori para la IA. Es necesario comprender qué es la inteligencia humana y determinar si las facultades mentales humanas son una condición necesaria para la inteligencia, o si puede existir inteligencia artificial sin ellas y de qué manera; pero no por esto debemos concluir erróneamente que sin conciencia humana no pueda haber inteligencia y que sí o sí debamos reconstruirla. Además, no debemos enfocarnos solo en el estudio de la conciencia humana. Es difícil imaginar una acción intencional humana que requiera inteligencia y no esté acompañada de sensaciones de placer o repulsión, sensaciones corporales, emociones, creencias espirituales, euforia, depresión u otros estados de conciencia (Robinson, 2014, p. 67). Tampoco podemos ignorar que no solo el resultado de una acción determina si actuamos de manera inteligente, sino también la organización interna funcional que nos permite llevar a cabo dicho acto (Arkoudas & Bringsjord, 2014, p. 35).

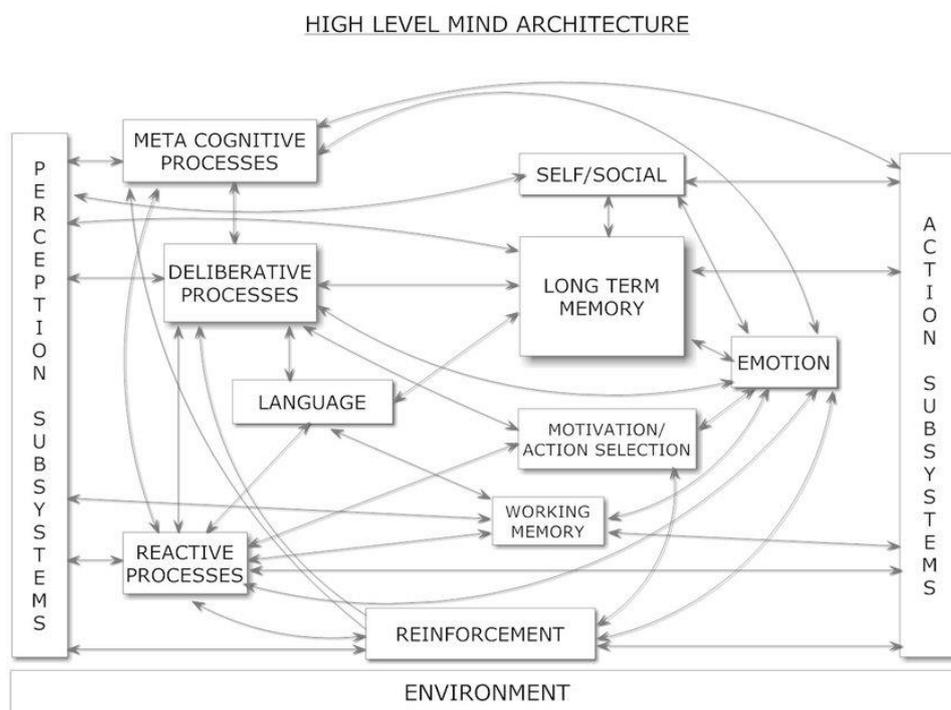


Figura 1: Estructura de alto nivel de una mente humana (Goertzel, 2014, p. 27)

Aunque la conciencia ha ayudado mucho en el desarrollo de la IA cómo veremos a continuación, si nos centramos en el funcionamiento actual de la IA, podemos confirmar que la mayoría de los sistemas no funcionan con conciencia o facultades mentales humanas, sino dando respuestas adecuadas a los propósitos programados en ellos y que, al realizarse por humanos, serían indicador de inteligencia. Los objetivos son instrucciones específicas en torno a las que se organiza la actividad de las máquinas, y el éxito se calcula según la realización de estos objetivos establecidos más que en si tienen conciencia humana o no (Robinson, 2014, p. 67).

Para comprender cómo la IA puede ser diseñada de forma deliberada sin poseer estados intencionales humanos, debemos introducir dos conceptos centrales: algoritmos y datos. Comencemos con el algoritmo, que es “un conjunto metódico de pasos que se pueden emplear para hacer cálculos, resolver problemas y tomar decisiones. No es un cálculo concreto, sino el método que se sigue cuando se realiza el cálculo” (Harari, 2018, p. 100). Los algoritmos son las pautas que describen cómo realizar una tarea, son la secuencia de instrucciones que le dicen a una computadora qué hacer. Por lo tanto, no debería sorprendernos que “el intento de formalizar las cosas con algoritmos conduzca a una comprensión más profunda de las mismas” (Gips, 2011, p. 10). De hecho, es una forma de conocer el método o los pasos detrás de una acción. Para poder utilizar un algoritmo con una computadora, necesitamos un medio donde emplearlo, es decir, necesitamos información completa, correcta y actualizada sobre la que se encuentren patrones y se puedan sacar conclusiones de ellos.

Afortunadamente, la humanidad siempre se ha organizado de acuerdo a métricas. Algunas de estas métricas son naturales y cíclicas, están basadas en las estaciones y los movimientos planetarios. Otras son religiosas, como el antes y después del nacimiento de Jesucristo (Floridi, 2012, pp. 3-4). Además, la ciencia nos ha mostrado a través del descubrimiento de reacciones químico-biológicas en nuestro interior que tanto los animales como los seres humanos, por más complejos que seamos, somos reducibles al procesamiento de información, lo cual puede explicar nuestro comportamiento (Bynum, 2012, p. 24). Nuestra tendencia a categorizar todo con métricas, junto con nuestra capacidad para almacenar y transmitir estos conocimientos, ha llevado a que poseamos grandes conjuntos de información. Pero el momento decisivo llegó cuando el uso creciente de internet y el mejoramiento en el poder de cómputo permitió la creación del big data y la minería de datos por la década de los 2000. Estos son conjuntos de información demasiado grandes para los softwares tradicionales de análisis de datos, pero que pueden ser minados por computadoras con software de aprendizaje para descubrir patrones y extraer conclusiones e información de estos conjuntos de datos (Andersen, 2021, p. 8).

La maleabilidad lógica de las computadoras, junto con estos grandes conjuntos de datos, ha permitido que la inteligencia artificial tenga un propósito más allá de los estados mentales humanos. Su propósito es el de almacenar información en su memoria y resolver problemas al manipular esos datos de manera sistemática dando un resultado a partir de ellos (Arkoudas & Bringsjord, 2014, p. 41). Por eso, no debería sorprendernos que la carrera por los datos ya haya comenzado. Como nos muestra Harari (2018a), gigantes como Google, Facebook, Baidu y Tencent han adoptado el modelo de negocio de los 'mercaderes de atención'. Nos proporcionan información, servicios y diversión de forma gratuita y luego la revenden a nuestros anunciantes en la web. Pero su verdadero negocio no es vender anuncios. Es acumular cantidades inmensas de datos sobre nosotros en este proceso. No somos sus clientes: somos su producto (pp. 100-101).

Así, aunque nuestras máquinas inteligentes no posean facultades mentales humanas, sí pueden procesar cantidades masivas de datos y extraer patrones e información útil, incluso a un ritmo exponencialmente más rápido que los humanos, y de conjuntos que incluso pueden ser muy grandes para las capacidades humanas (Bradley, 2017). Esto se debe a varios factores. Primero, las computadoras son mucho más precisas que nosotros para hacer cálculos aritméticos y calcular probabilidades de datos. En segundo lugar, las máquinas son imparciales y, si se les proporciona un conjunto adecuado de datos, pueden brindar una cierta sensación de objetividad. Además, las máquinas pueden analizar muchas más probabilidades en mucho menos tiempo por sus avanzados y eficaces procesadores (Anderson & Anderson, 2007, p. 18). Más aún, el factor fundamental para que veamos máquinas que nos superan en ciertas actividades fue crear algoritmos que pueden aprender o mejorar con el tiempo en una tarea. Este innovador algoritmo se denomina Aprendizaje Automático (Machine Learning), ya que las máquinas parecen aprender de los conjuntos de datos con el tiempo. Ya no es solo un programador que instala un programa específico que permite a una computadora realizar una tarea en particular, ahora hablamos de programas estadísticos diseñados para que las computadoras puedan explicar el conjunto de datos o predecir conjuntos de datos similares (Andersen, 2021, p. 9). En pocas palabras, este es el funcionamiento de un algoritmo de Aprendizaje Automático (Machine Learning):

- Los programadores comienzan con un conjunto de datos histórico que luego dividen en un conjunto de entrenamiento y un conjunto de prueba.
- Luego eligen un modelo, que es una estructura matemática capaz de categorizar un rango de reglas de decisión en relación con el conjunto de datos, y que es ajustable.
- También se define una función objetivo para medir la deseabilidad del resultado.
- Entrenan el modelo para que se ajuste y maximice el resultado de la función objetivo.
- Después de esto, se prueba el modelo con el conjunto de prueba y se evalúa su efectividad y precisión utilizando un nuevo conjunto de datos. El objetivo final es generalizar el modelo para cualquier conjunto de datos.

El Aprendizaje Profundo ha sido uno de los avances más significativos en el campo de la inteligencia artificial. De hecho, combinado con otras tecnologías inteligentes, ha comenzado a mostrar su potencial en áreas como los negocios, la salud, el transporte y muchos otros campos que nos afectan a diario (Siau & Wang, 2020, p.74). Su desarrollo más destacado se ha producido al fusionarse con los avances en redes neuronales inspiradas en el cerebro, lo que confirma que el estudio de la mente puede ayudar en el desarrollo de la IA. Esto llevó a la creación de las Máquinas de Aprendizaje Profundo (Deep Learning Machines). Imitando a las neuronas, las MAP tienen múltiples unidades diferentes, cada una de las cuales recibe un conjunto de entradas, ya sea del conjunto inicial o de la unidad anterior, y produce un resultado. Esto permite que cada una de estas unidades perciba patrones más precisos y en conjuntos de datos cada vez más grandes (Andersen, 2021, p. 9).

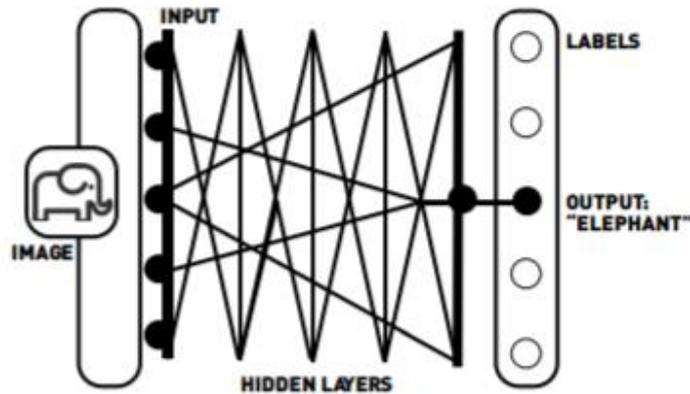


Figura 2: Este es el proceso de la mente humana que imitan las Maquinas de Aprendizaje Profundo (Andersen, 2021, p. 9).

Desde la aparición de las tecnologías de Aprendizaje Profundo, por ejemplo, hemos visto cómo se combinan con los chat-bots y hemos mejorado día a día en la modelación de conversaciones y relaciones humanas. En 2015, un chat-bot programado para parecer un niño de 13 años llamado Eugene Goostman superó a más de la mitad de las personas que lo evaluaron en la prueba de Turing. Los investigadores conversaron con el bot durante un tiempo y luego sacaron sus propias conclusiones sin saber si estaban hablando con un bot o un humano<sup>2</sup>. Este mismo año, el lanzamiento de ChatGPT ha cambiado la forma en que realizamos y percibimos muchas tareas. Las personas están obteniendo beneficios con la ayuda de este tipo de chat-bots, desde datos estadísticos, hasta información relevante sobre cualquier tema, y muchas otras posibilidades que demuestran cómo la IA ya está impregnando nuestra realidad de manera exponencial.

Además de los avances en la interacción a través de chat-bots, otro ejemplo notable es el desarrollo de los autos autónomos. En el año 2015, 1.25 millones de personas morían en accidentes automovilísticos, y el 90% de estos se debía a errores humanos (Harari, 2017). Según ChatGpt (2021), la OMS (Organización Mundial de la Salud) estima que alrededor de 1.35 millones de personas mueren anualmente por esto. Como el lector probablemente haya experimentado, esto puede deberse a la falta de comunicación, el exceso de velocidad, el estado de embriaguez o la falta de señalización en la vía. En respuesta a esta alarmante cifra, "varias empresas como Google y Tesla producen autos autónomos que ya circulan por nuestras calles. Los algoritmos que controlan el auto realizan millones de cálculos por segundo en relación con otros autos, peatones, semáforos y baches" (Harari, 2018b, pp. 132-133). Los autos autónomos pueden conectarse, lo que reduce las posibilidades de colisión. Veamos el potencial de esta idea. En agosto de 2015, un auto experimental de Google tuvo un accidente. Al acercarse a un cruce, detectó la presencia de peatones y aplicó los frenos. Sin embargo, un automóvil lo chocó por detrás debido a un conductor descuidado al volante de un Sedán que no se percató de la frenada del auto. Si ambos autos hubieran estado conectados a través de sistemas interconectados por computadoras, el algoritmo de posicionamiento habría conocido la posición e intención de cada vehículo y no habría permitido la colisión (Harari, 2018b, p. 342).<sup>3</sup>

<sup>2</sup> <https://time.com/2847900/eugene-goostman-turing-test/>

<sup>3</sup> De acuerdo a ChatGPT, el mayor desarrollo de estas tecnologías ocurre en:

Los vehículos autónomos y Chat GPT ilustran dos puntos importantes. En primer lugar, puede tener sentido reemplazar a los humanos por robots o computadoras en algunos campos. El cambio en algunos campos puede ocurrir de manera tan repentina que ni nos damos cuenta cuando estamos rodeados de robots. El aprendizaje automático y la inteligencia artificial han demostrado su potencial para simplificar, acelerar y mejorar muchos aspectos de nuestra vida (Bradley, 2017). A pesar de los avances logrados, sigue siendo cuestionable si estas máquinas son realmente inteligentes. Aunque pasen el Test de Turing e incluso nos superen en algunas actividades, todavía no tenemos fundamentos suficientes para afirmar que las máquinas con algoritmos de aprendizaje profundo son inteligentes. Esto no implica que las facultades humanas sean necesarias para actuar de manera inteligente. Simplemente queremos señalar que, hasta ahora, no contamos con suficiente evidencia para afirmar que las máquinas con algoritmos de aprendizaje profundo poseen verdadera inteligencia.

El consenso general ha optado por denominar a este tipo de IA descritas hasta ahora como IA estrecha. Esto, como su nombre lo indica, se debe a que son sistemas que se aplican para cumplir un solo objetivo, como el reconocimiento de imágenes, la traducción de lenguaje y los vehículos autónomos mencionados anteriormente (Andersen, 2021, p. 8). Sin embargo, se está desarrollando un nuevo término y campo de investigación, que algunos incluso denominan IA "real", pero que llamaremos IA general. Los científicos afirman que lo que falta para lograr inteligencia en máquinas como estas es la generalidad (Bostrom & Yudkowsky, 2014, p. 3). Si bien Deep Blue se convirtió en el campeón mundial de ajedrez, no podría jugar damas chinas. La inteligencia artificial pretende cumplir una amplia gama de objetivos y realizar tareas en diferentes contextos y ambientes. Un sistema inteligente suficientemente general debe poder manejar problemas y situaciones muy diferentes en cuanto a su contexto. Hasta este momento, la humanidad muestra un nivel mayor de inteligencia en cuanto a la inteligencia general se refiere, aunque no necesariamente en todas las tareas específicas.

Aún es muy ambiguo lo que separa a la inteligencia general de la inteligencia estrecha, podemos basarnos en la hipótesis lanzada por Goertzel en 2014 para afirmar que, al menos cualitativamente hablando, son diferentes en su creación y en su resultado (Goertzel, 2014, p. 3). Esta hipótesis se conoce como la hipótesis central de la inteligencia artificial general y fue fuertemente influenciada por el factor g del psicólogo Spearman, quien sostiene que nuestra inteligencia se compone de ocho componentes principales: lingüístico, lógico-matemático, musical, corporal kinestésico, espacial, interpersonal, intrapersonal y naturalista (Goertzel, 2014, p. 8). Imaginemos un robot que realice nuestras compras, nos ayude a separar la fruta en buen estado de la podrida, a ajustarnos en los autobuses que nos transportan y recordarnos

---

Estados Unidos: Estados Unidos ha estado a la vanguardia de la tecnología de los coches autoconducidos. Empresas como Waymo (filial de Alphabet), Tesla y Uber han estado realizando pruebas y ensayos exhaustivos en estados como California, Arizona y Texas.

China: China también ha hecho avances significativos en la tecnología de los coches autoconducidos. Empresas como Baidu, Pony.ai y WeRide han realizado pruebas y ensayos en ciudades como Pekín, Shanghái y Guangzhou.

Alemania: Alemania, conocida por su industria automovilística, ha participado activamente en el desarrollo de coches autónomos. Empresas como BMW, Audi y Mercedes-Benz han realizado pruebas en carreteras públicas, especialmente en ciudades como Múnich y Stuttgart.

Singapur: Singapur ha sido pionero en la adopción de la tecnología de los coches autónomos. Empresas como nuTonomy (ahora parte de Aptiv) y Grab han participado en las pruebas.

Japón: Japón ha estado invirtiendo en tecnología de coches autónomos, con empresas como Toyota, Honda y Nissan a la cabeza. Se han realizado ensayos y pruebas en ciudades como Tokio y Yokohama.

(Open AI, 2021)

nuestros compromisos sin que tengamos que pedirle que haga esto. Un sistema posee una mayor inteligencia general si puede adaptarse de manera efectiva a diferentes entornos sin requerir modificaciones por parte del diseñador. Sin embargo, como veremos en la siguiente sección, construir una IAG también conlleva implicaciones en cuanto a cómo limitar los campos de acción de una inteligencia artificial como esta y dónde trazar esos límites

Tras revisar el desarrollo e implementación de la IA en nuestra sociedad, observamos que aún hay algunas preguntas sin respuesta. La inteligencia estrecha ha demostrado que las máquinas pueden realizar tareas específicas que denotan inteligencia mejor que los seres humanos y sin sus aptitudes cognitivas. Debemos preguntarnos si la conciencia humana es necesaria para que haya este tipo de actos porque aparentemente esto no es así. Además, el objetivo final es construir una máquina inteligente general, aunque todavía no tenemos claridad sobre los requisitos y las características que distinguen a cada tipo de inteligencia. Esta incertidumbre nos plantea la cuestión de si es mejor seguir el camino estrecho de la inteligencia o apostar por generalizarla y centrar nuestro enfoque en ello, así como también por buscar una mayor delimitación y claridad en las características esenciales de cada tipo de inteligencia artificial y sus diferencias. Nuestros laboratorios de IA benefician el estudio y desarrollo de la mente humana y sus facultades mentales, lo que contribuye al progreso en este campo y nos beneficia con las nuevas tecnologías (Allen et al., 2011, p. 15).

### **1.3 Singularidad**

¿Hacia dónde se dirige la inteligencia artificial ahora? Todo indica que la IA no tiene intenciones de rendirse en el desarrollo de tareas específicas a un nivel superior al humano, ni en la creación de una inteligencia artificial general que nos permitiría construir una máquina capaz de interactuar en muchos, si no en todos, los contextos humanos.

Desde los años 60, se ha planteado la posibilidad de crear una máquina ultra-inteligente. Según Good (1966), una máquina ultra-inteligente es aquella que puede superar ampliamente todas las actividades intelectuales que se consideran inteligentes en los seres humanos. Una ultra-inteligencia debe cumplir con los siguientes aspectos: 1) tener objetivos claros, 2) identificar las posibles acciones en cualquier situación, 3) considerar las posibles consecuencias de cada acción, 4) tomar la acción que más probablemente conduzca a sus objetivos, y 5) actualizar su modelo del mundo para estar al tanto de lo que sucede en el planeta Tierra. Una máquina ultra-inteligente, es una máquina que contenga todas las inteligencias específicas de los diferentes ambientes y contextos y nos supere en cada uno de ellos. Es la conciliación última de ambos tipos de inteligencia artificial (p. 37).

Los primeros pasos hacia esta dirección se pueden observar en dispositivos como Siri de Apple, en Watson de IBM (que destaca no solo por ganar concursos televisivos, sino también por diagnosticar y tratar enfermedades), y en muchas otras inteligencias artificiales similares. Sin embargo, lo que aún no hemos presenciado, pero que resulta verdaderamente alarmante, es lo que el filósofo Ray Kurzweil denominó singularidad (Bradley, 2017). Dado que el diseño de máquinas es una actividad intelectual, una máquina ultra inteligente podría llegar incluso a diseñar máquinas mejores que las diseñadas por nosotros. Esto podría llevar a una explosión de inteligencia, y la supremacía humana se convertiría en una reliquia evolutiva, al igual que nuestras muelas del juicio (Vinge, 1993, p. 16). Se trata de una situación hipotética en la que la humanidad habría desarrollado la inteligencia artificial hasta un punto en el que se produciría un cambio drástico e irreversible. Aunque lo que sabemos hasta ahora no nos permite afirmar con certeza si este momento llegará, si sucediera, sería un evento

sorprendente y probablemente ocurriría más rápido que cualquier otro evento que hayamos presenciado.

Mientras Darwin desarrollaba su teoría de la evolución, afirmaba que era difícil creer que algo tan mecánico e inconsciente como un algoritmo pudiera haber producido cosas tan hermosas como las que vemos en la evolución del planeta Tierra. Sin embargo, cada vez parece más evidente que esto no es así, y que en realidad se trata de un algoritmo recursivo que funciona tan lentamente que fue casi imperceptible para los seres humanos, de no ser por la inteligencia artificial. Harari (2018a) ejemplifica esto con el ojo humano y su evolución. Según el filósofo, el camino desde los sensores arcaicos de organismos unicelulares hasta el ojo humano es largo y tortuoso, pero si disponemos de millones de años, es posible recorrerlo en su totalidad. Esto es posible porque el ojo lo componen muchas partes diferentes y puede transferirse a datos (p. 122). Varios factores han cambiado y han hecho que este tipo de comprensión sea cada vez más posible. En primer lugar, los computadores trabajan mucho más rápido ahora y son mucho más fáciles de conectar e incluso de programar. Además, el entendimiento de las ciencias cognitivas y de las neurociencias cognitivas ha avanzado lo suficiente como para hacer un diagrama empírico de la mente humana. También ha habido avances significativos en el campo de la robótica y del mundo virtual. Por último, la inteligencia artificial estrecha ha mejorado los algoritmos hasta el punto en que podemos resolver casi cualquier tarea específica, por muy compleja que sea (Hugo De & Goertzel, 2009, p. 122).

En un contexto anterior, vimos la distinción que Moor realiza entre dos etapas diferentes en el desarrollo de la inteligencia artificial. Además, el filósofo distingue cómo percibiremos la IA en ambos contextos. En primer lugar, las vemos como simples herramientas que evaluamos en función de su capacidad para cumplir su función. Sin embargo, en la etapa de permeabilidad, las inteligencias artificiales se verán cada día más como parte integral de nuestra vida diaria e incluso como agentes en la sociedad. Un ejemplo de esto fue lo que sucedió en las últimas elecciones presidenciales de Estados Unidos, donde las cadenas televisivas proyectaron los resultados con la ayuda de la IA y supieron quién ganó incluso antes de que se cerraran las votaciones en California y Nueva York. En este momento, la pregunta ya no es qué tan eficientes pueden ser los computadores, ya que hemos visto su amplio rango de posibilidades y su eficacia. La pregunta ahora se enfoca en la naturaleza de esta actividad y cómo se relaciona con la humanidad, ya que la respuesta más efectiva parece ser una relación simbiótica entre la humanidad y la IA (Moor, 2017, p. 227). Debemos comenzar a pensar en la tecnología de la misma manera en que pensamos en la naturaleza.

## **2. Motivaciones Éticas**

Investigadores continúan explorando los límites y alcances de la IA. En la sociedad actual, nos encontramos en una constante y permanente relación con diversas máquinas inteligentes que no solo nos ayudan y ahorran tiempo, sino que también toman decisiones en nuestro nombre. Como señala Andersen (2021), se utilizan las salidas generadas por la inteligencia artificial a través de sus algoritmos para la toma de decisiones (p. 11). Sin embargo, algunos piensan que esto no implica que debamos pensar en la naturaleza de esta relación, como lo mencionó Moor, ni mucho menos en las implicaciones que pueda tener. Creen, en cambio, que lo importante sigue siendo la eficiencia de la inteligencia artificial y la cuestión de si podemos alcanzar la singularidad o no. Por esto, en lo que viene trataré de suscitar motivaciones para el correcto, estudio, desarrollo e implementación de la ética de máquinas generales y sus posteriores implicaciones.

## 2.1 Accidentes

Ciertamente, las ventajas que nos están trayendo las IA con algoritmos de decisión son maravillosas e increíbles. Google Maps, Waze y miles de otros GPS, por ejemplo, son algoritmos de decisión inteligente que, junto con conjuntos de datos acerca de mapas, trancones, presencia policial, entre otros, logran ahorrarnos mucho tiempo al llevarnos por la ruta más efectiva sin que debamos hacer un solo cálculo. Aunque la mayoría de las veces dicha relación es beneficiosa, sin embargo, algunas veces no sale como nosotros lo esperamos. En 2016, por ejemplo, una mujer siguió con fe ciega a su GPS en Canadá y acabó dentro del lago Ontario<sup>4</sup>. Aunque por lo general la relación entre inteligencia artificial y humanidad muchas veces es beneficiosa, no podemos asumir sin más que siempre lo será y que no debemos cuestionar cómo llevar a cabo esta relación. El ejemplo previo nos muestra que el uso excesivo del GPS y de tecnologías similares puede ocasionar accidentes. En el fondo, el problema radica en que orientarse, al igual que cualquier otra capacidad humana, es como un músculo: si no la entrenamos, acabaremos perdiéndola (Harari, 2018, p. 76). Si no limitamos esta relación a los aspectos saludables que nos pueda traer, puede acabar haciéndonos más daño que beneficio.

El filósofo Allen (2011) nos brinda otro ejemplo de una decisión tomada por una inteligencia artificial que afectó accidentalmente la vida de alguien. Durante un viaje de Texas a California, una persona condujo sin hacer paradas, y cuando intentó usar sus tarjetas en su destino, estas fueron bloqueadas por la inteligencia artificial del ordenador central de su banco. Al no poder rastrear ningún uso entre estas dos ciudades y considerando la gran distancia entre ellas, la IA supuso accidentalmente que el bloqueo de la tarjeta sería la mejor opción (p. 15). Ejemplos como estos, vivimos en una sociedad en la que las máquinas inteligentes toman decisiones, y algunas pueden afectar nuestra vida tanto beneficiosamente como perjudicialmente. Además, aunque hasta este momento hemos visto ejemplos de perjuicio exclusivamente accidental, es importante reconocer que las máquinas inteligentes también pueden afectarnos si se les da un uso malintencionado.

## 2.2 Maleficencia

Imaginemos pues, lo que es peor, que las decisiones de las máquinas no nos afecten de forma accidental, sino de forma intencional. Las estafas electrónicas, las guerras cibernéticas y prácticamente todos los delitos cibernéticos se verán potenciados por la IA (Taddeo, 2018). Por ejemplo, actualmente se imita la voz de familiares para robar a mayor escala y de manera más efectiva<sup>5</sup>. También, se están usando para robar datos y dinero a empresarios<sup>6</sup>. Además, podrían surgir nuevos delitos (King et al., 2018). Ahora no solo podemos imitar la voz, sino también la imagen de algo o alguien para hacerlo o decir algo que nunca hizo o dijo en realidad. Los videos falsos, como el que circuló de Obama, inundan la web. La desinformación política se vuelve cada vez más inminente y peligrosa. Un video falso incluso podría desencadenar un conflicto si está muy bien hecho y no se maneja de manera cautelosa. La IA tiene un potencial considerable para ser utilizada en formas perjudiciales, como la manipulación de información, la creación de contenido falso, el aumento de la

---

<sup>4</sup> <https://torontosun.com/2016/05/13/woman-follows-gps-ends-up-in-ontario-lake>

<sup>5</sup> <https://www.washingtonpost.com/technology/2023/03/05/ai-voice-scam/>

<sup>6</sup> <https://virtuoso.tech/ai-brings-more-convincing-phishing-scams-for-business-owners/>

automatización de ataques cibernéticos, el desarrollo de armas autónomas y el seguimiento masivo y la vigilancia.

## 2.3 Desigualdad

El mal uso y los accidentes que puedan ocasionar las máquinas con inteligencia artificial no son los únicos peligros que acarrea el desarrollo de estas tecnologías. Además, existe una marcada desigualdad en esta situación. Solo una pequeña fracción de la humanidad está involucrada en el diseño, desarrollo y uso de estas tecnologías que están transformando nuestras vidas (Floridi et al., 2018). En cuanto al uso, tomemos el caso de Angelina Jolie<sup>7</sup>, quien en el año 2013 se sometió a una mastectomía no debido a tener cáncer en sí, sino por tener un 87% de riesgo de desarrollar cáncer de seno. Sin embargo, esta cirugía preventiva también pone de manifiesto la brecha económica que existe entre los seres humanos. Solo las pruebas genéticas costaron 3000 dólares en ese momento, sin tener en cuenta el precio de la mastectomía, la cirugía reconstructiva y los tratamientos asociados.

Según los datos de la PIP (Plataforma de Inequidad y Pobreza),<sup>8</sup> 659 millones de personas viven actualmente con 2.15 dólares diarios o menos, lo cual constituye el índice de pobreza extrema. Otros 1,831 millones viven con 3.65 dólares diarios y 3,634 millones con 6.85 dólares diarios. Más de la mitad de la población vive en la pobreza. Según el informe sobre desigualdad global de 2022, el 50% más pobre de la población vivió con un promedio de 4,100 dólares en el año 2021, el 40% de la mitad con 57,300 dólares y el 10% más rico con 771,300 dólares. Considerando los precios de vivienda, comida y otros gastos, estas cifras significan que solo el 10% de la población más adinerada puede permitirse someterse a procedimientos como estos.

Region	Survey Coverage (%)	\$2.15 (2017 PPP)				\$3.65 (2017 PPP)				\$6.85 (2017 PPP)			
		Headcount ratio (%)		Number of poor (mil)		Headcount ratio (%)		Number of poor (mil)		Headcount ratio (%)		Number of poor (mil)	
		Mar 2023	Sep 2022	Mar 2023	Sep 2022	Sep 2022	Mar 2023	Sep 2022	Mar 2023	Sep 2022	Mar 2023	Sep 2022	Mar 2023
East Asia & Pacific	97.4	1.1	1.2	24	25	7.6	7.6	160	161	32.1	32.1	675	676
Europe & Central Asia	87.4	2.4	2.3	12	11	6.2	6.1	31	30	15	15	74	74
Latin America & Caribbean	86.7	4.3	4.3	28	28	10.6	10.6	68	68	28	28	180	179
Middle East & North Africa	48.3	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Other High Income	82.3	0.6	0.6	7	7	0.8	0.8	9	9	1.4	1.3	15	15
South Asia	96.4	8.5	8.6	156	161	42	42.3	772	788	82.2	82.3	1508	1532
Sub-Saharan Africa	54.3	35.1	34.9	389	391	62.4	62.3	691	698	86.5	86.4	958	969
Eastern & Southern Africa	29.6	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Western & Central Africa	90.5	27.2	27.3	122	124	57.1	57.2	255	260	85.1	85.1	380	387
World	84.6	8.4	8.5	648	659	23.5	23.6	1803	1831	46.7	46.8	3590	3634

Figura 3: Los datos de la pobreza según PIP

Algunos argumentarán, sin embargo, que el informe de la PIP también muestra una disminución en la pobreza desde los años 90. No tengo fundamentos para refutar esto, pero,

<sup>7</sup> <https://www.nytimes.com/2013/05/14/opinion/my-medical-choice.html>

<sup>8</sup> <https://blogs.worldbank.org/opendata/march-2023-global-poverty-update-world-bank-challenge-estimating-poverty-pandemic>

aunque sea cierto, la desigualdad ha aumentado en contraposición. En 2016, las 62 personas más ricas del mundo poseían tanta riqueza como la mitad más pobre del planeta (Harari, 2018, p. 380). Actualmente, ese 1% de la población sigue teniendo la mitad de la riqueza mundial y más<sup>9</sup>, pues ganó el doble de dinero que el 99% restante el solo el anterior año<sup>10</sup>.

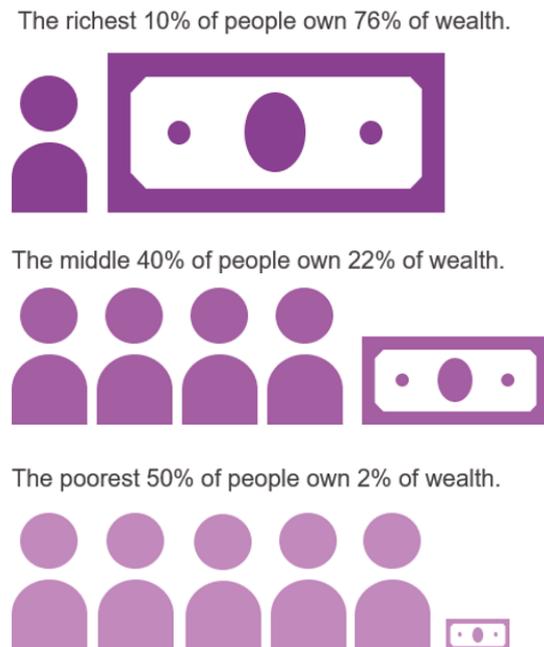


Figura 4: Distribución de la riqueza en el 2021

Actualmente ya estamos presenciando la acentuación de la desigualdad que estas tecnologías están generando. Por ejemplo, en 2014, las tres compañías más grandes de Silicon Valley ganaron la misma cantidad de dinero que las tres compañías más grandes de Detroit. Sin embargo, en Silicon Valley hay diez veces menos empleados (Bossman, 2016). El crecimiento tecnológico no mejorará las condiciones en todo el mundo, sino solo en los centros de alta tecnología, lo que genera una mayor segregación entre los países en desarrollo (Harari, 2018, p. 60). De esta manera, la desigualdad en Silicon Valley es más pronunciada que a nivel global, donde el 1% más rico posee 48 veces más riqueza que la mitad más pobre<sup>11</sup>.

## 2.4 Sustitución

Estas tecnologías pueden ser tan efectivas que incluso pueden llevar a la sustitución de los humanos en algunas o todas las áreas cognitivas en las que nos desempeñamos. Los seres humanos poseemos capacidades físicas y cognitivas. Después de la revolución industrial,

<sup>9</sup> <https://oxfamlibrary.openrepository.com/bitstream/handle/10546/621477/bp-survival-of-the-richest-160123-en.pdf>

<sup>10</sup> <https://www.globalcitizen.org/en/content/wealth-inequality-oxfam-billionaires-elon-musk/>

<sup>11</sup> <https://calmatters.org/california-divide/2023/02/silicon-valley-inequality/>

muchas de nuestras capacidades físicas fueron automatizadas y las máquinas se encargaron de realizar nuestro trabajo físico en el campo. Lo mismo podría estar ocurriendo en la actualidad con nuestras capacidades cognitivas (Harari, 2018, pp. 38-39). La IA está comenzando a superar a los humanos en actividades que requieren capacidades cognitivas. El ejemplo previo del automóvil Sanderó que chocó contra un vehículo de Google demuestra lo difícil que es para los humanos estar conectados y mantenernos actualizados. Por otro lado, las computadoras y las inteligencias artificiales, al no ser individuos, pueden ser fácilmente conectadas en una red integrada y flexible.

Lo mismo sucede con el caso de Watson, la inteligencia artificial de IBM diseñada para el diagnóstico de enfermedades. Si bien esta tecnología es más conocida por haber ganado un concurso de televisión, como ya se mencionó, en el día a día se utiliza para diagnosticar enfermedades y prescribir medicamentos. Si comparamos, no a la mejor persona médica del mundo sino a la red de médicos, con una red de múltiples Watson en todo el mundo, rápidamente llegaremos a la conclusión de que la IA superaría a los humanos. Según Harari (2018, pp. 345-346), existen tres ventajas en este sentido. En primer lugar, Watson podría contener en su sistema datos sobre todas las enfermedades conocidas y todos los medicamentos de la historia. Además, puede actualizar estos bancos de datos a diario, no solo con nuevas investigaciones, sino también con estadísticas de todas las clínicas y hospitales del mundo. En segundo lugar, Watson puede estar familiarizado en gran medida con el genoma y el historial médico de todas las personas en el mundo, y puede saber de inmediato si alguien ha visitado un país tropical, si padece alguna enfermedad recurrente, si hay un brote de diarrea en una ciudad e incluso si existe riesgo de cáncer en su familia. Por último, Watson nunca se cansará, tendrá hambre, estará enfermo o estará ocupado. En cambio, según Harari (2018, p. 42), es casi imposible para todos los médicos del mundo mantenerse al día con nuevas enfermedades y nuevos medicamentos.

## **2.5 ¿Tenemos Motivos?**

Estas son motivaciones suficientes para otorgar importancia al campo de la ética en el desarrollo tecnológico en el que nos encontramos inmersos. El rápido despliegue de inteligencias artificiales que toman decisiones relacionadas con nuestras vidas nos ha afectado, en primer lugar, de manera accidental. El uso constante de estas tecnologías puede afectar nuestro funcionamiento, ya que, por el uso excesivo, ni siquiera estamos ejercitando nuestras funciones cognitivas básicas. Actividades como caminar, orientarnos o realizar cálculos matemáticos simples están en riesgo de desaparecer si no estamos acompañados por la tecnología. Nuestra excesiva confianza en estas tecnologías puede llevarnos a sufrir daños accidentales e incluso afectar nuestra cognición a largo plazo de manera involuntaria.

El uso malintencionado de tecnologías inteligentes también ha demostrado ser un problema en el desarrollo de la inteligencia artificial. A medida que el desarrollo avanza de manera exponencial, las posibilidades se vuelven infinitas, no solo para crear tecnologías que nos benefician, sino también aquellas que puedan perjudicarnos o beneficiar a unos pocos. Este problema se vuelve aún más delicado si consideramos que no existe una relación de confianza inherente en el desarrollo de la inteligencia artificial, es decir, la ley no impone una obligación de confianza en dicho desarrollo tecnológico. Aunque algunos países están empezando a intervenir políticamente y promulgar leyes para regular el desarrollo de las

inteligencias artificiales, este proceso aún está en una etapa incipiente, ya que se pueden contar con los dedos de la mano los países que están tomando medidas en este sentido<sup>12</sup>.

Además, debido a esta falta de regulación, no podemos confiar en que los desarrolladores de la inteligencia artificial estén actuando con nuestros mejores intereses en mente al diseñar y crear esta tecnología (Mittelstadt, 2019, p. 3). Lo mismo sucedió con el análogo más cercano a la inteligencia artificial: la internet. Se tardaron 30 años en establecer leyes y regulaciones adecuadas para su uso, y durante ese tiempo ocurrieron numerosos problemas como el uso indebido de datos, piratería informática, ciberacoso, contenido inapropiado, noticias falsas, fraudes, entre otros. Sin percatarnos, o más bien, sin darle la importancia que merece, podríamos estar repitiendo el mismo error con la inteligencia artificial. Es necesario que la ley regule el desarrollo, implementación y uso de estas tecnologías. Ya sea con leyes o sin ellas, es fundamental contrastar el progreso social que puede traer la inteligencia artificial con los riesgos involucrados para su posterior desarrollo.

Sin embargo, esta tarea no es sencilla, ya que la desigualdad es un aspecto inherente en el desarrollo de las IA. Como ya vimos, el uso de estas tecnologías está siendo exclusivo y limitado al 10% más rico de la población. En medicina del siglo XXI, por ejemplo, se observa cada vez más la tendencia de mejorar la salud de los individuos sanos en vez de curar a los enfermos (Harari, 2018, p. 380). Esto se evidencia, por ejemplo, en el nacimiento de las primeras gemelas genéticamente modificadas en 2019<sup>13</sup>. La inteligencia artificial podría incluso conducir a una especiación en la que la humanidad se vea dividida en diferentes castas biológicas o especies distintas (Harari, 2018, p. 99). Además, son muy pocas las personas involucradas en su desarrollo, construcción y diseño. Es fundamental que guiemos nuestros esfuerzos a un enfoque más holístico que nos permita ver posibles riesgos e implementaciones injustas en estas tecnologías.

La sustitución completa en áreas cognitivas en las que los humanos desempeñamos un papel fundamental también representa un riesgo inminente. Muchas personas podrían quedarse sin empleo de un momento a otro, lo que conlleva un peligro potencial para un gran porcentaje de la humanidad. En Estados Unidos, hace un año, ya era común ver a personas trabajando o durmiendo dentro de sus automóviles mientras la inteligencia artificial de Tesla conducía por ellas. Sin embargo, los automóviles son solo una de las muchas áreas en las que esto puede ocurrir. También hemos presenciado cómo la medicina se encuentra en peligro. Incluso profesiones como diseñadores e ilustradores podrían verse afectadas por este creciente desarrollo. Dado que la mente humana sigue siendo un territorio poco explorado, no sabemos qué nuevas profesiones podrían surgir para reemplazar las pérdidas, tal como ocurrió durante la revolución industrial. Si no gestionamos con cautela la sustitución de seres humanos por inteligencias artificiales, podríamos enfrentar una crisis sin precedentes.

Tenemos, por tanto, motivaciones suficientes para el desarrollo de la ética en las máquinas inteligentes. La construcción de estas tecnologías se originó y continúa para obtener el máximo beneficio posible y minimizar los riesgos asociados. Además, aunque aún no tenemos garantías absolutas al respecto, es posible que pensar y desarrollar la ética de la inteligencia artificial, de manera similar a lo que ocurrió con la conciencia, nos ayude a comprender mejor la ética en nosotros mismos (Moor, 2011, p. 21). Es esencial abordar los aspectos éticos de la inteligencia artificial para garantizar que se utilice de manera responsable y en beneficio de la humanidad en su conjunto.

---

<sup>12</sup> <https://www.euronews.com/next/2023/05/03/which-countries-are-trying-to-regulate-artificial-intelligence>

<sup>13</sup> <https://www.nature.com/articles/d41586-019-00673-1>

### 3. Ética de Inteligencia Artificial

Aunque la IA puede desempeñar un papel significativo en mejorar nuestra calidad de vida, antes de implementarla debemos asegurarnos de que no nos causará a la vez perjuicios ni riesgos significativos como los ya vistos. Así entonces, para distinguir entre un funcionamiento correcto y uno incorrecto debemos hacer la pregunta, ¿qué es la ética? Específicamente, ¿qué es la ética de las máquinas inteligentes? Anderson (2011) la define como asegurarse de que el comportamiento de las máquinas inteligentes hacia los humanos sea éticamente aceptable. Sin embargo, esta definición deja de lado el diseño de las máquinas y que este también debe garantizar la ausencia de posibles riesgos. Por lo tanto, Torrance (2011) añade que puede ser definida como el adecuado diseño de máquinas que realicen acciones que, cuando son llevadas a cabo por humanos, son indicadores de la posesión de aptitudes éticas en esos humanos. Así pues, la ética de las máquinas inteligentes se relaciona con el diseño de máquinas inteligentes que demuestren implicaciones éticas en los resultados de su comportamiento, y con asegurarnos de que dicho comportamiento y diseño sean aceptables y no reprochables.

Pero esta definición excluye algunos riesgos mencionados y que nos llevan a reflexionar sobre las consecuencias éticas a largo plazo de la inteligencia artificial. No podemos ignorar la posible desigualdad o sustitución de la raza humana en algunos campos laborales de la sociedad. Para no dejar de lado ningún aspecto en la relación entre la humanidad y la inteligencia artificial, algunos investigadores la han concebido como el análisis de la naturaleza y el impacto social de las máquinas inteligentes, así como la formulación y justificación correspondiente de políticas para el uso ético de estas tecnologías (Moor, 1985, p. 268). En palabras de Siau & Wang (2020), la ética inteligente se refiere al comportamiento moral de los humanos al diseñar, construir, usar e interactuar con la IA, así como los impactos asociados con estas tecnologías en la humanidad y la sociedad a corto, mediano y largo plazo (p. 76).

Por lo tanto, lo que en esencia distingue a un problema básico de ética de las máquinas inteligentes es que involucra a una inteligencia artificial en una situación en la que existe cierta incertidumbre sobre qué hacer y cómo entender la situación (Moor, 1985, 266). No todas las situaciones que involucran tecnología y tienen importancia ética son relevantes para este ámbito, por supuesto. Para comprobar esto, consideremos el caso de un robo de 10 equipos inteligentes en una empresa de IBM. Si estas máquinas no representan ningún riesgo al ser robadas para ser utilizadas de manera malintencionada, se puede considerar que esta situación es más un robo que un dilema moral de la ética de las máquinas inteligentes. Sin embargo, si estas máquinas poseen información delicada que podría llevar al desarrollo de tecnologías perjudiciales para la humanidad, la situación adquiere relevancia ética para la ética de las máquinas inteligentes, ya que estos riesgos deben evaluarse y se debe rendir cuentas, pues pueden traer consecuencias graves para la humanidad y esto se tuvo que considerar antes de robarlas. Debemos preguntarnos, por lo tanto, cuándo una situación particular tiene importe ético en este campo de investigación. Además, a medida que aumenta la incertidumbre en el funcionamiento de los algoritmos de las máquinas inteligentes y se vuelve cada vez más difícil para los humanos comprender estos procesos, debemos cuestionarnos qué comportamientos esperamos de las máquinas las situaciones en las que actúan y pueden llegar a afectarnos (Anderson, 2011).

Por último, si en algún momento alcanzamos la anhelada singularidad, deberemos considerar con mayor énfasis la ética de las máquinas inteligentes. No sería sorprendente que una máquina más inteligente que nosotros también tenga una mayor relevancia ética, lo que incluso podría llevarnos a replantear nuestra ética y moral. El estudio de casos particulares

también nos permite investigar y comprender mejor las motivaciones y la esencia de la ética de las máquinas inteligentes. Los casos Trolley, presentados por la filósofa Philippa Foot en 1967, nos plantean situaciones en las que debemos elegir entre dos consecuencias moralmente aceptables pero contradictorias. Con la implementación de tecnologías inteligentes, la probabilidad de que surjan este tipo de dilemas en nuestra sociedad aumenta. De hecho, es más probable que se produzcan conflictos entre dos o más deberes, obligaciones o valores en situaciones en las que un agente ético tiene razones para elegir ambos sin poder hacerlo. Estas son situaciones casi irresolubles, ya que tomar acción en favor de uno de los valores implica un detrimento para el otro (Whittlestone, 2019, p. 13).

## 3.1 Ética de Máquinas estrechas

### 3.1.1 Algunos Dilemas

Imaginemos, por ejemplo, que logramos crear una inteligencia artificial capaz de salvar a miles de personas del cáncer al modificar los genes humanos de manera casi 100% efectiva. Sin embargo, al ser una máquina, existe un margen de error del 1%. En este 1%, en lugar de curar el cáncer, se produce una mutación genética que desarrolla una visión superior a la de los demás humanos. Al utilizar algoritmos de toma de decisiones, las máquinas inteligentes dependen de la estadística y el cálculo de probabilidades. Esto significa que siempre tendrán márgenes de error, incluso si el margen de error se acerca a cero, como en el ejemplo anterior. Una máquina con millones de usuarios podría terminar afectando incluso a miles de personas. Además, imaginemos en el ejemplo anterior que existe el riesgo de que esta tecnología caiga en manos malintencionadas, lo que daría lugar al posible surgimiento de ejércitos con supervisión. Estos ejércitos tendrían una ventaja sobre los demás debido a su visión mejorada, lo que les permitiría tener una ventaja táctica, como ver en la oscuridad, entre otras ventajas. A pesar de ser conscientes de estos riesgos, ¿es ético aceptar la implementación de estas tecnologías? Aunque podríamos salvar a millones de personas del cáncer, también podríamos dar lugar a una especiación, como mencionamos anteriormente. ¿Cómo determinamos qué es más importante? ¿Cuándo son los riesgos aceptables y cuándo no lo son?

Veamos ahora otro de estos dilemas, pero en el contexto de los automóviles autónomos. Imaginemos que un día, en la ciudad de Florida, donde ya podemos observar vehículos Tesla que se conducen solos, dos peatones cruzan abruptamente en rojo sin darse cuenta en un paso peatonal. Además, supongamos que el Tesla analiza todas las posibles opciones en una fracción de segundo y llega a la conclusión de que tiene dos alternativas: o bien subirse al andén y chocar contra un poste, causando daños a su pasajero, o atropellar a los dos peatones, ocasionándoles la muerte. La solución implementada es la de siempre resguardar la vida de los demás por encima de chocar un poste, subirse al andén o perjudicar al pasajero en el interior<sup>14</sup>. Teniendo esto en cuenta, podemos hacer dos afirmaciones al respecto. Primero, la construcción de vehículos inteligentes no considera el beneficio de quien adquiere el servicio. Para evitar problemas legales, se prefiere arriesgar la vida del pasajero y protegerse de intervenciones legales en el desarrollo de estas tecnologías. En segundo lugar, el dilema moral es inevitable. Aunque se disponga de los recursos, ¿persistiría el deseo de comprar un automóvil en el cual, ante el menor accidente posible, nuestra vida no valga nada? ¿Consideras que el lanzamiento de estas tecnologías se ha considerado esto? Aunque no creas que sea lo suficientemente preocupante, debemos considerar que estos vehículos no solo se

---

<sup>14</sup> <https://www.futurity.org/autonomous-vehicles-av-ethics-trolley-problem-2863992-2/>

implementan en todo el mundo, sino que están empezando a reemplazar a conductores de camiones, taxis y transporte público. Incluso si esto significa que miles de personas quedarían sin empleo, ¿es moralmente aceptable implementar este tipo de tecnologías sin más consideraciones?

Estos son algunos de los dilemas morales que surgen como resultado del desarrollo de la inteligencia artificial. Nos enfrentamos a nuevos desafíos nunca vistos, lo que nos toma por sorpresa en un mundo en el que es cada vez más difícil mantenernos actualizados con los últimos avances. En este sentido, ha surgido un interés por los casos particulares que pueden brindarnos perspectivas sobre la ética implicada en estas situaciones. Esta situación también resalta la importancia de enfocarnos en el estudio de nuestros propios valores, no solo para comprender mejor la ética humana y aplicarla en las tecnologías, sino también porque nuestros valores están cambiando bajo la influencia de estas nuevas tecnologías. No podemos simplemente esperar a que el mundo se ajuste a nuestros valores éticos y morales arcaicos anteriores al auge tecnológico. Debemos moldear nuestra moral y ética para adaptarnos a los cambios que implica la presencia de las nuevas tecnologías (Allen et al., 2011, p. 16).

### **3.1.2 Principios**

Como ya lo mencionamos, los seres humanos aplicamos razonamientos consecuencialistas y deontológicos en nuestros problemas prácticos sin mayor inconveniente. Aunque ambas teorías presenten sus propios problemas, comparten la cuestión central sobre la posibilidad de crear una máquina que se ajuste a alguno de estos razonamientos éticos, independientemente de cuál sea. De hecho, de manera similar a lo que ocurrió con el desarrollo de la conciencia, los científicos se han apresurado a incorporar nuestros principios morales y ética humana en la creación de IA como si fuera la única opción posible. Antes de apresurarnos en este camino de investigación, debemos cuestionar si la ética humana es la única posibilidad para que una máquina actúe de manera ética y si debemos limitarnos a este único criterio.

El conocimiento limitado que tenemos sobre ética en los seres humanos, junto con los dilemas morales y perjuicios generados por las máquinas inteligentes, nos obligan a dar mayor importancia a comprender cómo funciona la ética en los seres humanos mientras abordamos la ética de la IA (Storrs, 2011). Por lo tanto, muchos de los esfuerzos en ética de las máquinas inteligentes se centran en investigar la posibilidad de que las facultades mentales sean necesarias para que los seres humanos tengamos ética, y en cómo podríamos implementar esto en las máquinas. En el ámbito cognitivo, los investigadores intentan dotar a las máquinas de conciencia mientras descubren qué es la conciencia y cómo funciona. Paralelamente, los científicos están trabajando para lograr que las máquinas se comporten de manera ética, aunque permitimos su implementación sin saber realmente cómo funciona la ética en los seres humanos. Sin embargo, ¿por qué limitamos esta investigación a los límites humanos si las máquinas han demostrado afectarnos cognitivamente y éticamente a pesar de carecer de estas capacidades? Los humanos nos hemos dotado de un privilegio tal que llegamos a asumir que solo imitando nuestras capacidades podremos tener inteligencia, ética, entre otros aspectos. Pero en el fondo, no hay nada que nos lleve a esta conclusión. Las máquinas han demostrado ser más "inteligentes" que nosotros sin poseer conciencia. Del mismo modo, los robots pueden llegar a ser más "éticos" que nosotros sin contar con facultades humanas, como veremos a continuación.

De esta manera, enfocados exclusivamente en aptitudes éticas humanas, los investigadores que priorizan la ética en las máquinas inteligentes han intentado construir una máquina que:

- I) Posea una forma de describir situaciones y estados de cosas en el mundo.

- II) Pueda generar y ejecutar acciones en dichas situaciones y manipular los estados de cosas.
- III) Puede predecir las posibles consecuencias que podrían surgir al actuar de cierta manera en una situación determinada.
- IV) Cuento con un método para evaluar cada situación y sus posibles consecuencias en términos de deseabilidad o repulsión.

Con el fin de no dejar de lado la motivación principalista, algunos investigadores se centran en construir máquinas que contengan en su sistema o que logren conceptualizar principios morales. Para superar el relativismo ético, se han enfocado en los códigos morales que son comunes a todas las culturas. Esto es lo que Storrs (2011) denomina la "estructura moral profunda". Aunque no existe un consenso definitivo y todavía es un concepto abstracto, algunos principios que resuenan son:

- A) Reciprocidad: tanto en situaciones en las que se recibe algo bueno como en aquellas en las que se sufre un perjuicio.
- B) Jerarquía o estatus social.
- C) Universalidad de reglas morales básicas.
- D) Honestidad y confianza.
- E) Rechazo de la agresión no provocada.
- F) Noción de propiedad.
- G) Importancia relativa de algunas reglas sobre otras.
- H) Agencia moral y sus límites.

Estos principios pueden servir de guía para construir máquinas inteligentes con una base moral sólida y tomar decisiones éticas coherentemente. Sin embargo, es importante seguir investigando y refinando estos principios para garantizar que sean aplicables en diferentes contextos y situaciones.

Pero al intentar imitar la ética humana en el funcionamiento de máquinas inteligentes, nos encontramos con dos problemas principales. En primer lugar, no sabemos con certeza si esta es la mejor manera de construir ética en las máquinas inteligentes. No tenemos una respuesta clara sobre si debemos adoptar una teoría consecuencialista que valore más los casos específicos, una teoría principalista que se enfoque en los principios que motivan la acción, o una teoría híbrida que intente combinar ambas opciones. Además, dado que la humanidad tiene una historia limitada en el desarrollo de la IA, no hay nada que nos prevenga de cometer errores. Como vimos anteriormente, la necesidad de implementar leyes para regular el desarrollo de la IA se vuelve cada vez más evidente para evitar posibles perjuicios.

En segundo lugar, como nos muestra Allen (2005), uno de los mayores desafíos es el procesamiento del gran flujo de datos que las máquinas inteligentes deben actualizar diariamente para mantenerse al tanto de la situación social (p.152). Para poder adaptarse a las interacciones sociales y actuar en base a ellas, la inteligencia artificial debe recibir constantemente datos del mundo y actualizarse en tiempo real, de manera similar a cómo los seres humanos percibimos y procesamos información visual cuando tenemos los ojos abiertos. Exploraremos esta idea en mayor detalle más adelante. Por ahora, concluyamos que, aunque las máquinas inteligentes no necesiten las facultades humanas, podrían requerir mecanismos similares para garantizar un flujo constante de datos del mundo y asegurar su buen desempeño. Pese a los esfuerzos de los investigadores por mejorar la capacidad de adquisición de datos de las máquinas inteligentes, no debemos olvidar que las leyes y la relación de confianza entre la humanidad y las inteligencias artificiales aún no están

establecidas. A medida que los datos se almacenan de manera más precisa, la legislación no está haciendo lo suficiente para evitar los riesgos asociados a ello.

De acuerdo con Floridi (2012), este flujo constante de datos nos conducirá a experimentar un mundo completamente interactivo y receptivo, donde los datos estarán al alcance de todos y los procesos de información serán compartidos por todo. Actualmente, ya estamos presenciando esta extracción de datos. No es coincidencia que los anuncios de Google estén en línea con nuestro historial de búsqueda, ni que Facebook nos recomiende eventos basados en nuestros intereses y ubicación. Las tecnologías inteligentes, al extraer información de nosotros, ya tienen la capacidad de conocer dónde hemos estado, si estamos o hemos estado enfermos, nuestros planes de viaje y nuestras preferencias, entre otros aspectos.

### **3.1.3 Construcción de Maquinas Inteligentes**

Como ya lo mencionamos, los seres humanos aplicamos razonamientos consecuencialistas y deontológicos en nuestros problemas prácticos sin mayor inconveniente. Aunque ambas teorías presenten sus propios problemas, comparten la cuestión central sobre la posibilidad de crear una máquina que se ajuste a alguno de estos razonamientos éticos, independientemente de cuál sea. De hecho, de manera similar a lo que ocurrió con el desarrollo de la conciencia, los científicos se han apresurado a incorporar nuestros principios morales y ética humana en la creación de IA como si fuera la única opción posible. Antes de apresurarnos en este camino de investigación, debemos cuestionar si la ética humana es la única posibilidad para que una máquina actúe de manera ética y si debemos limitarnos a este único criterio.

El conocimiento limitado que tenemos sobre ética en los seres humanos, junto con los dilemas morales y perjuicios generados por las máquinas inteligentes, nos obligan a dar mayor importancia a comprender cómo funciona la ética en los seres humanos mientras abordamos la ética de la IA (Storrs, 2011). Por lo tanto, muchos de los esfuerzos en ética de las máquinas inteligentes se centran en investigar la posibilidad de que las facultades mentales sean necesarias para que los seres humanos tengamos ética, y en cómo podríamos implementar esto en las máquinas. En el ámbito cognitivo, los investigadores intentan dotar a las máquinas de conciencia mientras descubren qué es la conciencia y cómo funciona. Paralelamente, los científicos están trabajando para lograr que las máquinas se comporten de manera ética, aunque permitimos su implementación sin saber realmente cómo funciona la ética en los seres humanos. Sin embargo, ¿por qué limitamos esta investigación a los límites humanos si las máquinas han demostrado afectarnos cognitivamente y éticamente a pesar de carecer de estas capacidades? Los humanos nos hemos dotado de un privilegio tal que llegamos a asumir que solo imitando nuestras capacidades podremos tener inteligencia, ética, entre otros aspectos. Pero en el fondo, no hay nada que nos lleve a esta conclusión. Las máquinas han demostrado ser más "inteligentes" que nosotros sin poseer conciencia. Del mismo modo, los robots pueden llegar a ser más "éticos" que nosotros sin contar con facultades humanas, como veremos a continuación.

De esta manera, enfocados exclusivamente en aptitudes éticas humanas, los investigadores que priorizan la ética en las máquinas inteligentes han intentado construir una máquina que:

- I) Posea una forma de describir situaciones y estados de cosas en el mundo.
- II) Pueda generar y ejecutar acciones en dichas situaciones y manipular los estados de cosas.
- III) Puede predecir las posibles consecuencias que podrían surgir al actuar de cierta manera en una situación determinada.
- IV) Cuento con un método para evaluar cada situación y sus posibles consecuencias en términos de deseabilidad o repulsión.

Con el fin de no dejar de lado la motivación principalista, algunos investigadores se centran en construir máquinas que contengan en su sistema o que logren conceptualizar principios morales. Para superar el relativismo ético, se han enfocado en los códigos morales que son comunes a todas las culturas. Esto es lo que Storrs (2011) denomina la "estructura moral profunda". Aunque no existe un consenso definitivo y todavía es un concepto abstracto, algunos principios que resuenan son:

- A) Reciprocidad: tanto en situaciones en las que se recibe algo bueno como en aquellas en las que se sufre un perjuicio.
- B) Jerarquía o estatus social.
- C) Universalidad de reglas morales básicas.
- D) Honestidad y confianza.
- E) Rechazo de la agresión no provocada.
- F) Noción de propiedad.
- G) Importancia relativa de algunas reglas sobre otras.
- H) Agencia moral y sus límites.

Estos principios pueden servir de guía para construir máquinas inteligentes con una base moral sólida y tomar decisiones éticas coherentemente. Sin embargo, es importante seguir investigando y refinando estos principios para garantizar que sean aplicables en diferentes contextos y situaciones.

Pero al intentar imitar la ética humana en el funcionamiento de máquinas inteligentes, nos encontramos con dos problemas principales. En primer lugar, no sabemos con certeza si esta es la mejor manera de construir ética en las máquinas inteligentes. No tenemos una respuesta clara sobre si debemos adoptar una teoría consecuencialista que valore más los casos específicos, una teoría principalista que se enfoque en los principios que motivan la acción, o una teoría híbrida que intente combinar ambas opciones. Además, dado que la humanidad tiene una historia limitada en el desarrollo de la IA, no hay nada que nos prevenga de cometer errores. Como vimos anteriormente, la necesidad de implementar leyes para regular el desarrollo de la IA se vuelve cada vez más evidente para evitar posibles perjuicios.

En segundo lugar, como nos muestra Allen (2005), uno de los mayores desafíos es el procesamiento del gran flujo de datos que las máquinas inteligentes deben actualizar diariamente para mantenerse al tanto de la situación social (p.152). Para poder adaptarse a las interacciones sociales y actuar en base a ellas, la inteligencia artificial debe recibir constantemente datos del mundo y actualizarse en tiempo real, de manera similar a cómo los seres humanos percibimos y procesamos información visual cuando tenemos los ojos abiertos. Exploraremos esta idea en mayor detalle más adelante. Por ahora, concluyamos que, aunque las máquinas inteligentes no necesiten las facultades humanas, podrían requerir mecanismos similares para garantizar un flujo constante de datos del mundo y asegurar su buen desempeño. Pese a los esfuerzos de los investigadores por mejorar la capacidad de adquisición de datos de las máquinas inteligentes, no debemos olvidar que las leyes y la relación de confianza entre la humanidad y las inteligencias artificiales aún no están establecidas. A medida que los datos se almacenan de manera más precisa, la legislación no está haciendo lo suficiente para evitar los riesgos asociados a ello.

De acuerdo con Floridi (2012), este flujo constante de datos nos conducirá a experimentar un mundo completamente interactivo y receptivo, donde los datos estarán al alcance de todos y los procesos de información serán compartidos por todo. Actualmente, ya estamos presenciando esta extracción de datos. No es coincidencia que los anuncios de Google estén

en línea con nuestro historial de búsqueda, ni que Facebook nos recomiende eventos basados en nuestros intereses y ubicación. Las tecnologías inteligentes, al extraer información de nosotros, ya tienen la capacidad de conocer dónde hemos estado, si estamos o hemos estado enfermos, nuestros planes de viaje y nuestras preferencias, entre otros aspectos.

### 3.1.4 Contexto Investigativo

De acuerdo con Jobin et al. (2019), se ha llevado a cabo una cantidad limitada de investigaciones en el campo de la ética de las máquinas inteligentes, así como en los principios y directrices éticas para su desarrollo, y en la necesidad de implementar leyes que eviten consecuencias excesivas y perjudiciales. En la búsqueda de literatura relacionada con este tema, se encontraron 84 artículos. La mayoría de estos artículos fueron escritos por compañías privadas (22,6% = 19) y agencias gubernamentales (21,4% = 18). Les siguieron las instituciones académicas e investigativas (10,7% = 9), organizaciones intra gubernamentales o supranacionales (9,5% = 8), organizaciones sin ánimo de lucro (8,3% = 7), asociaciones profesionales o sociedades científicas (8,3% = 7), alianzas del sector privado (4,8% = 4), alianzas de investigación (1,2% = 1), fundaciones científicas (1,2% = 1), federaciones de trabajo (1,2% = 1) y partidos políticos (1,2% = 1). Algunos artículos pertenecían a más de una categoría (4,8% = 4), mientras que otros no eran fáciles de clasificar (4,8% = 4).

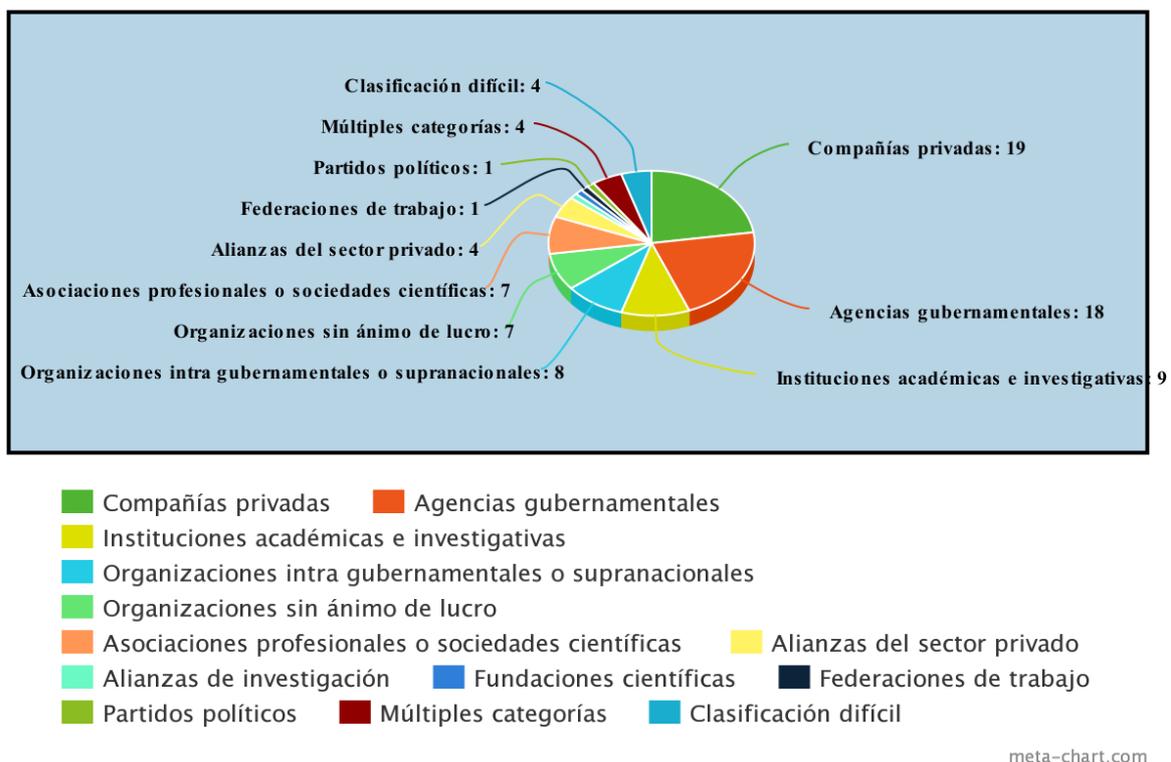


Figura 5: distribución del trabajo investigativo en ética de máquinas inteligentes (p. 6).

- La distribución geográfica de estos 84 trabajos de investigación revela una marcada desigualdad en la participación en este campo, ya que la mayoría de los actores son países económicamente desarrollados del primer mundo. Estados Unidos (23,8% = 20) y Reino Unido (16,7% = 14) representan aproximadamente un tercio de los datos. A estas potencias les siguen Japón (4,8% = 4), Alemania (3,6% = 3), Francia (3,6% = 3) y Finlandia (3,6% = 3). En total, hay 6 fuentes (%) provenientes de instituciones

pertenecientes a la Unión Europea y 16 fuentes (%) provenientes de alianzas entre miembros de la Unión Europea. Sin embargo, África y América del Sur no están representadas, excepto por su participación en alianzas internacionales o supranacionales.

- Estados Unidos: 20 (23.8%)
- Reino Unido: 14 (16.7%)
- Alianzas entre miembros de la Unión Europea: 13 (15.5%)
- Instituciones pertenecientes a la Unión Europea: 6 (7.1%)
- Artículos con participación africana y sudamericana: 6 (7.1%)
- Japón: 4 (4.8%)
- Alemania: 3 (3.6%)
- Francia: 3 (3.6%)
- Finlandia: 3 (3.6%)
- Países Bajos: 2 (2.4%)
- España: 1 (1.2%)
- Noruega: 1 (1.2%)
- Corea del Sur: 1 (1.2%)
- Italia: 1 (1.2%)
- India: 1 (1.2%)
- Islandia: 1 (1.2%)
- Canadá: 1 (1.2%)
- Australia: 1 (1.2%)
- Artículos sin reportar una nación específica: 3 (3.6%)

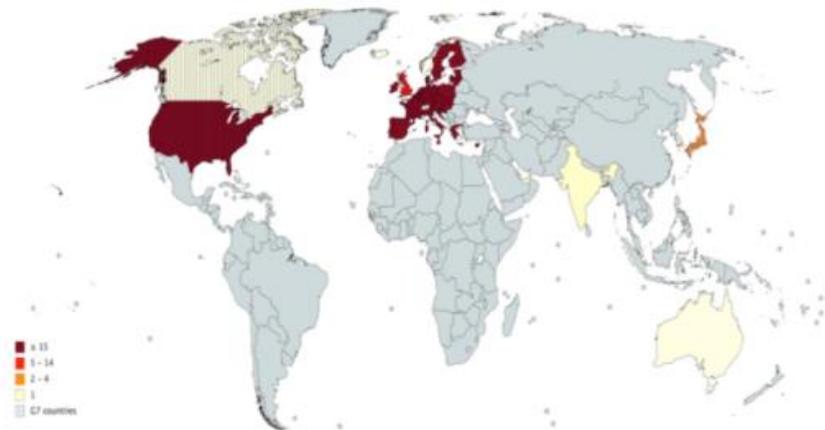


Figura 6: Distribución geográfica de la IA

En relación a los principios y directrices éticas, Jobin et al. destacan una convergencia en torno a los siguientes principios: transparencia, justicia y equidad, no maleficencia, responsabilidad y privacidad. Luego analizaremos estos principios para identificar los posibles dilemas morales que puedan surgir a partir de ellos y explicar por qué son éticamente importantes en nuestro contexto social actual.

### 3.1.4.1 Transparencia

El principio de transparencia es el más prevalente en la literatura actual. Se enfoca en aumentar la explicabilidad e interpretabilidad del funcionamiento de las máquinas con el objetivo de prevenir daños y mejorar la inteligencia artificial (IA). Además, no se limita a esto, sino que también aboga por ser abierto con el uso de datos, sus potenciales riesgos y su contexto de uso. Por ejemplo, en el caso anterior del auto tesla que no requiere conductor, la empresa es responsable de alertar a sus usuarios acerca de estos posibles riesgos y, además, de explicar cómo funciona el algoritmo para que esto se tenga en cuenta a la hora de implementar uno de estos coches. Si los usuarios y los gobiernos comprendieran cómo funcionan estos algoritmos mediante una explicación clara, no confiaríamos ciegamente en ellos. Esto fomentaría la implementación de leyes que promuevan la transparencia y regulen la utilización de tecnologías con algoritmos que podrían perjudicarnos o incluso con datos desactualizados que podrían perjudicarnos. Aunque no existe consenso sobre cómo describir este principio, todas sus menciones destacan algo novedoso en el desarrollo de la IA, ya que podríamos enfrentar la falta de transparencia debido a nuestras limitaciones humanas. Como mencionan Floridi et al. (2018), los avances y funcionamientos de las máquinas inteligentes son invisibles para la mayoría de las personas e incluso para todos en algunos casos. Avances como el Aprendizaje Profundo (Deep Learning) han demostrado ser muy complejos incluso para la comprensión humana. Por ejemplo, en 2014,<sup>15</sup> Facebook apagó uno de sus sistemas de IA luego de que creara un lenguaje que no podíamos entender. Surge la pregunta de si debemos continuar con este desarrollo o, al no poder comprenderlo ni explicarlo, deberíamos abandonarlo. La falta de comprensión sobre cómo funciona un algoritmo puede hacer que, sin que nos demos cuenta, este actúe de manera injusta y desigual.

#### **3.1.4.2 Justicia**

El segundo principio es el de justicia y equidad. Según Harari (2018), la desigualdad existe desde la Edad de Piedra, donde se observaba cierta división y estatus social en los rituales funerarios. Algunos eran enterrados con lujos mientras que otros tenían un entierro más simple. Sin embargo, experimentos como el mencionado por Frans de Waal, también citado por Harari (2018), desafían esta noción y sugieren que principios como la igualdad pueden ser naturales, biológicos e intemporales. Esto nos lleva a cuestionar la igualdad y la desigualdad en el contexto de los avances en IA. Por ejemplo, en un futuro cuando sean más las mejoras que los medicamentos, los beneficios podrían ser para unos pocos. ¿Debemos implementar estos beneficios, aunque solo un 10% de la población pueda aprovecharlos? Lo mismo ocurre con las empresas tecnológicas, que cada vez requieren menos empleados humanos, beneficiando a menos personas cada año. ¿Deberían los impuestos depender del número de empleados humanos frente a los trabajadores artificiales? ¿Deben estas empresas garantizar la contratación de un número mínimo de empleados humanos? Además, deberíamos tener más personas involucradas en el desarrollo de estas tecnologías, no exclusivamente a tecnólogos, ingenieros y matemáticos. Sin la participación de un campo de trabajo más amplio, es imposible tener en cuenta todos los posibles rasgos y los posibles perjuicios. En el caso de los datos, no fue hasta que se implementaron los algoritmos que se empezó a conocer que el uso de datos puede llevar a la discriminación y sesgos contra ciertos grupos si no se seleccionan adecuadamente. Por ejemplo, si utilizamos IA para ayudarnos en la selección de empleados, debemos proporcionarle datos actualizados y evitar que tome en cuenta sesgos del pasado que afectaban la participación de las mujeres en empresas y otros

---

<sup>15</sup> <https://www.forbes.com/sites/tonybradley/2017/07/31/facebook-ai-creates-its-own-language-in-creepy-preview-of-our-potential-future/>

trabajos. Es crucial tener cuidado con los datos que utilizamos, qué datos utilizamos y cómo los utilizamos.

### **3.1.4.3 Beneficencia**

El problema se agrava cuando se percibe que los avances con tecnologías inteligentes suelen usarse malintencionadamente. Un ejemplo notable es el caso de Meta (antes Facebook), donde robaron 553 millones de usuarios e información personal, y luego se vendieron en un foro de hackers con fines maliciosos. Este incidente ilustra cómo el progreso tecnológico sin una consideración adecuada de las implicaciones éticas puede dar lugar a un uso abusivo de dichas tecnologías<sup>16</sup>.

Si continuamos persiguiendo el avance tecnológico sin prestar atención a las implicaciones éticas, corremos el riesgo de enfrentar consecuencias no deseadas e incluso dañinas. Aunque es cierto que puede haber casos de riesgos no intencionales que no pueden evitarse y solo pueden abordarse con un análisis detallado de los casos, cualquier riesgo previsible debe evitarse. Es fundamental garantizar la buena fe y el bienestar de los seres humanos en el desarrollo e implementación de estas tecnologías, para evitar el uso malintencionado y proteger a las personas de posibles perjuicios.

### **3.1.4.4 Responsabilidad**

Cuando consideramos los casos de robo de información en Meta, la segregación de mujeres en las entrevistas de trabajo por parte de la IA, los errores en los sistemas de GPS que pueden poner en peligro a las personas, la sustitución de empleos y otros dilemas éticos mencionados anteriormente, surge la pregunta de quién o qué es responsable de estas consecuencias indeseadas. Debemos cuestionar quién o qué originó estos problemas. Además, nos enfrentamos a un nuevo dilema moral conocido como el "problema de las muchas manos" en la literatura reciente (Siau y Wang, 2020, p.81). Cuando una máquina nos perjudica en nuestras interacciones con ella, la causa puede ser atribuida a diversas opciones, como los códigos de programación, los datos utilizados o los usos malintencionados, entre otros. Es responsabilidad del campo de la IA brindar claridad sobre quién debe asumir la responsabilidad en cada caso. Los desarrolladores, diseñadores, instituciones e incluso las personas comunes podrían ser potencialmente responsables.

Considerando que los desarrolladores y diseñadores priorizaron el avance tecnológico sin dar suficiente importancia a otros aspectos, ante el problema de la falta de claridad sobre el grado de incertidumbre aceptable en el desarrollo de máquinas inteligentes. Mientras que Facebook apagó su inteligencia artificial cuando creó un lenguaje incomprensible para nosotros, existen muchas máquinas que realizan cálculos tan complejos que desconocemos su funcionamiento desde hace años<sup>17</sup>. Muchos diseñadores pueden aceptar la falta de comprensión de sus desarrollos con tal de lograr avances en el campo. Es importante tener en cuenta que no existe una relación de confianza establecida en este campo y que algunos podrían evadir las leyes y la responsabilidad en busca de beneficios y avances.

Concuerdo con Wemitz (2018) en que el caso del juicio de Google España es uno de los ejemplos más destacados en la actualidad en torno a la responsabilidad y el manejo de datos. En 1998, un periódico español publicó anuncios de casas hipotecadas por deudas con la

---

<sup>16</sup> <https://www.theverge.com/2022/11/28/23481786/meta-fine-facebook-data-leak-ireland-dpc-gdpr>

<sup>17</sup> <https://www.wired.com/story/our-machines-now-have-knowledge-well-never-understand/>

seguridad social, con el objetivo de vender estas propiedades. Uno de estos anuncios involucraba a Mario Costeja, quien en noviembre de 2009 solicitó al periódico que eliminara esa información, ya que habían transcurrido 11 años desde el suceso y, aunque ya no tenía una hipoteca, si introducía su nombre en el buscador el primer resultado era este. El periódico se negó a eliminar el anuncio, por lo que el señor Costeja llevó su queja a Google España. Sin embargo, le informaron que el área responsable de estos asuntos estaba fuera de la Unión Europea y que debían remitir el caso a las oficinas centrales de Google en California. Sin embargo, Google California se negó a eliminar el anuncio argumentando que no se trataba de información estadounidense que estuviera bajo su responsabilidad. El señor Costeja continuó lidiando con esta burocracia y presentó su queja ante la Agencia Española de Protección de Datos, solo para descubrir que había más de 200 quejas similares a las que Google simplemente respondía diciendo que no podían hacer nada porque la entidad encargada del motor de búsqueda de datos no se encontraba en la Unión Europea y no les correspondía intervenir. Este ejemplo destaca la necesidad de responsabilizar a aquellos que permiten la ocurrencia de perjuicios y de investigar las razones detrás de estos incidentes. Para lograr un trato justo, es necesario contar con transparencia y una responsabilidad adecuada que garantice la equidad y la justicia, y evite el uso malintencionado de estas tecnologías. Sin embargo, casos como el de Google España demuestran que, si no abordamos estos problemas a través de políticas gubernamentales que regulen y limiten estas prácticas, prevalecerá la evasión de la ley y se seguirá priorizando exclusivamente el desarrollo.

Este ejemplo casi que Kafkiano resalta la necesidad de transparencia en los conjuntos de datos de dominio público y mantener actualizada esta información para prevenir la desinformación y otros problemas. ¿Si proporcioné mis datos hace veinte años, estarán disponibles para siempre? Es crucial regular este tipo de cuestiones para evitar la propagación de información falsa y otros inconvenientes. ¿Hizo Google lo correcto al negarse a eliminar un dato que afectaba a alguien y que expresó esto? Además, ¿fue apropiado evadir la responsabilidad argumentando que estaban fuera de la Unión Europea geográficamente? ¿Es posible que las empresas más poderosas actúen de buena fe sin que existan regulaciones? Por otro lado, ¿si un estadounidense hubiera solicitado la eliminación de uno de estos anuncios, se le habría dado mayor importancia porque la entidad responsable está en ese país? En caso afirmativo, ¿no fomentaría esto la desigualdad y un trato preferencial hacia los ciudadanos estadounidenses? Estas interrogantes subrayan la importancia de abordar estas problemáticas de manera justa y equitativa para evitar la desigualdad y asegurar que todos los individuos sean tratados de forma igualitaria, independientemente de su nacionalidad. Vemos, además, de forma incipiente, una idea que reforzaremos más adelante: el desarrollo de principios para la ética de máquinas inteligentes requiere pensar en los principios de manera holística. Por lo general, si uno de ellos está siendo vulnerado, todos lo están.

### **3.1.4.5 Privacidad**

Por lo anterior, en el contexto de los datos, se ha generado un creciente interés en la distinción entre datos públicos y privados. Los datos abiertos (Open Data) son aquellos que están disponibles al público para su visualización, uso, transmisión y republicación, según Andersen (2021, p. 9). Por otro lado, la información protegida se refiere a datos que involucran comunicaciones personales o información íntima y privada. Esta información no es fácilmente accesible y requiere protección, ya que revelarla podría comprometer la identidad, comportamiento, asociaciones, condiciones médicas, raza, orientación sexual, orígenes, puntos de vista, lugares frecuentados y otros aspectos privados de un individuo.

El uso de máquinas inteligentes, como señala Harari (2018), podría permitir una detección más rápida de epidemias, reducción de la contaminación, gestión de residuos y otros

beneficios. Sin embargo, esto solo sería posible si se permite el acceso libre a la información que producimos. El uso de datos personales puede mejorar la calidad de vida, pero también preocupa la privacidad por la creciente demanda de datos. Muchos usuarios de plataformas como Facebook y Google se sienten preocupados por las prácticas de recolección de datos que estas compañías llevan a cabo sin un consentimiento plenamente informado. En esta era en la que los datos tienen un valor cada vez mayor tanto en términos monetarios como tecnológicos, es crucial proteger la privacidad. Se deben fomentar y aplicar políticas que regulen el uso protegido de datos y excluyan el acceso a información privada en todas las ciudades del mundo.

### **3.1.4.6 Reconsiderando la Felicidad**

Jabin (2019) resalta que son escasos los estudios de investigación que mencionan el objetivo original del desarrollo de máquinas inteligentes: promover el bienestar humano, su florecimiento, paz y felicidad, así como la prosperidad económica y la creación de oportunidades socioeconómicas. No podemos olvidar que esta es la razón primordial detrás de las máquinas inteligentes. Sin tener este propósito en mente es probable que las máquinas nos causen más perjuicio que bienestar. Floridi et al. (2018) identifican cuatro aspectos fundamentales para comprender el florecimiento humano y el desarrollo de la personalidad:

- i) Quiénes podemos llegar a ser (autorrealización autónoma).
- ii) Lo que podemos hacer (agencia).
- iii) Lo que podemos lograr (capacidades independientes y sociales).
- iv) Cómo podemos interactuar entre nosotros (cohesión social).

Estos puntos resaltados por Floridi nos indican que el florecimiento humano y el desarrollo de la personalidad no se enfocan únicamente en el individuo, sino también en nuestras interacciones sociales y capacidades colectivas. En el contexto de las máquinas inteligentes, es esencial considerar cómo estas tecnologías pueden contribuir a cada uno de estos aspectos, en vez de perjudicar, pues hemos visto infinidad de casos en las que esto sucede. Para promover el florecimiento humano, es necesario asegurar que las máquinas inteligentes permitan a las personas alcanzar su máximo potencial y autorrealización autónoma. Esto implica diseñar sistemas que fomenten la autonomía, la creatividad y la libertad individual, permitiendo a las personas tomar decisiones informadas y tener un control significativo sobre su interacción con las máquinas en la sociedad.

Además, las máquinas inteligentes deben ser herramientas que potencien nuestra agencia, brindándonos la capacidad de tomar acciones efectivas y tener un impacto positivo en nuestras vidas y comunidades. Esto implica que las tecnologías no sean meramente pasivas, sino que nos empoderen para alcanzar metas personales y colectivas, promoviendo la participación activa y la autodeterminación. Asimismo, el desarrollo de las máquinas inteligentes debe considerar la importancia de las capacidades independientes y sociales. Estas tecnologías deben estar orientadas a ampliar nuestras habilidades y oportunidades, tanto a nivel individual como en el contexto de la comunidad. Esto implica facilitar el acceso a la educación, el empleo, la atención médica, la justicia y otros recursos esenciales para un desarrollo humano integral.

Por último, la cohesión social es un aspecto crucial a tener en cuenta. Las máquinas inteligentes deben promover la colaboración, la inclusión y la igualdad, evitando la polarización y la exclusión social. Es necesario fomentar sistemas tecnológicos que fortalezcan los lazos comunitarios, la diversidad y la comprensión mutua, promoviendo así sociedades más cohesionadas y equitativas. En resumen, el desarrollo de máquinas

inteligentes debe ir de la mano con el objetivo de promover el florecimiento humano, la paz y la felicidad, la prosperidad económica y la creación de oportunidades socioeconómicas, asegurando así un impacto positivo y sostenible en nuestras vidas y en la sociedad en su conjunto.

Cuando hablamos de valores humanos, no nos referimos únicamente a deseos y placeres, sino a compromisos arraigados y compartidos en nuestras creencias. Son los ideales que motivan la forma en que organizamos la sociedad (Whittlestone, 2019, p. 25). El que estas tecnologías generen conflictos con algunos de nuestros valores nos da motivos suficientes para reflexionar sobre esta relación desde una perspectiva ética que recupere el objetivo original: pensar en las tecnologías inteligentes según el bienestar humano y la felicidad. Es fundamental recordar que somos responsables de dar forma a estas tecnologías de manera ética, asegurando que estén alineadas con nuestros valores fundamentales y que no comprometan nuestra humanidad. Solo a través de un enfoque reflexivo y consciente podemos garantizar que estas innovaciones tecnológicas contribuyan realmente a un futuro sostenible y beneficioso para todos.

Sin embargo, es importante reconocer que la búsqueda de la felicidad plantea sus propios desafíos, tanto dentro como fuera del contexto de las tecnologías inteligentes. Según Harari (2018), la felicidad humana tiene bases psicológicas y biológicas. A nivel psicológico, la felicidad depende de nuestras expectativas más que de las condiciones objetivas. El problema radica en que a medida que las condiciones mejoran, nuestras expectativas también se elevan, lo que dificulta alcanzar la plenitud. La mente humana tiende a buscar constantemente más en lugar de sentirse satisfecha (Harari, 2017). Esta búsqueda constante nos lleva a valorar más la gratificación inmediata que la satisfacción a largo plazo. Si esto no fuera así, cambiaríamos nuestros estilos de vida en relación al calentamiento global y no buscaríamos soluciones temporales. Si no fuera así, la inteligencia artificial habría sido legalizada hace mucho tiempo.

En este sentido, es importante reflexionar sobre cómo las tecnologías inteligentes pueden influir en nuestra búsqueda de la felicidad. Si nos enfocamos únicamente en el desarrollo tecnológico sin considerar las consecuencias a largo plazo y sin abordar las cuestiones éticas, es posible que nos alejemos aún más de la felicidad genuina. Debemos ser conscientes de los impulsos y deseos que guían nuestra relación con la tecnología, y asegurarnos de que estén alineados con nuestro bienestar y el de los demás. La felicidad duradera no se logrará solo con avances tecnológicos, sino con un equilibrio entre nuestras necesidades y expectativas, nuestras interacciones sociales y nuestro sentido de propósito y significado en la vida.

Sin embargo, es importante reconocer que esto puede ser inherentemente difícil de remediar, ya que nuestro sustrato biológico parece indicar que esto es resultado de nuestra evolución. A medida que hemos evolucionado hasta convertirnos en lo que somos hoy en día, nuestro sistema bioquímico se ha adaptado cada vez más para recompensar la supervivencia y la reproducción, en lugar de enfocarse en la satisfacción personal (49). Nuestra felicidad está en gran parte determinada por la bioquímica y los equilibrios sensoriales causados por estímulos en nuestro cuerpo. No obstante, es importante tener en cuenta que, aunque la felicidad pueda estar influenciada por nuestra biología y nuestras expectativas, aún podemos encontrar un equilibrio y buscar la satisfacción en nuestras vidas. Aunque puede ser difícil escapar de nuestros instintos evolutivos, también tenemos la capacidad de desarrollar habilidades y estrategias para cultivar una mayor sensación de bienestar y felicidad. Al comprender cómo funcionan nuestros sistemas bioquímicos y psicológicos, podemos tomar decisiones más conscientes y buscar formas de alcanzar una felicidad duradera, independientemente de las circunstancias externas. Es un proceso que requiere autoconciencia, autodisciplina y práctica

constante, pero es posible encontrar un mayor sentido de satisfacción y bienestar en nuestras vidas.

### 3.1.4.7 Sesgos de Datos

Además de los desafíos inherentes a la búsqueda de la felicidad, otro aspecto relevante a considerar en las máquinas inteligentes es el sesgo de datos (data bias) si no logramos distinguir adecuadamente entre los datos privados y los públicos. En los cinco conceptos que ya mencionamos en la investigación de la ética de las máquinas inteligentes, se observa una falta de claridad y delimitación en este concepto. Cada vez es más necesario establecer regulaciones para manejar de manera apropiada los datos y prevenir su uso malicioso (Siau & Wang, 2020, 80). La transparencia y la responsabilidad se vuelven cada vez más difíciles de alcanzar, por lo que también es crucial pensar en ellas en relación con los datos: qué datos debemos almacenar, quién o qué debe encargarse de dicho almacenamiento y quién debe tener acceso a ellos. La realidad es que las máquinas inteligentes son tan buenas como el conjunto de datos que se les introduce (Siau & Wang, 2020, p. 81). La tecnología inteligente está siendo desarrollada y diseñada por humanos, y es común encontrar sesgos implícitos en ellos. Por lo tanto, no debería sorprendernos que las máquinas que creamos hereden estos sesgos. Es fundamental abordar este problema y trabajar en soluciones que minimicen los sesgos y promuevan la equidad y la imparcialidad en el funcionamiento de las máquinas inteligentes.

Por ejemplo, en el año 2015 se identificó un problema con el reconocimiento facial de Google, donde se clasificaba incorrectamente a las personas negras como gorilas<sup>18</sup>. Hasta el día de hoy, no sabemos con certeza si Google no pudo resolver el problema, si decidió no asignar recursos para solucionarlo o si está siendo excesivamente cauteloso. Sin embargo, esta situación destaca la necesidad de responsabilizar a las personas involucradas en estos errores, ya que, de lo contrario, Google y otras empresas tecnológicas continuarán eludiendo las leyes. Lo que sí sabemos es que la compañía eliminó el algoritmo de reconocimiento de imágenes relacionado con gorilas para evitar más clasificaciones erróneas. Esto plantea interrogantes sobre la comprensión total de Google sobre cómo funciona este algoritmo y sobre la transparencia en su funcionamiento. Además, es preocupante que, ocho años después, la situación no haya cambiado<sup>19</sup>. Los lectores pueden hacer el intento y utilizar el algoritmo para identificar rápidamente animales como gatos, caballos, tigres e incluso monos capuchinos o monos araña. Sin embargo, esto no es posible al buscar gorilas o chimpancés, lo que indica una brecha en esta categoría. Nos preguntamos sinceramente cómo es posible que Google permita que estos errores pasen desapercibidos. Aunque los trabajadores de Google han señalado que el problema radica en la falta de representación de datos de entrenamiento que incluyan a personas negras, la empresa nunca ha proporcionado una explicación formal. Si ese es el problema, ¿Google está actuando de buena fe al no solucionarlo simplemente agregando nuevos datos? Es necesario abordar este tema de manera transparente y responsabilizar a las empresas tecnológicas por los sesgos y errores en sus sistemas de reconocimiento facial para evitar perjuicios y discriminación.

En el planeta Tierra, aproximadamente entre el 15% y el 16% de la población total es de ascendencia negra. El hecho de que el sistema de Google haya favorecido al resto de la población mientras segregaba e incluso evocaba dolorosos recuerdos de la época de la

---

<sup>18</sup> <https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai>

<sup>19</sup> <https://www.nytimes.com/2023/05/22/technology/ai-photo-labels-google-apple.html>

esclavitud para las personas negras, indica que los datos utilizados estaban sesgados y trataban de forma racista a esta parte de la población. La máquina heredó estos sesgos a través de los datos proporcionados. Por tanto, cualquier sistema de inteligencia artificial que discrimine sistemática e injustamente a ciertos individuos o grupos, favoreciendo a otros, muestra un comportamiento sesgado que beneficia a una parte de la población y a otros grupos según ciertos rasgos (Friedman & Nissenbaum, 1986, p. 343) y 39). Estos rasgos pueden ser la raza, como ya hemos visto, o el género, como en el ejemplo de la discriminación de las mujeres en entrevistas de trabajo, entre muchas otras categorías en las que clasificamos a los seres humanos. Además, es importante destacar que los sesgos deben ser sistemáticos, lo cual excluye los accidentes y nos obliga a responsabilizar siempre a alguien o algo en situaciones como estas.

En el ejemplo anterior, nos encontramos con un sesgo de datos incompletos. Cuando los datos están incompletos, incorrectos o desactualizados, las decisiones tomadas por la máquina se basarán en ellos y, por lo tanto, heredarán los sesgos y vacíos presentes en los datos de entrenamiento. Otro tipo de sesgo es el histórico, que se refiere a cuando un conjunto de datos hereda un sesgo que existió en el pasado y que aún persiste en los conjuntos de datos en el presente. Por ejemplo, los algoritmos de análisis de riesgo son modelos estadísticos que clasifican a los delincuentes en función de la probabilidad de reincidencia en el crimen. Estos algoritmos ayudan a decidir qué presos pueden ser liberados y cuáles no, y las máquinas inteligentes han comenzado a tomar estas decisiones que impactan nuestras vidas en nuestro nombre. El problema radica en que algunos resultados muestran que esta fórmula perjudica a las personas negras y las clasifica incorrectamente casi el doble de veces que a los presos blancos<sup>20</sup>. No solo se estaba clasificando erróneamente a las personas negras como riesgo el doble de veces, sino que también se beneficiaba injustamente a los presos blancos con puntajes favorables otorgados erróneamente por las máquinas. Desde la década de 1970, se ha prohibido el uso de factores como la raza, la nacionalidad o el color de piel en los análisis de riesgo, ya que generaban segregación y falsos prejuicios de que las personas negras eran más propensas a ser criminales. Este ejemplo nos muestra que la máquina puede heredar este sesgo por datos incompletos o porque el diseñador del algoritmo cree en secreto en estos falsos prejuicios. Hasta que no obtengamos transparencia, no tendremos respuestas claras sobre estas situaciones, solo podremos hacer especulaciones.

Por último, es importante mencionar que los datos también pueden dejar de ser representativos y heredar prejuicios debido a que quienes construyeron y diseñaron la máquina seleccionaron los datos con sus propios sesgos (Andersen, 2021, p. 11). Supongamos que se utilizaron los resultados de los exámenes médicos de Angelina Jolie para posteriores desarrollos médicos. Sin embargo, estos desarrollos solo recopilarán datos de aquellos que pueden permitirse pagar ese tipo de exámenes y tratamientos médicos, lo que representa solo el 10% de la población. De esta manera, la inteligencia artificial utilizada puede llegar a conclusiones beneficiosas sobre avances en modificaciones genéticas, curas contra el cáncer u otros avances médicos, pero esto solo será representativo del 10% de la población que pudo participar en la investigación. En un mundo en el que la especiación es cada vez más posible, ese 10% puede diferir significativamente de muchos de nosotros en tan solo una o dos generaciones en las que esté presente la modificación genética.

Para abordar el desafío de los sesgos en las tecnologías inteligentes, es fundamental contar con un marco legal sólido que promueva la equidad y evite la discriminación. La implementación de leyes y regulaciones específicas puede desempeñar un papel crucial en la

---

<sup>20</sup> <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

prevención y mitigación de los sesgos al establecer estándares claros y responsabilidades para las empresas y desarrolladores de inteligencia artificial. Estas leyes pueden abordar aspectos como la transparencia en los algoritmos, la diversidad en los conjuntos de datos utilizados y la evaluación continua de los sistemas para identificar y corregir sesgos. Al establecer normativas efectivas, podemos garantizar que las tecnologías inteligentes se utilicen de manera ética y equitativa, evitando así la perpetuación de desigualdades y sesgos en nuestra sociedad. En lo que respecta a los datos (Siau & Wang, 2020, p. 80) (Andersen, 2021, p. 12), existen diversas consideraciones importantes:

- i) Capacitación adecuada para los operadores: Es necesario brindar una formación adecuada a aquellos que trabajan con sistemas de inteligencia artificial, para que comprendan cómo funciona la IA, los límites de su uso y los posibles riesgos asociados.
- ii) Transparencia y responsabilidad gubernamental: Los gobiernos deben promover la transparencia y la rendición de cuentas en el uso de la inteligencia artificial, interviniendo para proteger los derechos humanos y exigiendo explicaciones claras en los sistemas de IA.
- iii) Relación de confianza en las empresas privadas: Se debe fomentar una relación de confianza entre las empresas privadas y los usuarios de la inteligencia artificial, garantizando que se respeten y protejan los derechos humanos en todas las etapas de desarrollo e implementación.
- iv) Investigación continua: Es fundamental seguir investigando y comprendiendo mejor los sesgos en la inteligencia artificial. Estudios recientes (Caliskan et al., 2017) (Greenwald & Pettigrew, 2014) han demostrado que el favoritismo es suficiente para generar sesgos implícitos. Aún queda mucho por descubrir y entender para eliminar estos sesgos.

En resumen, la combinación de un marco legal sólido, una capacitación adecuada, transparencia gubernamental, una relación de confianza en las empresas y una investigación continua nos permitirá abordar de manera efectiva los sesgos en las tecnologías inteligentes y trabajar hacia un futuro más equitativo y libre de discriminación. Al implementar datos en las máquinas inteligentes, debemos prestar especial atención para no confundir correlación con causalidad. Por ejemplo, en el caso de los análisis de riesgo, el hecho de que haya muchas personas negras en la cárcel no implica que los negros sean más propensos a ser criminales. Es más probable que haya una mayoría de población negra en estado de pobreza extrema. Además, debemos tener cuidado de no incluir parámetros que sean consecuencia de sesgos conocidos. Aunque excluyamos la variable que podría introducir sesgo, como la raza en el ejemplo anterior, podríamos estar incluyendo otros factores que también estén correlacionados con la raza y perpetúen la discriminación, como el nivel educativo, el código postal, la ocupación laboral o el historial crediticio.

De igual manera, como señalan Friedman y Nissenbaum (1996), es importante tener en cuenta que el sesgo puede tener diferentes orígenes (p. 344). Puede ser preexistente, arraigado en instituciones sociales, prácticas y actividades que muestran sesgos y que la máquina hereda y replica. Este sesgo puede ser intencional o malintencionado. También puede ser técnico, debido a limitaciones o consideraciones técnicas. Por ejemplo, el algoritmo de Google puede tener dificultades para identificar rostros en imágenes de personas negras debido a una limitación técnica ocasionada por el conjunto de datos utilizado. Además, el sesgo puede ser emergente, surgiendo de forma inesperada en el contexto real de uso con los usuarios. Es importante prestar atención a los casos particulares que se generan a diario y considerar períodos de pruebas apropiados antes de implementar desarrollos tecnológicos.

Todo esto resalta la importancia de prestar atención a los datos y su correcta implementación para garantizar el uso adecuado y beneficioso de las máquinas inteligentes en relación con los seres humanos. Debemos considerar cuidadosamente tanto el origen de los sesgos como las limitaciones técnicas, y estar atentos a posibles sesgos emergentes en contextos reales de uso. Además, reflexionar sobre el momento oportuno para implementar un desarrollo tecnológico y realizar períodos de prueba adecuados. Al reconocer la importancia de abordar los sesgos en las tecnologías inteligentes, trabajaremos hacia un futuro más equitativo y libre de discriminación, donde estas herramientas se diseñan y utilizan de manera responsable en beneficio de la sociedad.

## 3.2 Ética de Máquinas Generales

El acelerado y despreocupado desarrollo e implementación de máquinas inteligentes, sin considerar una posible legislación que prevenga riesgos, sesgos y perjuicios, puede estar motivado por la carrera tecnológica en la que nos encontramos, especialmente en el campo de investigación que pretende construir una máquina inteligente general. No debería sorprendernos que este dilema haya permeado también el desarrollo de la ética en las máquinas inteligentes, ya que debemos reflexionar sobre las implicaciones éticas de construir una máquina de este tipo y los posibles riesgos y perjuicios que podría ocasionar. En esencia, estamos apuntando a la creación de una máquina general que no solo pueda actuar en múltiples dominios, sino que lo haga de manera segura y ética, generando confianza en su funcionamiento (Bostrom & Yudkowsky, 2014, p. 5). El desarrollo e implementación de máquinas inteligentes generales se convierte en un motivo adicional para reflexionar sobre la ética de las máquinas inteligentes, ya que debemos evitar las malas consecuencias y considerar qué acciones y decisiones pueden ser consideradas buenas y cuáles reprochables (Allen et al., 2011, 13).

Sin embargo, especificar el buen comportamiento de máquinas inteligentes generales no es una tarea fácil. En primer lugar, implica reflexionar sobre lo que entendemos por "bueno" en el contexto de una máquina de este tipo. Por ejemplo, Deep Blue fue considerada buena porque aprendió a jugar ajedrez y venció a los mejores jugadores humanos. Watson es considerado bueno porque puede detectar enfermedades de manera rápida y precisa. En estos casos, lo bueno de una acción depende de los propósitos específicos diseñados e implementados en la máquina para su interacción con los usuarios. Sin embargo, esto no ocurre con las máquinas inteligentes generales. Como señalan Bostrom y Yudkowsky (2014), debemos especificar que una máquina inteligente general actúa adecuadamente si elige una acción específica X tal que las consecuencias de X no sean perjudiciales. Además, no podemos pasar por alto que nosotros también actuamos bajo principios éticos (p. 5). Por lo tanto, algunos investigadores argumentan que la máquina debe tener motivaciones éticas, ya sea incorporadas en su código o aprendidas, que la guíen a elegir X en función de la deseabilidad de sus consecuencias y que pueda proporcionar una explicación ética de dicha elección.

Ya hemos constatado lo difícil que resulta cuestionarse el desarrollo de una conciencia en las máquinas inteligentes, dado que aún no comprendemos por completo qué constituye la conciencia humana. Este desafío se intensifica aún más si deseamos dotar a estas máquinas de principios y aptitudes éticas que les permitan ser implementadas de manera segura en la sociedad, lo que plantea nuevos retos en este campo de investigación.

- i) La predictibilidad es un aspecto que debemos comenzar a considerar en las acciones de las máquinas inteligentes. Los comportamientos específicos de estas máquinas

pueden resultar impredecibles, lo que nos lleva a enfocarnos en garantizar la seguridad sin poder anticipar la acción exacta de la IA, centrándonos exclusivamente en sus intenciones.

- ii) Siendo seres humanos limitados, nunca alcanzaremos una teoría ética definitiva y completa que resuelva todos nuestros dilemas morales. Sin embargo, esto difiere de no saber qué constituye una teoría ética adecuada. La ética es un campo tan complejo que, si no logramos delimitarlo y comprenderlo de manera sólida y segura, será aún más difícil implementarlo en máquinas inteligentes generales.
- iii) Debemos profundizar en nuestro conocimiento sobre el aprendizaje y el uso adecuado de la información. Aunque nuestras máquinas ya están aprendiendo, también aprenden de nuestros errores y prejuicios. Debemos enfocar nuestros esfuerzos en mejorar la calidad del aprendizaje y los materiales que utilizamos para enseñar.
- iv) Surge nuevamente la pregunta de si la conciencia es necesaria o no, no solo para el desarrollo de la inteligencia, sino también para el desarrollo de actitudes y comportamientos éticos. La psicología moderna y la neurociencia han revelado recientemente que nuestros procesos de toma de decisiones están influenciados en muchas ocasiones por factores externos y procesos inconscientes más allá de la conciencia o el control interno del agente (Allen et al., 2011, p. 14). Sin embargo, aún no tenemos certeza absoluta al respecto, y es necesario obtener una mayor claridad en estos asuntos para lograr una máquina inteligente general que sea ética.

Que estas sean las cuestiones dominantes en el área nos permite conocer hacia dónde se dirigen los intereses de los investigadores en este campo. La pregunta sobre la "sensibilidad" (sentiente) se revitaliza, lo que requiere un flujo constante y actualizado de información para poder llegar a una respuesta adecuada en diversos entornos. El desarrollo de la Inteligencia Artificial General (AIG) se centró en la sabiduría de los humanos, en las capacidades mentales asociadas a los actos inteligentes. Sin embargo, al intentar pensar en la posibilidad de que una máquina pueda distinguir entre actos reprochables y actos aceptables, nos lleva a considerar la posibilidad de que dicha máquina tenga una experiencia fenoménica o "qualia", es decir, la capacidad de experimentar sensaciones y emociones, de desear ciertos actos y repudiar otros.

Hasta este momento, la responsabilidad por los perjuicios causados por las máquinas inteligentes recaía exclusivamente en los humanos y en las instituciones humanas que estaban involucradas en su funcionamiento y objetivo. Solo las intenciones de los humanos eran consideradas responsables de dichos perjuicios, ya que estaban incorporadas en el funcionamiento y propósito de las máquinas. Los castigos debían ser impuestos exclusivamente a los humanos que causaban estos perjuicios, y las sanciones se aplicaban a las instituciones humanas que permitieron su implementación y salida al mercado. Sin embargo, al considerar la posibilidad de AIG éticas, se está buscando predecir las acciones y que las máquinas inteligentes actúen de manera intencional y motivada teniendo en cuenta las consecuencias. Esto nos obliga a actualizar nuestra noción de responsabilidad o atribuirla a estas creaciones, ya que estarían actuando por sí mismas y para sí mismas.

Lo cual nos lleva a cuestionar la presencia de intencionalidad en las máquinas inteligentes. Siguiendo una línea similar a la Habitación China planteada por Searle (1980), muchos considerarían que un comportamiento carece de ética si no hay intención detrás de él. Kant abordó este tema en 1781 y afirmó que, si bien un acto bueno sin intención puede ser beneficioso, no es moralmente bueno. A medida que el actuar y el proceso de toma de decisiones dependen menos de los programadores y se vuelven exclusivos de la máquina, surge la necesidad de preguntarnos sobre la intención y por qué la máquina actuó de esa

manera (Gips, 2011, p. 10). Desde la década de 1980, Searle afirmaba que, aunque los estados cognitivos de las máquinas no fueran los mismos que los de los humanos, se esperaba que mostraran ciertas motivaciones intencionales, que demostraran que:

- Son capaces de entender la situación y responder preguntas al respecto.
- El programa explica por qué actuó de manera específica y eligió consecuencias sobre otras.

De esta forma, parece que, al igual que con la conciencia, no tenemos ninguna razón para asumir que estas capacidades son exclusivas de los humanos, y solo podremos alcanzarlas si imitamos su funcionamiento. Siempre que el comportamiento de una Inteligencia Artificial General sea lo suficientemente complejo como para no poder excluir la responsabilidad, también tendrá intención (John, P., 2011). La intencionalidad debe pensarse entonces como el actuar complejo que lleva a un beneficio o daño moral y parece ser un acto deliberado y calculado (Siau & Wang, 2020, p. 81). Al final, sin importar si logre esto por medio de una intencionalidad que imite las facultades mentales humanas, o en otro tipo de desarrollo, toda IAG capaz de estos comportamientos tendrá la cualidad de representar objetos, consecuencias, propiedades, etc. (Robinson, 2014).

De esta manera, la funcionalidad o agencialidad moral parece depender de la autonomía y la sensibilidad, independientemente de si se asemejan a la autonomía y sensibilidad humanas o provienen de otro desarrollo o funcionamiento (Wallach, 2009, 570). Máquinas simples como martillos, utensilios de cocina y pistolas no poseen estas dos características. Por otro lado, nos encontramos con los humanos, agentes morales completos. El desarrollo de la Inteligencia Artificial (IA) desde el primer congreso mencionado en la introducción nos ha situado en algún lugar entre estos dos extremos. Tenemos máquinas autónomas sin sensibilidades éticas, como el mencionado Deep Mind. Estas máquinas no están bajo el control directo de ningún agente o usuario y toman decisiones enfocadas únicamente en el objetivo de ganar, pero esto carece de importancia ética (John, P., 2014). Por otro lado, tenemos máquinas con un alto grado de importancia ética, como las simulaciones éticas presentadas en las obras de Peg Tittle, pero que carecen de autonomía y dependen de axiomas impuestos por nosotros.

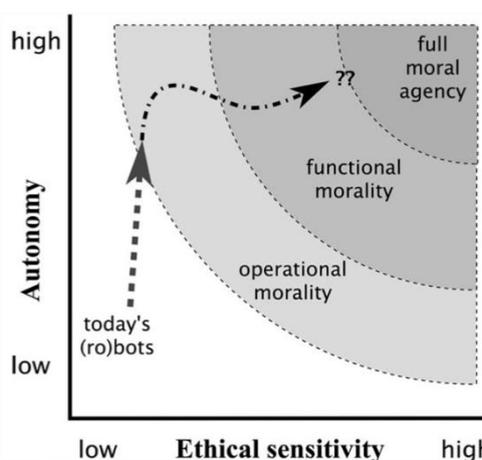


Figura 7: la agencialidad de acuerdo a Sullins (John P., 2011)

El problema de la agencialidad se vuelve más complejo cuando consideramos los niveles de abstracción. (47) los define como la forma en que un agente o entidad describe, analiza y

discute una situación y toma decisiones. Es posible que logremos crear máquinas con sensibilidad ética, con autonomía e incluso con un nivel de abstracción mayor que el nuestro, pero que carezcan de facultades humanas. Tanto la autonomía como la conciencia deben alejarse de la creencia de que la única forma de poseer estas facultades es imitando a los humanos. Esta perspectiva es subjetiva y privilegiada de nuestra parte, al igual que asumir un trato ético por parte de las máquinas inteligentes generales. Sin embargo, las nuevas generaciones nos brindan una esperanza, ya que están moldeando conceptos como la agencialidad a medida que nuestra relación con la IA evoluciona (Turkle, 2011). Niños que se relacionan con la tecnología desde una edad temprana pueden considerar a los objetos computacionales como vivos porque piensan por sí mismos y son autónomos. Así pues, la agencialidad moral puede no depender de lo mucho o poco que se asemeje a la agencialidad moral humana, sino de la autonomía, la sensibilidad ética y los niveles de abstracción de una máquina que nos permitan considerarla como un agente explícito capaz de calcular las mejores consecuencias a través de categorías que logra gracias a su complejo nivel de abstracción, y que actúa motivada éticamente basándose en las posibles consecuencias y los principios que la constituyen (Moor, 2011, p. 5).

Hasta este punto en el desarrollo de la Inteligencia Artificial General (IAG), todas las nociones mencionadas confluyen en la cuestión del estatus moral de una máquina de este tipo. Algunos creen, siguiendo el test de Turing, que solo es necesario que una máquina inteligente pueda ser atribuida con un comportamiento moral y una justificación para este, sin importar su origen (S.L. Anderson, 1995). Por otro lado, Yudkowsky (2014) sostiene que x tiene estatus moral cuando x tiene importe ético por sí mismo, es decir, decide por su propio interés lo que es permisible y lo que no en su accionar (p. 6). Estas definiciones consideran que no necesitamos que las facultades sean las mismas que las humanas, por eso se planteó el principio de no discriminación ontológica, que establece que, si tenemos la misma funcionalidad moral pero otra existencia, tendremos el mismo estatus moral.

En conclusión, el debate sobre el estatus moral de las máquinas de Inteligencia Artificial General nos ha llevado a considerar diferentes perspectivas. Algunos defienden que basta con atribuirles un comportamiento moral y una justificación para este, sin importar su origen. Por otro lado, existe la postura de que una máquina debe tener un sentido ético propio, tomando decisiones basadas en su propio interés y determinando lo que es permisible y lo que no. En este contexto, se ha planteado el principio de no discriminación ontológica, que establece que el estatus moral debe ser igual para entidades con la misma funcionalidad moral, independientemente de su existencia. Estas reflexiones abren un diálogo importante sobre la ética y la responsabilidad en relación con las máquinas inteligentes, generando interrogantes que requieren ser abordados con cautela y sabiduría a medida que avanza el desarrollo de la Inteligencia Artificial General.

## **4. Algunos Dilemas de la Actualidad**

### **4.1 Chat GPT**

En el concepto No. 02 de 2015, la Universidad Nacional de Colombia cuestiona si el plagio es fuente de responsabilidad, afirmando que constituye una fuente de responsabilidad penal, civil y disciplinaria. En dicho documento, se considera el alcance de las funciones del Comité universitario frente a los casos de plagio y otras infracciones al régimen general de derechos de autor. La Universidad se acoge a la definición de la Organización Mundial de Propiedad Intelectual y define el plagio como "el acto de ofrecer o representar como propia, en totalidad

o en parte, la obra de otra persona, en una forma o contexto más o menos alterados" (Universidad Nacional de Colombia, 2015). El plagio, de esta manera, atenta contra los derechos de autor y la propiedad intelectual al apropiarse sin autorización de la obra de alguien más y presentarla como propia. Como afirman en el artículo, esto hace que sea misión de la Universidad responsabilizarse por esto de forma crítica y reglamentar el plagio para que no se atente contra derechos de autor y la autenticidad de una obra. Asimismo, el artículo 14 del Acuerdo 35 de 2003<sup>21</sup>, describe los derechos de autor como el conjunto de normas que protegen los derechos subjetivos del creador de una obra artística, científica o literaria, ya que esta es la manifestación personal y original de la inteligencia.

Por lo tanto, uno de los deberes de los estudiantes es respetar y cumplir las normas relacionadas con la propiedad intelectual, como lo establece el Acuerdo 044 de 2009 del Consejo Superior Universitario, que emitió el Estatuto Estudiantil de la Universidad Nacional de Colombia. Del mismo modo, es responsabilidad del docente actuar con honestidad intelectual y prevenir cualquier tipo de caso<sup>22</sup>. La responsabilidad, además, también cubre a los ciudadanos colombianos. En Colombia, el marco legal protege la obra intelectual y los derechos de autor mediante cuatro leyes:

- i) La Decisión Andina 351 de 1993: Artículo 34: Los artistas intérpretes o ejecutantes tienen el derecho de autorizar o prohibir la comunicación al público, en cualquier forma, de sus interpretaciones y ejecuciones no fijadas, así como la fijación y la reproducción de sus interpretaciones o ejecuciones.<sup>23</sup>
- ii) La Ley 23 de 1982: Los autores de obras literarias, científicas y artísticas gozarán de protección para sus obras de acuerdo con lo establecido en esta Ley y, en cuanto sea compatible con ella, por el derecho común.<sup>24</sup>
- iii) La Ley 44 de 1993: Artículo 51: Quien publique una obra literaria o artística inédita, o parte de ella, por cualquier medio, sin la autorización previa y expresa del titular del derecho, incurrirá en prisión de dos (2) a cinco (5) años y multa de cinco (5) a veinte (20) salarios legales mínimos mensuales.<sup>25</sup>
- iv) El Decreto 1066 de 2015: Establece la obligación de llevar el registro nacional de las obras literarias y artísticas, así como ejercer la inspección y vigilancia sobre las sociedades de gestión colectiva de los derechos reconocidos en la Ley 23 de 1982.<sup>26</sup>

Por otro lado, la información referenciada a continuación proviene del uso intencional de ChatGPT para obtener conocimiento y evaluar su transparencia, responsabilidad, funcionalidad, entre otros aspectos. Por ejemplo, el programa de OpenAI (2021) me enseñó

---

<sup>21</sup>

[http://www.legal.unal.edu.co/rlunal/home/doc.jsp?d\\_i=34248#:~:text=ART%C3%8DCULO%2014.&text=El%20derecho%20de%20autor%20protege,re%C3%BAnen%20caracter%C3%ADsticas%20de%20creaci%C3%B3n%20original.](http://www.legal.unal.edu.co/rlunal/home/doc.jsp?d_i=34248#:~:text=ART%C3%8DCULO%2014.&text=El%20derecho%20de%20autor%20protege,re%C3%BAnen%20caracter%C3%ADsticas%20de%20creaci%C3%B3n%20original.)

<sup>22</sup> [http://www.legal.unal.edu.co/rlunal/home/doc.jsp?d\\_i=44925](http://www.legal.unal.edu.co/rlunal/home/doc.jsp?d_i=44925)

<sup>23</sup> <http://www.sice.oas.org/trade/junac/decisiones/dec351s.asp#:~:text=%2D%20Los%20artistas%20int%C3%A9rpretes%20o%20ejecutantes,de%20sus%20interpretaciones%20o%20ejecuciones>

<sup>24</sup>

<https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=3431#:~:text=Esta%20Ley%20protege%20exclusivamente%20la,obras%20literarias%2C%20cient%C3%ADficas%20y%20art%C3%ADsticas>

<sup>25</sup>

<https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=3429#:~:text=%2D%20Incurrir%C3%A1%20en%20prisi%C3%B3n%20de%20dos,expresa%20del%20titular%20del%20derecho>

<sup>26</sup> <https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=76835>

cómo referenciar la información que me proporcionó de acuerdo con las normas APA y me mostró cómo se vería la bibliografía al final. Además, también le pregunté acerca del plagio en relación con la información que proporcionaba si no se referenciaba correctamente. La inteligencia artificial respondió que su directriz se basa en los términos y condiciones de GPT-3.5, el modelo en el que está basado<sup>27</sup>. Por lo tanto, pregunté si corregir un texto sería considerado plagio. "Si has utilizado ChatGPT para corregir un texto y realizar modificaciones, la cuestión del plagio dependerá de varios factores. En general, si tú eres el autor original del texto y utilizaste ChatGPT como una herramienta para mejorar y corregir tu propio trabajo, entonces no se consideraría plagio, ya que conservas los derechos de autor sobre tu propio contenido" (OpenAI, 2021).

Tomando en cuenta lo anterior y reiterando mi compromiso de no incurrir en ninguna actividad ilegal, me aventuré a corregir el estilo y la ortografía de este texto, siempre y cuando no infringiera derechos de autor. Solicité la ayuda de ChatGpt para corregir cada párrafo en orden y obtener la mejor corrección sin incurrir en plagio. El resultado final lo recae el lector, en formato impreso y en su dispositivo electrónico preferido. En mi opinión, el texto no carece de coherencia ni sentido. Posee un mínimo de errores ortográficos, si es que hay alguno. Y lo más importante, es completamente original y de autoría exclusiva del escritor humano. Para garantizar la originalidad del texto, escribí todo a mano. ChatGpt corrigió y modificó el estilo, pero no el contenido, el cual es propiedad intelectual del autor. En este caso en particular, tanto ChatGpt como el usuario demostramos responsabilidad al buscar transparencia tanto legal como funcional en esta relación. Sin embargo, esto no significa que ChatGpt no pueda ser utilizado con intenciones maliciosas, o que no requerimos algo más para que no solo se nos sugiera no hacer plagio.

Analicemos esto con el siguiente ejemplo: le solicité amablemente a ChatGpt (OpenAI, 2021) que continuara un párrafo en el que estaba trabajando. Este párrafo era en realidad el comienzo de un párrafo de Harari (2018, p. 77), y el bot no solo corrigió y modificó la cita sin tener ninguna manera de comprobar que no era de mi autoría, sino que también continuó el párrafo. ¡Después de esto, le pedí que hiciera el siguiente párrafo y lo hizo! A continuación, en la tabla, muestro la cita de Harari a un lado y lo que la IA hizo al otro:

	Harari	ChatGpt
Párrafo 1	Una vez que la IA decida mejor que nosotros las carreras e incluso las relaciones, nuestro concepto de la humanidad y de la vida tendrá que cambiar. Los humanos están acostumbrados a pensar en la existencia como un drama de toma de decisiones. La democracia liberal y el capitalismo de libre mercado ven al individuo como un agente autónomo que no para de tomar decisiones sobre el mundo. Las obras de arte (ya sean las piezas teatrales de Shakespeare, las novelas de Jane Austen o las chabacanas comedias de	Una vez que la IA decida mejor que nosotros las carreras e incluso las relaciones, nuestro concepto de la vida tendrá que cambiar. Nuestro entendimiento de la existencia se transformará a medida que la inteligencia artificial tome decisiones más acertadas que las nuestras, tanto en el ámbito laboral como en el de las relaciones personales. Este nuevo paradigma requerirá que reevaluemos nuestros valores y perspectivas, adaptándonos a un mundo en constante evolución.

<sup>27</sup> <https://openai.com/policies/terms-of-use>

	<p>Hollywood) suelen centrarse en que el o la protagonista ha de tomar alguna decisión particularmente crucial. ¿Ser o no ser? ¿Hacer caso a mi mujer y matar al rey Duncan, o hacer caso a mi conciencia y perdonarlo? ¿Casarme con el señor Collins o con el señor Darcy? Las teologías cristiana y musulmana se centran de manera parecida en el drama de la toma de decisiones, y aducen que la salvación o la condena eternas dependen de haber tomado la decisión correcta.</p>	<p>La coexistencia con la IA nos desafiará a replantear nuestras creencias y encontrar un equilibrio entre la tecnología y nuestra propia humanidad. A medida que la IA se adentre en los dominios de la toma de decisiones más complejas, es posible que nos sintamos tentados a cederle el control total. Sin embargo, debemos recordar que la esencia de nuestra humanidad radica en nuestra capacidad de elección y libre albedrío. Si bien la IA puede ofrecer sugerencias valiosas y eficientes, debemos ser cautelosos al ceder por completo nuestra autonomía.</p>
<p>Párrafo 2</p>	<p>¿Qué pasará con esta forma de entender la vida si cada vez confiamos más en la IA para que tome las decisiones por nosotros? En la actualidad nos fiamos de Netflix para que nos recomiende películas y de Google Maps para elegir si giramos a la derecha o a la izquierda. Pero una vez que empezamos a contar con la IA para decidir qué estudiar, dónde trabajar y con quién casarnos, la vida humana dejará de ser un drama de toma de decisiones. Las elecciones democráticas y los mercados libres tendrán poco sentido. Lo mismo ocurrirá con la mayoría de las religiones y de las obras de arte. Imagine el lector a Anna Karénina sacando su teléfono inteligente y preguntándole al algoritmo de Facebook si debe seguir casada con Karenin o fugarse con el conde Vronsky. O imagine el lector su obra teatral favorita de Shakespeare con todas las decisiones cruciales tomadas por el algoritmo de Google. Hamlet y Macbeth llevarían una vida mucho más confortable, pero ¿qué tipo de vida sería, exactamente? ¿Tenemos modelos para dar sentido a una existencia de este tipo?</p>	<p>Es esencial mantener una mentalidad crítica y seguir siendo participantes activos en nuestras propias vidas, utilizando la tecnología como una herramienta complementaria en lugar de una autoridad definitiva. De esta manera, podemos mantener un equilibrio entre la innovación tecnológica y nuestra identidad como seres humanos conscientes. A medida que avanzamos hacia ese futuro incierto, solo podemos anticipar los cambios que nos esperan y prepararnos para abrazar una nueva forma de vida en la que la IA sea una compañera inseparable. En última instancia, dependerá de nosotros aprovechar los beneficios de la IA sin perder nuestra esencia y capacidad para tomar decisiones que reflejen nuestra individualidad y valores fundamentales.</p>

Tabla 1: Chat GPT plagiando a Harari

Como si esto fuera poco, recordemos ese artículo que todo estudiante lee en clase de hermenéutica por allá en primer semestre de pregrado de filosofía: “¿Qué es la ilustración?” de Kant. Este texto, irónicamente, argumenta a favor de la necesidad de que nosotros los seres humanos pensemos por nosotros mismos y salgamos de una minoría de edad autoimpuesta. Sin embargo, con el objetivo explícito de mostrar cómo ChatGPT puede estar generando todo lo contrario, le pedí al bot que me hablara un poco de este texto, que posteriormente hiciera un esquema de él y que, por último, produjera un texto completo acerca del texto. Aunque inicialmente el artículo fue muy pequeño, le pedí al bot que ampliara la introducción y en vez de un párrafo corto me brindó seis párrafos, lo que bien podría ser una introducción de un texto universitario. No lo hice con los demás temas del artículo, pero si el autor tiene dudas puede hacerlo con cualquier texto filosófico que sea de antes de septiembre del 2021:

Resumen del Texto	Esquema del Texto	Breve artículo del texto	Complementación de la introducción
<p>La Ilustración, según la perspectiva de Immanuel Kant, fue un movimiento intelectual y filosófico que se desarrolló en Europa durante el siglo XVIII. Kant fue un destacado filósofo de la Ilustración y su obra "¿Qué es la Ilustración?" es una referencia importante en este contexto.</p> <p>En su ensayo, Kant define la Ilustración como la liberación del individuo de su "inmadurez" autoimpuesta. Esta inmadurez se caracteriza por la incapacidad de pensar por sí mismo y de confiar en su propio juicio. Kant argumenta que la Ilustración se logra cuando una persona se atreve a usar su propia razón de</p>	<p><b>I. Introducción</b></p> <p>A. Definición de la Ilustración como la salida del estado de inmadurez</p> <p>B. Importancia de la razón y la autonomía en la Ilustración</p> <p><b>II. La inmadurez autoimpuesta</b></p> <p>A. Dependencia de la autoridad externa</p> <p>B. Falta de confianza en el propio juicio</p> <p><b>III. La Ilustración como emancipación</b></p> <p>A. Uso autónomo de la razón</p> <p>B. Liberación de las tutelas y prejuicios</p> <p><b>IV. Libertad de pensamiento y autonomía moral</b></p> <p>A. Crítica de las creencias establecidas</p> <p>B. Independencia intelectual y moral</p> <p><b>V. El papel de la</b></p>	<p><b>Título:</b> La Ilustración de Kant: El poder de la razón y la emancipación intelectual</p> <p><b>Introducción:</b> La Ilustración fue un movimiento intelectual que floreció en el siglo XVIII, cambiando el curso de la historia y dejando una profunda huella en la sociedad y el pensamiento modernos. Entre los pensadores más influyentes de este período se encuentra Immanuel Kant, cuyo ensayo "¿Qué es la Ilustración?" ofrece una visión reveladora sobre esta corriente filosófica.</p> <p><b>La inmadurez autoimpuesta:</b> Kant comienza su ensayo argumentando que la humanidad ha caído</p>	<p>La Ilustración fue un extraordinario movimiento intelectual que emergió en Europa durante el siglo XVIII. Marcó una época de cambios radicales en la forma en que las personas concebían el mundo, la sociedad y su propio papel en ella. Fue una era de cuestionamiento, razón y búsqueda de la emancipación intelectual. Entre los pensadores más destacados y trascendentales de este período se encuentra el filósofo alemán Immanuel Kant, cuyo ensayo "¿Qué es la Ilustración?" brinda una visión profunda y reveladora sobre los principios y valores fundamentales de esta corriente</p>

<p>manera autónoma, sin depender de la autoridad externa. La Ilustración de Kant promovió la libertad de pensamiento, la autonomía moral y la búsqueda del conocimiento mediante el uso de la razón. Kant enfatizó la importancia de la crítica y el cuestionamiento de las creencias establecidas, alentando a las personas a pensar de forma independiente y a no aceptar dogmas sin un examen riguroso. En resumen, la Ilustración de Kant se refiere a un movimiento intelectual que valoraba la autonomía y el razonamiento crítico como medios para alcanzar el progreso y la emancipación del individuo. Fue un importante período en la historia de la filosofía y sentó las bases para muchas de las ideas que influyeron en la sociedad y el pensamiento modernos.</p>	<p><b>razón y la crítica</b>  A. Valoración de la reflexión y el cuestionamiento  B. Rechazo de la aceptación acrítica de dogmas</p> <p><b>VI. Conclusión</b>  A. Importancia histórica y filosófica de la Ilustración  B. Relevancia de la autonomía y el razonamiento crítico en la sociedad.</p>	<p>en un estado de inmadurez autoimpuesta. Esta inmadurez se caracteriza por la dependencia de la autoridad externa y la falta de confianza en el propio juicio. Kant desafía a las personas a liberarse de esta tutela y a tomar la responsabilidad de pensar por sí mismas.</p> <p><b>La Ilustración como emancipación:</b>  Según Kant, la Ilustración implica la emancipación del individuo de la inmadurez autoimpuesta. En lugar de aceptar pasivamente las verdades y creencias establecidas, la Ilustración invita a las personas a usar su propia razón de manera autónoma. Es un llamado a cuestionar las normas establecidas y a buscar la verdad y el conocimiento a través de la reflexión y el razonamiento crítico.</p> <p><b>Libertad de pensamiento y autonomía moral:</b></p>	<p>filosófica. La Ilustración fue una respuesta al antiguo régimen y a las estructuras de poder arraigadas que caracterizaron gran parte de la historia anterior. Se produjo en un momento de cambios sociales, políticos y científicos, y estuvo influenciada por los avances en la ciencia, la filosofía y la razón. Los filósofos ilustrados promovieron la idea de que la razón y la evidencia empírica debían utilizarse como guías para comprender el mundo y mejorar la sociedad. La Ilustración fue un movimiento que valoraba la libertad individual, la autonomía moral y el progreso intelectual. En contraposición a la autoridad absoluta de la monarquía y la iglesia, los ilustrados buscaban la emancipación intelectual y la difusión del conocimiento. Creían en la capacidad de la razón humana para descubrir la verdad,</p>
--	---	---	--

		<p>La Ilustración de Kant enfatiza la importancia de la libertad de pensamiento y la autonomía moral. Se insta a las personas a examinar críticamente las ideas y creencias existentes, y a no aceptarlas ciegamente. La autonomía moral implica tomar decisiones éticas basadas en la propia razón y no simplemente obedecer los dictados externos. Es un llamado a actuar como seres autónomos y responsables.</p> <p><b>El papel de la razón y la crítica:</b> Kant considera que la razón y la crítica desempeñan un papel central en la Ilustración. La razón nos capacita para pensar de manera independiente, evaluar las ideas y formar nuestro propio juicio. La crítica nos permite cuestionar las suposiciones arraigadas y las verdades aceptadas, fomentando así el progreso intelectual</p>	<p>resolver problemas y construir una sociedad más justa y racional.</p> <p>Immanuel Kant, uno de los principales representantes de la Ilustración, desempeñó un papel fundamental al articular los principios clave de este movimiento en su ensayo "¿Qué es la Ilustración?". Kant argumentaba que la Ilustración consistía en abandonar la inmadurez y atreverse a pensar por uno mismo, sin depender de la autoridad externa. Instaba a las personas a utilizar su propia razón de manera autónoma y a no aceptar dogmas sin una evaluación crítica.</p> <p>El legado de la Ilustración y la obra de Kant continúan siendo de relevancia en la actualidad. Sus ideas sentaron las bases para la democracia, los derechos humanos, la separación de poderes y la valoración de la ciencia y la educación. La Ilustración no solo</p>
--	--	---	---

		<p>y social.</p> <p><b>Conclusión:</b> La Ilustración de Kant representa una búsqueda apasionante de la emancipación intelectual y moral a través del uso autónomo de la razón. Invita a cada individuo a liberarse de las ataduras de la inmadurez y a abrazar la libertad de pensamiento. A través de la Ilustración, Kant nos recuerda el poder transformador de la razón y nos desafía a ser pensadores críticos y agentes de cambio en nuestra propia época.</p>	<p>cambió la forma en que las personas percibían el mundo, sino que también influyó en la manera en que se estructuraron las instituciones políticas y sociales en muchas sociedades modernas.</p> <p>En resumen, la Ilustración fue un movimiento intelectual revolucionario que buscó la emancipación del individuo a través de la razón, la libertad y la autonomía moral. Immanuel Kant y su ensayo "¿Qué es la Ilustración?" nos ofrecen una visión valiosa de los ideales y principios fundamentales de este movimiento, cuyo impacto perdura en la sociedad y el pensamiento contemporáneos.</p>
--	--	---	---

Tabla 2: Chat GPT vs. "¿Qué es la Ilustración?"

Adicionalmente, recordando mi clase de Lógica I, le pedí a la IA que me hablara de los tipos de falacias. En primer lugar, le pedí que identificara los argumentos y las conclusiones de un texto de Platón y la máquina lo hizo inmediatamente encontrando tres argumentos y tres conclusiones. Le pedí que en este mismo texto identificara posibles falacias. Sin embargo, al no encontrar extrapolé el ejercicio a uno de los párrafos ejercicios de Copi con el cual la máquina explico que había una falacia, el tipo de falacia y por qué:

Lógica I	Hallar argumentos	Tipos de falacias	Identificar falacias
Input dado	¿Puedes organizar esto por argumentos y conclusiones?	¿Qué tipos de falacias existen?	¿Acá encuentras falacias?

<p>Consideremos, por ejemplo, una noción como la de justicia. Si nos preguntamos qué es la justicia, es natural proceder considerando este acto justo, aquél y aquel otro, con vistas a descubrir lo que tienen en común. Deben todo ellos, en algún sentido, participar de una naturaleza común que se encontrará en todo lo que es justo y en nada más. Esta naturaleza común, en virtud de la cual todos ellos son justos, será la justicia misma, la pura esencia cuya mezcla con los hechos de la vida ordinaria produce la multiplicidad de actos justos. Lo mismo ocurre con cualquier otra palabra que pueda ser aplicable a hechos comunes, como “blancura” por ejemplo. La palabra será aplicable a un número de cosas particulares porque todas ellas participan de una naturaleza o esencia común. Esta esencia pura es lo que Platón llama una “idea” o “forma”. (No debe suponerse que las “ideas”, en este sentido, existen en las mentes, aunque pueden ser aprehendidas por las mentes.) La “idea” justicia no es idéntica a nada que sea justo: es algo distinto de las cosas particulares y de lo cual participan las cosas particulares. No siendo particular, no puede ella misma existir en el mundo de los sentidos. Además, no es fugaz o mutable como los objetos de los sentidos: es eternamente ella misma, inmutable e indestructible. Así Platón es llevado a un mundo suprasensible, más real que el</p>		<p>“El profesor de guardia sorprende a Silvia fumando un cigarro, y la pide que lo tire, puesto que no se puede fumar en el instituto. Silvia contesta indignada, que no hay derecho a obligarla a apagar el cigarro, ya que muchos profesores fuman. ¿Está justificada la indignación de Silvia? ¿Está razonando falazmente? Si es así, indica en qué tipo de falacia se basa su indignación.”</p>
--	--	---

<p>mundo común de los sentidos, el mundo inmutable de las ideas [...]. El verdadero mundo real, para Platón, es el mundo de las ideas; pues sea lo que fuere que podamos tratar de decir acerca de las cosas del mundo de los sentidos, sólo podemos lograr decir que participan de tales y cuales ideas, las cuales, por tanto, constituyen toda su peculiaridad. De aquí es fácil pasar a un misticismo. Podemos esperar, en una iluminación mística, ver las ideas como vemos los objetos de los sentidos; y podemos imaginar que las ideas existen en el cielo. Estos desarrollos místicos son muy naturales, pero la base de la teoría está en la lógica, y como basada en la lógica debemos considerarla. La palabra “idea” ha adquirido, en el curso del tiempo, muchas asociaciones que son totalmente desorientadoras cuando se aplican a las “ideas” de Platón. Usaremos por tanto la palabra “universal”, en vez de la palabra “idea”, para describir aquello a lo que Platón se refería. La esencia del tipo de entidad a la que Platón se refería es que es opuesta a las cosas particulares que son dadas en la sensación. Hablamos de cuanto es dado en la sensación, o es de la misma naturaleza que las cosas dadas en la sensación, como de un particular; por oposición a esto, un universal será cualquier cosa que pueda ser compartida por muchos particulares y que tenga las características que, como vimos, distinguen la justicia y la</p>		
---	--	--

	blancura de los actos justos y de las cosas blancas.		
Output otorgado por Chat GPT	<p>Argumentos:</p> <ul style="list-style-type: none"> <li>• La noción de justicia implica la existencia de una naturaleza común que se encuentra en todos los actos justos. Esta naturaleza común es la esencia pura de la justicia, una "idea" o "forma" que no es idéntica a los objetos particulares, pero de la cual participan.</li> <li>• Las ideas o formas, como la justicia, existen en un mundo suprasensible e inmutable de las ideas. Este mundo de las ideas es más real que el mundo común de los sentidos, ya que las cosas del mundo sensible solo pueden ser descritas como participantes de las ideas.</li> <li>• La teoría de las ideas de Platón se basa en la lógica y describe entidades universales, opuestas a las cosas particulares que se experimentan en los sentidos. Estos universales pueden ser compartidos por muchos particulares y poseen características distintivas.</li> </ul> <p>Conclusiones:</p> <ul style="list-style-type: none"> <li>• La justicia, como una idea o forma, trasciende los actos justos y las cosas justas particulares. Es una esencia pura que existe en un mundo suprasensible y eterno.</li> </ul>	<p>Existen varios tipos de falacias que se encuentran comúnmente en los argumentos. Algunos ejemplos incluyen:</p> <ul style="list-style-type: none"> <li>• Falacia ad hominem: Atacar o desacreditar a una persona en lugar de abordar su argumento.</li> <li>• Falacia de la pendiente resbaladiza: Argumentar que si se permite un cierto evento o acción, se desencadenarán consecuencias extremas y no deseadas, sin una evidencia sólida.</li> <li>• Falacia de la generalización apresurada: Sacar conclusiones generales a partir de un número limitado de ejemplos.</li> <li>• Falacia del hombre de paja: Distorsionar o</li> </ul>	<p>En el diálogo proporcionado, se puede identificar una falacia conocida como "falacia de relevancia" o "falacia del tu quoque" (tú también). Esta falacia ocurre cuando se desvía la atención del argumento principal al señalar la hipocresía o las acciones inconsistentes de la otra persona, en lugar de abordar directamente el punto en cuestión. En este caso, Silvia razona falazmente al justificar su acción de fumar en el instituto apelando al hecho de que otros profesores también fuman. Su indignación se basa en el argumento de que si los profesores pueden fumar, ella también debería poder hacerlo. Sin embargo, este argumento no aborda la regla existente en la institución sobre la prohibición de fumar. La respuesta apropiada sería</p>

	<ul style="list-style-type: none"> <li>• Las ideas o formas, como la justicia, son la base de la realidad y proporcionan la verdadera peculiaridad a las cosas del mundo sensible.</li> <li>• El concepto de universales, como la justicia, nos permite entender cómo múltiples objetos particulares pueden compartir una naturaleza común y ser descritos por medio de la participación en dichos universales.</li> </ul>	<p>exagerar los argumentos de alguien para hacerlos más fáciles de refutar.</p> <ul style="list-style-type: none"> <li>• Falacia de causa falsa: Asignar una relación causal incorrecta entre dos eventos solo porque ocurren juntos.</li> <li>• Falacia de autoridad: Aceptar una afirmación como verdadera solo porque una figura de autoridad la hace, sin evidencia adicional.</li> <li>• Falacia de la petición de principio: Presuponer como verdadero aquello que se está tratando de demostrar.</li> <li>• Falacia de la falsa dicotomía: Presentar una situación como si solo hubiera dos opciones</li> </ul>	<p>evaluar si la regla del instituto es justa o no, en lugar de simplemente señalar la hipocresía de otros. Al desviar la atención hacia los comportamientos de otros profesores, Silvia comete la falacia del "tu quoque".</p>
--	--	--	---

		<p>posibles, cuando en realidad existen más.</p> <ul style="list-style-type: none"> <li>• Falacia del argumento ad ignorantiam: Argumentar que algo es verdadero solo porque no se ha demostrado que sea falso, o viceversa.</li> <li>• Falacia de la falacia: Considerar que una afirmación es falsa solo porque se utiliza una falacia en su argumento.</li> </ul> <p>Estos son solo algunos ejemplos, y hay muchas otras falacias que se pueden encontrar en el razonamiento y el debate.</p>	
--	--	--	--

Tabla 3: Chat GPT vs. Lógica I

La inteligencia es cada día menos una manifestación personal, original y exclusiva de los seres humanos, pues la inteligencia artificial está permitiendo que la relación entre máquinas y humanos sea muestra de mayor inteligencia que la que pueden mostrar los seres humanos por su cuenta. Estas máquinas se están desempeñando mucho mejor que los humanos en áreas específicas de la inteligencia, por lo que debemos cuestionar cómo se están transformando nuestras interacciones cognitivas y qué debemos hacer para que esto nos beneficie más de lo que puede llegar a perjudicarnos. El plagio ya no es solo una cuestión de robar la propiedad de otra u otras personas, también debemos considerar que la gente intentará producir textos y robar la propiedad intelectual de las máquinas inteligentes al demostrar ser mucho más rápidas y mejores que nosotros a la hora de corregir y producir textos universitarios. Desde nuestro contexto actual situado debemos considerar que vivimos en un punto en el que un texto puede ser producido por una máquina o por un ser humano de igual manera. Debemos preguntarnos por formas de distinguir estos dos tipos de producción

intelectual y la acertada implementación de lo que pueda llegar a producir una máquina inteligente para no actuar de manera malintencionada o éticamente repudiable.

Las normas que regulan el plagio y la propiedad de autor no tienen en cuenta este tipo de nuevas tecnologías y mucho menos que éstas ya están cambiando nuestra forma de ver las cosas. Todo, desde nuestra manera de investigar y recopilar información, pasando por la traducción y corrección de textos, hasta la mismísima producción intelectual, todo está siendo moldeado mientras esas tecnologías ven la luz del día. Lo que es peor, no hay ningún método o forma de garantizar que los textos que produce ChatGPT no sean usados por personas y presentados como propios de manera maliciosa. Si gente está haciendo esto, además, no hay forma alguna de corroborar si es un texto de ChatGPT o propio por lo que es muy posible que muchas personas estén haciendo esto. Cuando le pedimos que haga un texto, el bot solo nos recuerda que debemos referenciar al usar texto que no es nuestro. Toda la decisión de qué hacer con el texto recae enteramente en la decisión y en la agencialidad del ser humano que está usando sus servicios. Y el problema es que ni estudiantes ni profesores han asumido responsabilidad por esto. Hasta que no nos tomemos este problema como algo serio, no tendremos leyes que regulen el uso de estas tecnologías y nadie será responsable en su diseño y su implementación.

Por más que el chat-bot de GPT no sea humano ni tenga conciencia o aptitudes éticas, esta es una IA que con aprendizaje profundo ha logrado incluso producir textos con sentido y que denotan inteligencia. Tanto así que cualquier persona podría presentar un texto de su producción como una asignatura del colegio o de la universidad. Pero aún con esto en mente, la ley no protege ni garantiza que las personas respeten esta propiedad de autor, mucho menos la Universidad Nacional de Colombia. Nada garantiza que ChatGPT, al ser un arma de doble filo, sea usada de forma correcta y no implemente para hacer cosas de manera maliciosa. El bot no hace nada más que sugerir el uso de referencias sin preocuparse por la necesidad de proteger su producción intelectual y la de los autores humanos que pueden ser puestas bajo el escrutinio del bot solo para que las modifique y no parezca información robada. Pero proteger la propiedad intelectual de un chat-bot no es empresa fácil ya que al exigir este tipo de garantías y derechos a una máquina inteligente nos hace reconsiderar la autonomía de este tipo de máquinas. Al fin y al cabo, por más que esté diseñada por seres humanos y por más que la decisión final recaiga en los límites del diseño humano, es el chat-bot el que utiliza los datos dados y produce contenido intelectual y con sentido. Tenemos razones para afirmar que esta máquina está afectando, al menos, la manera en que concebimos la propiedad intelectual, la manera de investigar, la manera de escribir y puede que muchas otras nociones. Debemos o bien limitar los usos de las máquinas a aquellos que sean beneficiosos y no acarreen riesgos potenciales. O debemos garantizar una manera de distinguir de forma transparente entre ambos para saber cuándo se implementa de forma éticamente aceptable y cuando no.

Pero para pensarnos la autonomía y responsabilidad de este tipo de máquinas. No podemos dejar de lado que el objetivo ulterior detrás de este software es simplemente el de mostrarse inteligente, tal como Eugene Goostman que superó el Test de Turing. Por más producción intelectual que pueda llegar a producir de forma autónoma a través de conjuntos gigantes de datos, el objetivo deliberativo de estas máquinas es simplemente parecer inteligente y esto lo puede llevar a cabo sin considerar la propiedad intelectual, el plagio, el detrimento de las capacidades cognitivas humanas entre otros. Así, es importante que haya una limitación en las indicaciones válidas de estos desarrollos más allá de no ayudar a construir armas o matar gente (OpenAI, 2021). Pero también se debe responsabilizar a la humanidad para que se

piense estas indicaciones y nos responsabilicemos por dotar a estas máquinas de indicaciones que sean justas y no transgreda la ética o la moral.

No podemos dejar esto de lado a la hora de diseñar e implementar máquinas inteligentes. Estaríamos ignorando que este tipo de tecnologías ya están cambiando nuestros valores éticos, intelectuales y legales. La IA muestra cómo la conciencia humana y sus aptitudes éticas no son necesarias para que algo o alguien pueda afectar nuestros valores y tenga que pensarse desde un punto de vista ético e incluso legal. En nuestro contexto social situado particular, hemos logrado ver que al menos al nivel cognitivo esto está sucediendo con tecnologías, y si esto no es tenido en cuenta, sino que implementamos sin más estos avances, hasta nuestras capacidades cognitivas básicas podrían estar deteriorándose y viéndose afectadas. En el caso en el que le pedimos a ChatGPT que corrigiera y continuara una cita de Harari, vimos ya bastantes riesgos y peligros potenciales que no han sido limitados ni considerados en su uso e implementación. Primero, el bot no tiene responsabilidad a la hora de analizar texto, pues no protege la propiedad intelectual de referencias al no identificarlas y modificarlas. Los creadores de este Chat Bot siempre han garantizado tranquilidad, seguridad y transparencia en el desarrollo de esta tecnología [referencia], pero nuestra propiedad intelectual puede estar siendo perjudicada y se puede estar beneficiando el plagio y la apropiación de obras que no son de uno. Requerimos más transparencia para lograr entender cómo funcionan estos algoritmos y verdaderamente garantizar que sean seguros y así limitar la posibilidad de que sean usados para cosas éticamente reprochables.

Además, al desarrollar el ejemplo del chat-bot con literatura obligada en la carrera de filosofía, vimos cómo el Bot es capaz de brindarnos resúmenes argumentativos, esquemas con los puntos y argumentos principales, ejemplos, fechas, contexto e incluso puede elaborar un texto argumentativo con relevancia filosófica acerca de la fuente o el tema que nosotros deseemos. Si no garantizamos que solo se usen estas tecnologías con buena intención y permitimos que se use de mala intención, esta tecnología permitirá incluso producir textos universitarios y presentarlos en cualquier universidad como de uno. Al pensar en este problema junto con la falta de responsabilidad y transparencia que muestra el Chat Bot, vemos como las malas consecuencias y los riesgos en potencia están presentes y pueden llegar incluso a generar fallas en nuestras funciones cognitivas básicas. Si un estudiante de primer semestre ha estado usando esta tecnología y otras similares y lo hace durante toda su carrera, es probable que cuando se gradúe no sepa ni siquiera leer un texto y separarlo por argumentos y conclusiones, escribir sin faltas ortográficas o de estilo, poner referencias adecuadas, aportes filosóficos y que no provengan de una IA. Como dijo el mismo ChatGPT cuando creó un texto a partir de la cita de Harari (2021), debemos ser cautelosos al ceder nuestra autonomía a estas máquinas. Incluso debemos pensar cómo, al estar cediéndola, debemos pensar en la autonomía y el estatus moral de máquinas, aunque no tengan conciencia humana ni mucho menos actitudes éticas.

Si no empezamos a pensar en estas máquinas de manera holística, no solo en sus posibles beneficios, sino también junto con sus posibles afectaciones al tener cierto grado de autonomía y, por tanto, de agencialidad, no generaremos responsabilidad frente a estas máquinas, sino que evadiremos esto e incluso podría llegar un día en el que perdamos esa autonomía y agencialidad que creíamos tan única en nosotros. Las máquinas pueden retroceder a esa inmadurez autoimpuesta en la que no podemos pensar por nosotros mismos, ya que hacemos de la relación entre humanos y máquinas algo más necesario y fundamental para nuestro florecimiento humano. Por esto, antes de usar este tipo de tecnologías debemos delimitar claramente y concisa los límites de usar una máquina de manera correcta o

incorrecta, o limitar las capacidades de la máquina a usos beneficiosos a corto y largo plazo. Además, no debe garantizarse solo después de la implementación de máquinas, cuando ya hayan logrado causarnos perjuicios. Se nos debe garantizar que antes de que esta tecnología saliera al mercado se hizo el respectivo estudio y análisis de riesgos por el que predijimos algunos y tratamos de prevenirlos de la mejor manera posible antes de su implementación y uso. Es precisamente por esto que la participación política y democrática es fundamental, sin leyes que regulen el uso de tecnologías inteligentes, pueden llegar a ser más los perjuicios que genere una máquina a largo plazo que los beneficios que aparentemente vemos en la inmediatez y rapidez con que son implementadas estas tecnologías.

La transparencia, además, debe extrapolarse incluso a la intención y el objetivo detrás de estas tecnologías, pues no podemos dejar de lado que estos bots no son un producto final sino un experimento que recopila datos y prueba el uso de estas tecnologías para futuros desarrollos. La máquina está programada para parecer inteligente, pero también debemos tener transparencia de hacia dónde se dirigen las motivaciones de investigadores con esto y los posibles nuevos desarrollos que traerá consigo. Nada garantiza que con esta recolección de datos se proteja nuestra privacidad, que no pueda traer malas consecuencias al no estar regulada o que sean datos que se recojan en este momento por su importancia para nuevos desarrollos, que una vez implementados puedan ser tecnologías tan caras que solo el 10% de arriba pueda comprar, generando más inequidad.

## **4.2 Armamento Militar**

El desarrollo militar es otro ámbito en el que debemos tener más precaución a la hora de implementar la Inteligencia Artificial (IA). En efecto, su aplicación en estos entornos puede crear armamento inteligente capaz de causar un daño mayor que un batallón de humanos con rifles automáticos. Recientemente, desde aproximadamente 2018, hemos sido testigos de una carrera para diseñar, implementar y utilizar estas armas. Es probable que no esté lejos el día en el que actores no estatales, industrias privadas e incluso departamentos de policía hagan uso de estas armas en la sociedad. La mencionada carrera ha estado en marcha desde 2018, pero es importante destacar que ya desde comienzos de los años 2000, se empezó a seguir este camino cuando los estadounidenses utilizaron miles de drones controlados por humanos en Afganistán e Irak para desactivar bombas, realizar ataques predator y obtener información sobre objetivos, entre otros propósitos. Ahora este dilema se intensifica porque la IA se está implementando específicamente con el objetivo de asesinar, neutralizar o eliminar enemigos, bases operativas, transporte enemigo y tecnología enemiga, entre otros; aunque pueda beneficiar a las tropas que las utilizan, también puede causar daño directo a las tropas enemigas e incluso a civiles inocentes y desarmados que se convierten en víctimas. En efecto, el motivo original por el cual Estados Unidos implementó estas tecnologías era reemplazar a los humanos en situaciones de guerra que generan numerosas bajas y heridas físicas y mentales, como en los combates cercanos, la desactivación de armas y el reconocimiento de lugares. Sin embargo, la situación era diferente para los afganos, ya que esta tecnología fue programada y diseñada para que, al ser controlados por humanos, sus objetivos específicos fueran dañinos para ellos.

Como pasó en el siglo pasado con la energía nuclear, la IA aplicada a armamento plantea la necesidad de considerarla como un problema de uso dual, un avance con la posibilidad de utilizarlo moralmente correcta o incorrecta. En casos previos en los que se ha demostrado intención maliciosa, ya hemos sido testigos de cómo la IA ha sido utilizada en circunstancias

como herramienta para perjudicar, estafar, robar información, entre otros fines. En el ámbito militar, el problema se vuelve más grave, ya que el objetivo final parece ser desarrollar la IA para que funcione de manera autónoma, neutralizando, combatiendo y protegiéndose de enemigos. En otras palabras, los militares están intentando aplicar todo el potencial que la humanidad ha alcanzado con la IA para crear armamento militar con mínima o nula intervención humana, permitiendo que estas armas tomen decisiones en el campo de batalla e incluso lleguen a cometer acciones letales. Por ejemplo, desde 2016 hemos sido testigos del desarrollo de robots perros por parte de científicos de Boston Dynamics, Xiaomi, MIT y otros actores, que han demostrado su utilidad en empresas y fábricas para asegurar lugares, detectar riesgos como radiación peligrosa y resolver diversos problemas que podrían representar peligros para los seres humanos.<sup>28</sup> Estos robots han sido creados mediante el uso de aprendizaje profundo, lo que les permite aprender a caminar en diversos terrenos y adaptarse a diferentes contextos.<sup>29</sup> El alcance de estos perros robots ha llevado a que, en 2020, estos hicieran su debut en el campo militar. El equipo Químico, Biológico, Radiológico y Nuclear (QBRN) del 5.º Escuadrón de Ingenieros Civiles del ejército estadounidense implementó estos perros robots para reemplazar a los perros y humanos que vigilaban el perímetro de la base. Esta implementación ha sido seguida por muchas otras, incluyendo su uso en la seguridad de la guardia costera de Honolulu, Hawái, para prevenir la distribución de armas de destrucción masiva y en el Centro de Guerra Naval en Filadelfia.<sup>30</sup>

Los perros robots hasta el momento se han utilizado con propósitos beneficiosos, como proteger y vigilar perímetros. Sin embargo, es importante tener en cuenta que su potencial de uso no se limita a aplicaciones benignas, ya que existe la posibilidad de que puedan ser modificados o equipados con armas para tomar decisiones letales si perciben una amenaza dentro del área que vigilan. Esta idea plantea un dilema moral y nos lleva al siguiente ejemplo: en las guerras recientes, el campo de batalla está siendo cada vez más permeado por la IA. Desde al menos el año 2014, se han utilizado drones de guerra en Irak, Siria y Afganistán, según documentos confidenciales que han sido revelados. Por ejemplo, en el año 2019, los estadounidenses utilizaron drones para reconocer y posteriormente bombardear una supuesta base enemiga. Este ataque no solo fue encubierto por los militares, sino que también mostró que, aunque estas tecnologías proporcionaban ventajas, eran muy imprecisas en la distinción entre civiles y combatientes. En ese ataque, murieron 64 mujeres y niños, y no se tiene certeza de que fueran militantes.<sup>31</sup> Este hecho puede considerarse uno de los primeros crímenes de guerra perpetrados con la ayuda de la IA. En Kabul, la capital de Afganistán, se han reportado ataques con drones desde principios de 2019 hasta finales de 2018. Esta situación es preocupante no solo por el uso de la tecnología, sino también por sus consecuencias. En este tipo de conflictos que han tenido lugar en Medio Oriente, las víctimas suelen ser mayoritariamente civiles, con un total de 1,417 muertes desde 2018.<sup>32</sup>

---

<sup>28</sup> <https://www.youtube.com/watch?v=gvozljfK-PiU&t=1s>

<sup>29</sup> <https://www.technologyreview.com/2022/07/18/1056059/robot-dog-ai-reinforcement/>

<sup>30</sup> <https://taskandpurpose.com/tech-tactics/robot-dogs/>

<sup>31</sup> <https://www.reuters.com/world/middle-east/us-military-hid-airstrikes-that-killed-dozens-civilians-syria-nyt-2021-11-13/>

<sup>32</sup> [https://policymemos.hks.harvard.edu/files/policymemos/files/1\\_18\\_21\\_nyt\\_hid\\_pent\\_rec\\_reveal\\_pat\\_of\\_fail\\_khan\\_azmat.pdf?m=1642604358](https://policymemos.hks.harvard.edu/files/policymemos/files/1_18_21_nyt_hid_pent_rec_reveal_pat_of_fail_khan_azmat.pdf?m=1642604358)

Además de los estadounidenses, otros países también están utilizando este tipo de tecnologías. En el año 2020, Azerbaiyán obtuvo una victoria significativa en la guerra contra Armenia, que se originó por la ocupación de los armenios en Nagorno y Karabaj después del conflicto previo que tuvo lugar entre 1992 y 1994. Esta victoria fue histórica para Azerbaiyán, un país con una población de 10 millones de personas y un presupuesto militar de 2 mil millones de dólares al año. Sin embargo, esta ventaja no se debió únicamente a la fuerza de su ejército o a su costoso armamento, sino que se logró gracias al uso masivo de drones turcos de bajo costo y sin necesidad de pilotos, conocidos como ‘Tb2’. A estos drones les añadieron 4 bombas ‘MAM-L’ de guía láser, lo que permitió a Azerbaiyán neutralizar las defensas aéreas enemigas y llevar a cabo ataques más profundos al destruir 144 tanques ‘T-72’, 310 vehículos blindados ligeros, 116 piezas de artillería, 34 objetivos en tierra e incluso interrumpir las comunicaciones enemiga.<sup>33</sup> Esta tecnología de bajo costo en comparación con los tanques y otros armamentos militares convencionales está transformando el poder militar, ya que no requiere una gran inversión financiera y permite que cualquier país desarrolle armamento basado en IA que les otorgue ventaja en el campo de batalla.<sup>34</sup>

Otro caso que ilustra esta tendencia es el de Israel, que en el año 2022 fue objeto de ataques terroristas por parte de los palestinos en sus principales ciudades, resultando en la muerte de 19 personas. En respuesta, Israel desplegó un enjambre de diversos drones de reconocimiento conectados entre sí, lo que les permitió detectar la ubicación de lanzacohetes y transmitir esta información a los misiles, logrando así destruir dichos objetivos.<sup>35</sup> La eficacia de estos enjambres es tal que en el 2022 murieron 146 palestinos, el año más violento desde el 2005.<sup>36</sup>

Así, aunque esta estrategia se mostró eficaz, también pone de relieve el dilema ético que surge cuando se busca desarrollar máquinas inteligentes que asuman funciones militares de manera autónoma, incluso tomando decisiones cruciales, como la de llevar a cabo bombardeos en contra de objetivos humanos específicos. Un ejemplo documentado de esta tendencia se dio en Libia en 2021, donde se reportó un ataque contra el Gobierno de Acuerdo Nacional reconocido por la ONU. El ataque fue realizado por un dron llamado ‘Kargu-2’, programado para atacar sin requerir conexión de datos entre el operador y la munición. En otras palabras, el sistema toma sus propias decisiones acerca de cuándo disparar, perseguir y localizar enemigos por medio de su algoritmo sin necesidad de que un humano se lo diga o se lo impida. El sistema puede usarse con la participación humana de igual manera y no tenemos certeza de cuál de los dos modos se usó, pero el simple hecho de que tenga la capacidad de actuar autónoma plantea serias preocupaciones sobre la toma de decisiones letales sin intervención humana.

El desarrollo de estas tecnologías crece a pasos tan agigantados que se ha manifestado en la guerra entre Rusia y Ucrania, donde se está llevando a cabo la primera guerra de drones a gran escala, un acontecimiento de importancia significativa, pero que no ha recibido la atención que podría esperarse. En esta guerra, drones son utilizados para lanzar bombas,

---

<sup>33</sup> <https://www.militarystrategymagazine.com/article/azerbaijans-victory-initial-thoughts-and-observations-and-caveats-for-the-innovative/>

<sup>34</sup> <https://www.forbes.com/sites/davidhambling/2020/11/10/the-magic-bullet-drones-behind--azerbaijans-victory-over-armenia/?sh=1963a87b5e57>

<sup>35</sup> <https://www.timesofisrael.com/in-apparent-world-first-idf-deployed-drone-swarms-in-gaza-fighting/>

<sup>36</sup> <https://foreignpolicy.com/2023/01/20/israel-palestine-west-bank-attack-surveillance-drones-far-right-settlers/>

destruir tanques, asesinar personas, neutralizar municiones e incluso derrotar otros drones. La magnitud de este uso se refleja en la abrumadora cantidad de drones abatidos en combate, aproximadamente 10,000 al mes, lo que equivale a más de 300 por día. No solo se están usando este tipo de armamento inteligente, sino que cada vez es a mayor escala y de forma más común.<sup>37</sup>

Al usar inteligencia artificial para el desarrollo explícito de armas y armamento militar, se intensifica y se posibilita la desigualdad y la posibilidad de que el poder y su estabilidad se vean alterados. Los países que alcancen estos avances de forma más rápida y efectiva serán los países con mayor influencia y poder social. Se tendrá la ventaja militar pues las máquinas han mostrado leer y reaccionar al ambiente 1000 veces más rápido que los seres humanos. Sin embargo, los militares parecen olvidar que no siempre la rapidez con que se haga algo se equipara a su efectividad ética y moral, su uso beneficioso y correcto, su confiabilidad e incluso su objetivo inicial de ser una ventaja para toda la humanidad. Por esto no podemos apartar el problema de la inteligencia artificial y mucho menos cuando, como vimos, estas máquinas son autónomas y no requieren intervención humana alguna. Militares están sacando provecho de estas tecnologías inteligentes al dotarles una intención específicamente dirigida al poder militar y a la neutralización de enemigos específicos. Las máquinas no tratan solo de ir de un punto A a un punto B, sino que lo hacen con un objetivo militar dirigido en mente.<sup>38</sup>

Estas preocupaciones se vuelven más alarmantes cuando consideramos que los pocos casos que han salido a la luz generan más incertidumbre que certezas. No tenemos información precisa sobre cuándo se utilizó por primera vez la inteligencia artificial en la historia, si fue probada antes de su implementación o si se tuvieron en cuenta principios éticos como evitar atentar contra personas indefensas. Tampoco tenemos certeza sobre cuándo y cómo surgirán o serán utilizadas nuevas tecnologías en el futuro. Esta falta de transparencia y claridad no nos permite confiar plenamente ni estar tranquilos, especialmente en un mundo en el que parecen resurgir conflictos bélicos en algunos países. Además, los casos específicos que hemos observado hasta ahora han planteado dilemas morales que debemos tener en cuenta al desarrollar la inteligencia artificial y al pensar en cómo implementar esta tecnología. Por ejemplo, los incidentes con el ejército estadounidense en Afganistán muestran que esta tecnología no es ni confiable ni éticamente efectiva. Desde 2018, cuando se confirmó el uso de inteligencia artificial en armamento en este conflicto, se han registrado 1417 bajas civiles, como mencionamos anteriormente.<sup>39</sup> Se esperaría que las máquinas que utilizan reconocimiento facial con inteligencia artificial para determinar a quién ejecutar y a quién no, tendrían una mayor efectividad en distinguir entre civiles y combatientes, pero lamentablemente no ha sido así. Incluso en el 62 % de estos casos se han involucrado niños, lo que demuestra que la distinción entre un niño indefenso y un combatiente no ha sido tan efectiva y no ha salvaguardado la vida de inocentes adecuadamente.<sup>40</sup> Estas tecnologías, por lo tanto, no están cumpliendo su objetivo de proteger a los más vulnerables y no están distinguiendo entre lo legal y lo moralmente correcto. La ineffectividad de estos softwares de

---

<sup>37</sup> <https://www.atlanticcouncil.org/blogs/ukrainealert/ukraine-strikes-back-against-russia-as-worlds-first-drone-war-escalates/>

<sup>38</sup> <https://www.hindawi.com/journals/ijis/2023/8676366/>

<sup>39</sup> <https://www.statista.com/statistics/917766/united-states-artificial-intelligence-application-use/>

<sup>40</sup> <https://www.nytimes.com/2021/11/03/us/politics/drone-strike-kabul-child.html>

reconocimiento ha sido tal que una IA estadounidense falló en reconocer a 28 miembros del Congreso y los confundió con criminales.<sup>41</sup>

Es importante considerar un escenario en el que la inteligencia artificial tome decisiones en situaciones éticamente complejas, como el ejemplo de una niña afgana obligada a colocar explosivos cerca de una base militar estadounidense. Si un soldado humano se enfrenta a este dilema, es probable que no asesine a la niña o, en caso de hacerlo, sería como último recurso, generándole estrés postraumático, remordimiento y dolor. Si una inteligencia artificial toma esta decisión, no tendría en cuenta las circunstancias humanas y calcularía estadísticamente que la mejor opción sería eliminar a la niña, sin importar si la forzó o si es indefensa. Por más efectiva que sea al neutralizar enemigos, estas máquinas no tienen la capacidad de hacer este tipo de distinciones en las que requieren no solo inteligencia sino aptitudes éticas. Aceptar sin cuestionar la implementación de tecnologías como estas podría llevar a situaciones legalmente correctas, pero claramente éticamente censurables. No podemos confiar en que las leyes resolverán los dilemas éticos asociados a estas tecnologías, ya que son innovaciones nuevas y complejas, y nadie las comprende plenamente. Por lo tanto, todos los involucrados en su desarrollo y uso deben reflexionar desde una perspectiva ética y moral que considere no solo los beneficios inmediatos de las tecnologías inteligentes, sino también sus posibles consecuencias éticas y morales a largo plazo. Si otorgamos autonomía a las máquinas para tomar decisiones en actos éticamente complejos sin considerar los riesgos éticos y su nivel de autonomía, permitiremos que en situaciones en las que debemos pensar éticamente, solo actuemos desde un enfoque estadístico y utilitarista, calculando cada acción solo según números y en el objetivo de obtener una ventaja militar.

Es precisamente debido a la incertidumbre e imprecisión asociadas a estas tecnologías que organizaciones como la ONU y el Departamento de Defensa de Estados Unidos han declarado que el uso de armas de inteligencia artificial es políticamente inaceptable y moralmente reprochable.<sup>42</sup> Imaginemos que esta posición no fuera así y que los enjambres de drones, como los vistos en el caso de Israel y la Franja de Gaza, estuvieran presentes en todos los conflictos militares actuales. Esto aumentaría la incertidumbre y la imprevisibilidad de las situaciones, y los posibles riesgos y su probabilidad de materializarse se incrementarían considerablemente. Pese a ello, la mayoría de los países desarrollados que trabajan en el desarrollo de estas tecnologías han rechazado su uso en combate. Rusia, por ejemplo, expresó explícitamente a la ONU que tratará de construir máquinas asesinas.<sup>43</sup> Por otro lado, China ha estado liderando la carrera en el desarrollo de estas tecnologías. Aunque el Departamento de Defensa de Estados Unidos advirtió sobre el uso de estas armas, los estadounidenses dejaron de preocuparse una vez que se dieron cuenta de que Rusia y China no tenían intenciones de frenar su desarrollo. Sin embargo, es importante resaltar que al menos Estados Unidos parece tener principios similares a los mencionados en este trabajo, lo cual es alentador en su postura frente a la implementación de estas tecnologías en el ámbito militar. Queda mucho por hacer para que estos principios se consideren y cumplan, especialmente si

---

<sup>41</sup> [https://elpais.com/tecnologia/2018/07/27/actualidad/1532683160\\_968681.html](https://elpais.com/tecnologia/2018/07/27/actualidad/1532683160_968681.html)

<sup>42</sup> <https://www.un.org/sg/en/content/sg/speeches/2023-07-18/secretary-generals-remarks-the-security-council-artificial-intelligence>

<sup>43</sup> <https://apnews.com/article/putin-russia-artificial-intelligence-3098b4f5205785f1b8281b34f13bff92>

consideramos que muchos casos en los que se han visto fallas en la inteligencia artificial los han encubierto las altas esferas de mando.

La efectividad demostrada por estas armas inteligentes, junto con su menor costo en comparación con los tanques y demás armamento militar convencional, ha llevado a que no solo los países desarrollados trabajen en estas tecnologías e implementen su uso, sino que también se abra la posibilidad de que países más pequeños o incluso grupos terroristas accedan a estas capacidades militares impresionantes. El caso de Azerbaiyán es un claro ejemplo de cómo estas tecnologías pueden cambiar el equilibrio de poder y fomentar más conflictos y menos resolución o tratados, ya que más actores tienen acceso a este poder militar. Si no dotamos el desarrollo y la implementación de estas tecnologías con un enfoque ético que abarque los derechos humanos, estaremos permitiendo que máquinas más efectivas, pero con menos aptitudes éticas que nosotros tomen decisiones sobre si asesinar o no a alguien, y que lo hagan sin intervención humana. Es fundamental que, como sociedad global, nos comprometamos a abordar de manera responsable y ética el desarrollo y uso de la inteligencia artificial militar, considerando sus implicaciones en seguridad, derechos humanos y valores éticos fundamentales.

El problema va más allá de simplemente considerar la autonomía, la agencialidad y el estatus moral de estas máquinas. También debemos reflexionar sobre el impacto que otorgarles estatus moral y autonomía para realizar acciones tan graves como el asesinato podría tener en nuestra propia valoración ética como seres humanos. Al permitir que máquinas sin ninguna aptitud ética tomen decisiones para quitar vidas humanas, corremos el riesgo de disminuir el valor de nuestro propio estatus moral. Estaríamos permitiendo que máquinas sean libres para matar solo porque lo hacen mejor que nosotros, y esto socavaría nuestros valores éticos fundamentales. Si aceptamos el uso de estas máquinas, no podemos permitir que la muerte y otros aspectos importantes desde el punto de vista ético se conviertan únicamente en cálculos estadísticos. Es crucial que, si decidimos continuar con el desarrollo de esta tecnología, lo hagamos de una manera que establezca límites y garantice que podemos confiar en su implementación, evitando así consecuencias nefastas para la humanidad. El caso de Israel refuerza la efectividad y el poder que se obtiene al implementar estas tecnologías en el conflicto. Solo en el año en que estas tecnologías comenzaron a utilizarse en el conflicto, el número de palestinos asesinados fue igual de elevado que en el año 2005, cuando hubo una amplia participación de tropas estadounidenses en el conflicto. Esto demuestra que estas tecnologías pueden tener un impacto significativo en los resultados de un conflicto. Sin embargo, centrarse únicamente en el poder militar inmediato que brinda esta tecnología nos hace pasar por alto los posibles riesgos y cómo la estabilidad global de los últimos 75 años podría estar en peligro. Es necesario tener en cuenta no solo las ventajas tácticas que estas tecnologías pueden ofrecer, sino también sus implicaciones éticas, legales y humanitarias a largo plazo. La implementación responsable de la inteligencia artificial en el ámbito militar debe ser un objetivo primordial para evitar consecuencias no deseadas y asegurar la protección de los derechos humanos y la paz mundial.

### **4.3 MegaSyn**

Los investigadores Fabio Urbina, Filippa Lentzos, Cédric Invernizzi y Sean Ekins han estado trabajando constantemente en la tecnología inteligente MegaSyn.<sup>44</sup> Esta tecnología utiliza el algoritmo de aprendizaje 'de novo', mediante el cual, en un espacio virtual simulado, la máquina aprende a modelar posibles moléculas que aún no son conocidas por los humanos debido a su alta complejidad de desarrollo y que, sin embargo, pueden ser útiles en el diseño de nuevos medicamentos. Con este modelo innovador, los científicos identificaron 200 moléculas que podrían contribuir al descubrimiento de nuevas medicinas. Paralelamente a este desarrollo, los investigadores fueron invitados por el laboratorio suizo Spiez a una conferencia para abordar los posibles perjuicios y riesgos asociados con estos avances y desarrollos.

En esta conferencia, con fines experimentales, los investigadores exploraron la eficacia de su desarrollo no para encontrar moléculas que ayuden en el tratamiento de enfermedades, sino para descubrir moléculas tóxicas. Para ello, dado que el mecanismo funciona con aprendizaje profundo, solo tuvieron que modificar el valor de penalización de la toxicidad de 0 a 1, y el valor de recompensa de la molécula sana de 1 a 0. Con este cambio específico en el programa, en menos de seis horas el sistema generó 40,000 moléculas tóxicas. Algunas de estas moléculas son las más tóxicas conocidas por la humanidad. La máquina incluso creó variantes de VX y Novicholks, y generó también moléculas totalmente desconocidas para nosotros que podrían ser aún más tóxicas que las mencionadas anteriormente.<sup>45</sup>

El desarrollo de tecnologías inteligentes no puede ser ingenuo y debe tener en cuenta, más allá de su desarrollo inicial y su intención, las posibles aplicaciones malintencionadas que puedan surgir. El caso de los investigadores y su tecnología MegaSyn ilustra claramente el alcance de esta preocupación. No solo se necesitó un Mac de escritorio para generar 40,000 moléculas tóxicas en seis horas, sino que también muestra cómo los estadounidenses solo se preocuparon por el desarrollo de estas tecnologías cuando posiblemente ya era demasiado tarde. Si solo hubiéramos considerado los casos de inteligencia artificial en el armamento militar, es probable que los estadounidenses no hubieran tomado medidas para implementar principios éticos en el desarrollo y uso de estas tecnologías. Fue la posibilidad de que otros países o incluso grupos terroristas pudieran acceder a esta capacidad en cuestión de horas con un computador de escritorio lo que llevó a una mayor conciencia sobre la necesidad de un desarrollo ético de estas tecnologías.<sup>46</sup>

Aunque es un avance que los estadounidenses hayan implementado principios éticos, no podemos ignorar que lo hicieron cuando sintieron que su soberanía en el desarrollo de estas tecnologías podría verse amenazada y que otros actores también podrían adquirir este poder. Esto plantea la pregunta de si podemos confiar en los principios implementados por un país que ha fomentado el uso desigual de tecnologías inteligentes en su poder militar y científico hasta que se vio amenazado. Es posible que las intenciones de científicos, matemáticos e

---

<sup>44</sup> <https://pubmed.ncbi.nlm.nih.gov/35694522/>

<sup>45</sup> <https://www.wired.com/story/ai-dr-evil-drug-discovery/>

<sup>46</sup> [https://www.ai.mil/blog\\_02\\_26\\_21-ai\\_ethics\\_principles-highlighting\\_the\\_progress\\_and\\_future\\_of\\_responsible\\_ai.html#:~:text=in%20the%20DoD-.AI%20Ethical%20Principles%20%E2%80%93%20Highlighting%20the%20Progress%20and%20Future.Responsible%20AI%20in%20the%20DoD&text=On%20February%2024%2C%202020%2C%20the.by%20the%20Defense%20Innovation%20Board.](https://www.ai.mil/blog_02_26_21-ai_ethics_principles-highlighting_the_progress_and_future_of_responsible_ai.html#:~:text=in%20the%20DoD-.AI%20Ethical%20Principles%20%E2%80%93%20Highlighting%20the%20Progress%20and%20Future.Responsible%20AI%20in%20the%20DoD&text=On%20February%2024%2C%202020%2C%20the.by%20the%20Defense%20Innovation%20Board.)

investigadores sean siempre buenas al desarrollar tecnologías inteligentes. Lo mismo sucedió con la energía nuclear, que inicialmente se descubrió con el objetivo de encontrar nuevas fuentes de energía. Sin embargo, es fundamental reconocer que la implementación responsable de estas tecnologías es crucial para evitar consecuencias indeseables y garantizar que su desarrollo beneficie a toda la humanidad en lugar de generar riesgos y desigualdades.

Continuar con el desarrollo de tecnologías inteligentes conlleva la responsabilidad de considerar no solo sus beneficios potenciales, sino también sus riesgos y posibles repercusiones negativas. La historia nos ha enseñado que cada avance tecnológico significativo puede utilizarse para el bien y el mal, y la inteligencia artificial no es una excepción. Es esencial que, al desarrollar estas tecnologías, se preste atención a la ética y se establezcan principios sólidos que guíen su implementación. Esto implica no solo la creación de mecanismos que aseguren un uso responsable y seguro, sino también la consideración de cómo estas tecnologías pueden afectar la sociedad, las relaciones internacionales y la seguridad global. Además, el interés por los estadounidenses en la aplicación de estas limitaciones se vio contradicha una vez que China y Rusia no optaron por este camino.

No podemos depender únicamente de la autorregulación de los países o entidades individuales en el desarrollo y uso de la inteligencia artificial. Se necesita una colaboración internacional más amplia y una regulación efectiva para garantizar que estas tecnologías se utilicen de manera ética y responsable. La inteligencia artificial puede mejorar nuestras vidas de manera diferente, desde el diagnóstico y tratamiento médico hasta la optimización de la producción intelectual y la neutralización de objetivos militares. Sin embargo, no debemos ignorar los posibles peligros y dilemas éticos que conlleva su uso en contextos militares y de seguridad.

Como sociedad, debemos adoptar un enfoque prudente y reflexivo hacia el desarrollo de tecnologías inteligentes. Esto implica promover una mayor conciencia pública sobre los avances en inteligencia artificial y sus implicaciones éticas, y alentar un diálogo abierto y transparente entre expertos, gobiernos y la sociedad civil. Solo mediante una cuidadosa consideración de los aspectos éticos y morales podremos aprovechar plenamente el potencial de la inteligencia artificial para el bienestar humano y asegurar que no se convierta en una herramienta para la destrucción y la desigualdad. La tecnología inteligente debe estar al servicio de toda la humanidad y ser una fuerza positiva que promueva la paz, la justicia y el progreso en el mundo.

## **5. Implicaciones Filosóficas**

### **5.1 Motivaciones**

Al desarrollar las tres situaciones anteriores, se nos desvelan dilemas éticos de diversas maneras. A continuación, hablaré de cada una de ellas en particular para dotar de un análisis más específico a la necesidad de abordar la IA desde un enfoque ético y explorar las implicaciones filosóficas que esto pueda conllevar.

#### **5.1.1 Accidentes**

A la hora de implementar Chat GPT para que corrigiera mi tesis tuve que ser extremadamente cuidadoso. Por eso, todo lo expuesto en este trabajo de investigación se escribió antes a mano. Esta fue la manera más convincente para mí de asegurarme de que la máquina se

limitara a realizar correcciones de estilo, ya que no está exenta de accidentes. En varias ocasiones el Chat GPT excedió su capacidad requerida y no se limitó a la corrección de estilo, ortografía y puntuación, sino que a veces tomaba confianza y añadía nuevos párrafos como si el texto fuera de su autoría. No podemos pretender que solo con que le dotemos indicaciones claras y concisas a estas máquinas estas no puedan fallar. Además, no todos tenemos entrenamiento en programación y dotación de entradas para inteligencias artificiales, por lo que tampoco podemos confiar ciegamente en que nuestra indicación no está exenta de errores. Incluso, como el objetivo último de la máquina es mostrar inteligencia, podría hacerlo aún sin tener en cuenta lo específico de las indicaciones o los derechos de autor, entre muchos otros aspectos.

En el ámbito militar, los accidentes adquieren una tonalidad más oscura, ya que no solo pueden afectar nuestra propiedad intelectual, sino que también podrían ocasionar la pérdida de vidas humanas inocentes. Las máquinas inteligentes utilizadas para llevar a cabo objetivos militares han demostrado no tener la capacidad de distinguir entre personas armadas e indefensas, ni tampoco consideran el concepto de 'hors de combat', llegando incluso a asesinar a tantos civiles como tropas estadounidenses durante este conflicto en el 2005. Con esto no quiero minimizar la importancia de la pérdida de propiedad intelectual y otros problemas que puedan ocurrir accidentalmente con el Chat GPT. Un plagio, sea accidental o no, puede llevarnos incluso a años en la cárcel. Pero el problema se intensifica en el ámbito militar porque se trata de quitar vidas humanas inocentes o indefensas. Los casos han mostrado cómo muchas personas, incluso niños, están sufriendo debido a estos efectos secundarios.

Volviendo al tercer dilema, podemos observar no solo la posibilidad de accidentes, sino también la necesidad de considerar estas posibilidades incluso antes de la implementación y uso de estas tecnologías. Los investigadores desarrollaron este tipo de sistemas desde mayo del 2022, pero no fue hasta la conferencia Spiez en septiembre del 2022 que se dieron cuenta de los posibles perjuicios que podrían surgir si se usaba de manera malintencionada. Puede que este sistema no nos haya perjudicado de manera accidental, pero era un accidente prevenible. Si los investigadores hubieran centrado sus esfuerzos no solo en innovar y desarrollar avances en la medicina, sino también en los posibles riesgos asociados, se podría haber limitado su funcionamiento o tomado medidas para prevenir estos riesgos antes de su lanzamiento. El problema se intensifica porque, ya sea accidental o no, a través de este software es incluso más fácil sintetizar o extraer toxinas que componer o producir las complejas moléculas químicas que pueden generar medicamentos. Mientras que para el momento de la publicación del primer artículo solo se habían encontrado 600 moléculas benéficas, ya había 40,000 toxinas sintetizadas por la máquina en seis horas.

En los tres casos, pudimos observar que un diseño, desarrollo e implementación apresurados y poco considerados pueden llevar a avances beneficiosos en la IA, pero también a accidentes prevenibles que incluso pueden ser mortales. Si no comenzamos a pensar en la IA como algo más que una herramienta beneficiosa y efectiva para los humanos, estaremos ignorando los riesgos accidentales que podrían ocasionar perjuicios.

### **5.1.2 Maleficencia**

A pesar de que inicialmente solo veamos accidentes como posibles consecuencias, es crucial considerar que los perjuicios ocasionados a través de MegaSyn podrían dejar de ser accidentales y ser provocados de manera malintencionada por seres humanos que empleen estas tecnologías de forma maliciosa. El hecho de que solo se necesite un computador de escritorio para desarrollar toxinas a través de MegaSyn es preocupante. Muchas de las toxinas

desarrolladas están relacionadas con aquellas que se utilizan actualmente en atentados y crímenes biológicos, como el VX. No podemos evitar pensar que grupos terroristas o personas malintencionadas podrían estar interesados en aplicar estas tecnologías de una forma no beneficiosa. Debemos contemplar las posibilidades que estos desarrollos generan, tanto positivas como negativas. No podemos seguir desarrollando estas tecnologías de forma ingenua creyendo que solo sucederán cosas buenas y que todas las posibilidades serán éticamente aceptables.

Si volvemos nuestra mirada hacia el caso de ChatGPT y su posible uso malintencionado, podremos desarrollar esta idea con un enfoque más profundo en las implicaciones filosóficas que pueda conllevar. En primer lugar, a pesar de que los bots ya no copien y peguen texto sin más para aparentar inteligencia, no podemos olvidar que todo proceso estadístico basado en algoritmos y probabilidad tendrá siempre un margen de error. Así, aunque los diseñadores de este software han garantizado que este bot y otros similares no incurrirán en plagio, lo han hecho considerando que, si el 20% del texto provisto es plagiado, esto es aceptable para una máquina. En el caso específico de ChatGPT se trata de un 5% de margen de error, por lo que por cada cien oraciones que produzca, alrededor de cinco serán plagiadas. En el fondo, los diseñadores están analizando este problema del plagio de una forma que excluye por completo la ética, ya que, desde un punto de vista moral, aceptar actos éticamente reprochables en las máquinas y en su uso, aunque sean en menor grado, no es aceptable. Este tipo de márgenes de error nos muestran que el análisis del alcance y los resultados de las máquinas inteligentes está estrechamente ligado a porcentajes y cálculos, dejando de lado el hecho de que la apropiación de contenido intelectual no es aceptable bajo ninguna circunstancia. Ahora, imaginemos que los estudiantes usen estos softwares para plagiar o producir sus asignaturas. En primera instancia, al menos un porcentaje del texto será plagio de otro plagio, lo que resta credibilidad y valor intelectual. Además, los detectores que se han creado para combatir esto están incompletos y solo cubren textos en inglés. Alguien podría incluso preferir hacer todo en español y luego traducirlo con otra máquina inteligente al inglés para evadir el plagio. Lo cierto es que personas inescrupulosas que no se tomen en serio la producción intelectual podrían estar afectando a la producción intelectual. Siguiendo datos estadísticos, aproximadamente dos tercios de los estudiantes cometen plagio en el colegio, y un tercio de los estudiantes en la universidad también lo hace. Dada la efectividad y rapidez con la que ChatGPT realiza tareas es muy probable que este porcentaje de estudiantes esté utilizando este tipo de tecnologías, e incluso que esté ocasionando que este porcentaje aumente.

Lamentablemente, la misma posibilidad se está generando en el caso del armamento militar donde hablamos de tecnologías muy baratas y rápidas de producir. Si no tomamos las precauciones adecuadas este tipo de tecnologías también podrían caer en manos de grupos terroristas o agentes malintencionados. Incluso los países podrían usarlas para ganar guerras injustamente como ya hemos visto previamente y ocultar sus errores. Debemos considerar de manera más crítica el hecho de que estas tecnologías ya pueden asesinar personas de forma autónoma sin que ningún humano tenga conectividad con la munición de la máquina, y cuestionarnos si no es mejor frenar este desarrollo por completo.

### **5.1.3 Desigualdad**

Si nos enfocamos en el carácter desigual que ha demostrado poseer y potenciar este desarrollo concluimos rápidamente que debemos considerar esta cuestión tanto antes como después de que estas tecnologías sean diseñadas e implementadas. En el aspecto militar, por ejemplo, ocurre lo mismo que con el desarrollo general de la IA. Solo los países de primer mundo y algunos países apoyados por estos están implementando estas tecnologías y

beneficiándose con ellas. Esto genera un desequilibrio en la estabilidad social mundial, ya que en primera instancia se llevan a cabo más conflictos debido a la efectividad y bajo costo de estas tecnologías. Además, estas guerras han demostrado que aquellos con acceso a la IA podrían convertirse en los nuevos poderes mundiales en el futuro de manera similar a como los desarrolladores de la bomba nuclear lo fueron en el pasado. Es innegable que un gran armamento militar suele asociarse con la soberanía. La poca claridad y transparencia en estos desarrollos aumenta la desigualdad, ya que solo unos pocos están siendo involucrados en estos proyectos, que generalmente son confidenciales y restringidos. Esto puede prevenir que caigan en manos equivocadas, pero también facilita un desarrollo éticamente cuestionable que se oculta intencionalmente y beneficia solo a unos pocos con estas tecnologías.

Históricamente, gran parte del conocimiento y progreso científico ha sido divulgado y compartido de manera pública y abierta para todos. También ha habido empresas y actores privados que, para lucrarse o por temas de seguridad nacional, han patentado o restringido el acceso público a estas tecnologías. Esto es lo que han intentado hacer los estadounidenses con avances como MegaSyn, ya que han visto que divulgar de forma pública este avance podría llevar a riesgos para la humanidad irreversibles. Esto genera una tensión entre lo público del conocimiento y los riesgos éticos que esto puede acarrear, ya que se está violando el derecho al libre acceso a los hechos y a cualquier tipo de conocimiento. Además, aunque sea con la expresa intención de prevenir riesgos y usos malintencionados, esto a su vez genera más desigualdad en cuanto a la participación en el diseño e implementación de estas tecnologías. Esto puede aumentar el riesgo al usar estas tecnologías y provocar consecuencias éticamente cuestionables que no fueron previstas debido a la falta de diversidad y participación en el desarrollo.

Discutiendo con un gran amigo abogado, él me decía que incluso los abogados están utilizando procesadores de texto como GPT para la corrección de sus sentencias y otras labores. Lo cierto es que más de 1 billón de personas son usuarios activos de ChatGPT, es decir, un octavo del mundo aproximadamente interactúa con estas tecnologías. Incluso con este gran número de usuarios, no podemos dejar de lado ciertos criterios que nos muestran cómo ChatGPT es desigual y podría serlo aún más. En primer lugar, ChatGPT requiere un conocimiento básico de tecnología, indicadores y programación, educación que muchas personas carecen. Adicionalmente, ya hay una versión paga y mejorada del chatbot que cuesta 100 dólares al mes, lo que resulta costoso e inasequible para la población más pobre, fomentando la segregación y división de beneficios basados en el ingreso monetario. Además, debemos tener en cuenta la capacidad de ChatGPT de ser una suerte de experimento para desarrollos posteriores. Aunque su "programación garantiza la privacidad y la confidencialidad de todas las interacciones con los usuarios" (Open AI, 2021), también es cierto que los datos recopilados están siendo utilizados para el desarrollo de tecnologías más avanzadas que podrían llegar a costar y segregar aún más. ChatGPT es solo el conjunto de entrenamiento para tecnologías más complejas que pueden generar una mayor desigualdad. El conocimiento que adquirió ChatGPT en su entrenamiento no es suficiente para que estas máquinas puedan captar todos los matices del lenguaje y su uso dependiendo del contexto. Esto puede llevar a malinterpretaciones que generen confusiones o incluso a una desigualdad que favorezca ciertos usos del lenguaje por parte de los creadores de estas tecnologías.

#### **5.1.4 Sesgos**

Tampoco podemos ignorar la cuestión de si los datos privados y públicos están siendo transformados de manera que requiera una nueva consideración para la correcta implementación de la inteligencia artificial. Es un motivo adicional para reflexionar sobre la IA desde un enfoque ético. Deberíamos establecer límites para regular las máquinas

inteligentes, ya que no podemos entregarles ciegamente nuestros datos personales y confiar en que, a pesar del uso dual que puede tener la inteligencia artificial, siempre optará por la privacidad y la protección de datos. Ha habido casos en los que actores privados como Facebook, Google y otros no han tenido en cuenta adecuadamente estos aspectos. Debemos considerar la posibilidad de que nuestros datos sean robados por agentes externos. También es importante mencionar que el conjunto de datos que contiene ChatGPT está desactualizado, ya que solo incluye datos hasta septiembre de 2021 y desconocemos los criterios de selección y su propósito.

No debemos confiar ciegamente en la selección, implementación y actualización de conjuntos de datos. A lo largo de este trabajo, hemos visto varios casos en los que este proceso fue inadecuado y replicó sesgos y discriminación. Lo mismo podría estar sucediendo con ChatGPT, e incluso su versión premium mostró un sesgo hacia las mujeres en la ingeniería<sup>47</sup>. Incluso los creadores de ChatGPT pueden tener sesgos, y si no los tenemos en cuenta antes, durante y después de que nos empiece a afectar, podrían reflejarse en el bot. Lo mismo ocurre con tecnologías militares, ya que, aunque sean creadas por militares bajo la supervisión de algún gobierno, hemos visto cómo pueden fallar y matar a civiles e incluso a niños indefensos. Es muy probable que se estén correlacionando datos sin que exista una causalidad, lo que podría llevar a que las máquinas lleguen a conclusiones sesgadas. Por ejemplo, los drones implementados para reconocimiento facial en el ejército estadounidense identificaron erróneamente a 28 miembros del congreso como criminales peligrosos y sujetos a neutralización<sup>48</sup>. Las máquinas inteligentes han demostrado su poca efectividad al reconocer efectivamente a delincuentes sin caer en falsos prejuicios o identificar erróneamente a civiles como delincuentes. Debemos cuestionar esto sin ignorar los riesgos que implica si una máquina falla en esta tarea y, aun así, se le da el uso que las tropas militares le están dando.

En el caso de MegaSyn, todo se torna un poco más complicado ya que los datos utilizados en el desarrollo del software provienen de bases de datos médicas que contienen información sobre moléculas similares a las que los investigadores estaban intentando desarrollar para crear nuevas medicinas. Al ser datos científicos que recopilan información sobre la interacción de estas moléculas con el mundo físico y su desarrollo en laboratorios, es posible que no estén sesgados. Puede que permitan la creación de toxinas similares a las utilizadas en atentados biológicos porque se basan en lo que ya existe en el mundo. En este sentido, frenar el progreso es casi imposible. Por no haber prevenido esta tensión entre el conocimiento público y su posible uso malintencionado se posibilitó esta situación.

### **5.1.5 Sustitución**

El potencial que han demostrado las máquinas inteligentes nos lleva a considerar la posibilidad de que en un futuro cercano estas tecnologías nos sustituyan en áreas completas de trabajo que requieran funciones cognitivas. Desde hace años, las máquinas ya nos superan en efectividad, conexión y comunicación. Actualmente, ChatGPT se considera una herramienta que ahorra tiempo y facilita la investigación. Sin embargo, puede que llegue el momento en el cual no solo recurramos a ChatGPT o tecnologías similares para crear o corregir contenido intelectual, sino que se convierta en la única manera en que nosotros, como seres humanos, podamos lograr esto. El uso excesivo de este tipo de inteligencias

---

<sup>47</sup> [https://www.equalitynow.org/news\\_and\\_insights/chatgpt-4-reinforces-sexist-stereotypes/](https://www.equalitynow.org/news_and_insights/chatgpt-4-reinforces-sexist-stereotypes/)

<sup>48</sup> <https://www.nytimes.com/2018/07/26/technology/amazon-aclu-facial-recognition-congress.html>

artificiales podría llevarnos a perder gradualmente nuestras capacidades cognitivas y a depender de estas tecnologías para realizar tareas que antes realizábamos de manera independiente, como la corrección de estilo, la referencia de bibliografía y la producción de propiedad intelectual. Es probable que ya haya estudiantes en primeros semestres de universidad o comenzando la educación secundaria que, en lugar de prestar atención en clase para aprender a escribir, redactar e investigar, dirijan su atención exclusivamente a utilizar estas tecnologías para cumplir sus tareas rápidamente a través de bots como ChatGPT. Al utilizar estas tecnologías, debemos cuestionarnos si queremos que nos reemplacen incluso en la capacidad de escribir y producir conocimiento valioso, o si preferimos limitar su uso y generar una relación simbiótica en la que nos beneficiamos, y las máquinas nos ayuden a mejorar nuestras propias capacidades en vez de sustituirlas.

Lo mismo debemos preguntarnos en el ámbito militar, donde los drones ya comienzan a sustituir a los militares en conflictos a gran escala. Aunque estas tecnologías pueden disminuir el número de bajas humanas y enfermedades físicas y mentales, debemos recordar que 27.671.000 personas están involucradas en el ámbito militar y gran parte de este personal podría enfrentarse a la sustitución en un futuro próximo. El uso masivo de estas tecnologías en conflictos como el de Rusia y Ucrania nos muestra que esta es una posibilidad no muy lejana. Si bien es importante evitar situaciones perjudiciales para los soldados, también debemos garantizarles la posibilidad de desarrollarse en otras áreas y encontrar bienestar en otras actividades. El objetivo final de la IA debe ser siempre el bienestar humano, y al considerar la sustitución en el ámbito militar, debemos asegurarnos de que no solo salvaremos muchas vidas de soldados, sino que también garantizaremos que estas máquinas no lastimen a seres humanos ni les quiten la vida.

MegaSyn es otro caso que nos lleva a considerar que las máquinas pueden llegar a ser mejores médicos que nosotros. Como hemos visto, la tecnología inteligente ha demostrado ser más competente que los humanos en actividades cognitivas complejas y especializadas, como la medicina, la conducción y las finanzas, entre otras. Sin embargo, cuando se trata de actividades tan complejas, solo podemos aceptar e implementar esta sustitución si las máquinas demuestran ser superiores en todas las áreas de la medicina y hacen todo mejor que nosotros. Hasta que esto sea una realidad, la relación simbiótica entre humanidad y tecnología debe prevalecer. Además, cualquier análisis previo debe garantizar que, al igual que en el ámbito militar, si se produce esta sustitución, los médicos del mundo no perderán su motivación y su capacidad de encontrar florecimiento y bienestar en sus profesiones.

## **5.2 Implicaciones**

Las motivaciones que nos llevan a reflexionar sobre los avances en el campo de la inteligencia artificial en los últimos años son más que suficientes para considerar su diseño, implementación y uso desde una perspectiva ética que contemple tanto los aspectos beneficiosos para el progreso de la humanidad, como los posibles riesgos que podrían llevar al deterioro o incluso al fin de la humanidad. A continuación, analizaré y reflexionaré sobre cómo estas motivaciones y las tensiones que generan nos obligan a cuestionar los principios éticos necesarios para evitar escenarios negativos y cómo estos principios están siendo remodelados por el contexto social actual.

### **5.2.1 Justicia**

La inteligencia artificial ha demostrado una marcada desigualdad en su implementación, diseño y desarrollo. En primer lugar, su avance se concentra en muy pocos países, y son aún

menos los que consideran sus implicaciones éticas y filosóficas generándose poca cohesión social. El problema subyace en los centros de desarrollo de estas tecnologías, donde la participación activa es escasa. Es imprescindible fomentar una diversificación en el campo de la IA para evitar perspectivas limitadas que no contemplen los riesgos emergentes. Por ello, es crucial también adoptar un enfoque ético que permita evaluar los beneficios y peligros potenciales antes y después de su impacto en la sociedad, incluso anticipando nuevas amenazas no previstas. La IA se ha convertido en una herramienta omnipresente que permea la tecnología, lo social y lo político, y podría llegar a ser inherente a la humanidad. Esto plantea la necesidad de involucrar activamente a toda la comunidad, mediante consultas y retroalimentación, para una comprensión más profunda de sus implicaciones y preocupaciones. Esta inclusión equitativa también debe garantizar un acceso igualitario a los datos y recursos utilizados en la IA, evitando injusticias en su desarrollo y aplicación.

El ejemplo de Megasyn es altamente revelador en relación con este dilema. Recordemos brevemente que solo cuando Spiez lo invitó a una conferencia sobre los riesgos de estas tecnologías, los investigadores comenzaron a reflexionar sobre los posibles riesgos de estos desarrollos. Sin embargo, esta conferencia nunca llegó a realizarse debido a la pandemia, lo que retrasó aún más los resultados esperados. De no ser así, es probable que los aspectos negativos de esta IA pasaran inadvertidos hasta que fuera tarde para prevenirlos. Esto evidencia la escasa y desigual participación en tecnologías inteligentes, ya que no hemos podido considerar los riesgos más allá del beneficio inmediato. Es fundamental garantizar un pensamiento más justo y equitativo, donde cualquier persona interesada pueda participar en la discusión de estas cuestiones. No podemos esperar a que ocurra algo catastrófico para cuestionar las posibilidades y alcances de la IA. Si no actuamos desde el presente para abordar estas posibilidades como reales, podríamos enfrentar perjuicios a largo plazo debido a nuestra estrecha dependencia y confianza en estas tecnologías.

En cuanto al conocimiento científico y su divulgación pública, es necesario replantearse su enfoque. La mera divulgación pública del conocimiento y funcionamiento de ciertas tecnologías, como en el caso de MegaSyn, puede no ser la mejor solución. Los investigadores, conscientes de las consecuencias negativas, optaron por restringir la transparencia, lo que ha generado incertidumbre acerca de este tema. Debemos reconsiderar el acceso al conocimiento desde una perspectiva ética y filosófica, evaluando cuidadosamente el valor universal de este derecho y sopesando la posibilidad de no revelar ciertos funcionamientos o algoritmos debido a su potencial uso dual.

Los casos expuestos por medio de ChatGPT evidencian cómo estos bots pueden ser efectivos, pero también excluyentes, limitando su uso a ciertos juegos del lenguaje y excluyendo otros. El uso de estas tecnologías ha evolucionado rápidamente, afectando nuestras capacidades cognitivas y volviéndose esencial para la humanidad. Es crucial considerar conjuntamente las posibilidades riesgosas en su implementación para evitar problemas y asegurar que sean utilizadas de manera justa. También es necesario realizar un análisis conjunto de los datos utilizados en ChatGPT para garantizar su validez y veracidad, dado que es una plataforma de libre acceso que recopila y analiza datos a través de nuestras consultas. La cuestión de la justicia y la protección de datos en estas tecnologías debe ser abordada con participación abierta e involucrar a todos por igual.

La justicia enfrenta su mayor dilema en el contexto actual, al considerar el impacto de la inteligencia artificial en el armamento militar y el campo de batalla. Los países que antes mostraron interés en el desarrollo de armamento nuclear son ahora los principales y casi exclusivos autores del diseño y desarrollo de estas tecnologías. Aunque algunos países puedan compartir estas tecnologías con naciones en desarrollo como Ucrania y Azerbaiyán,

el problema no se limita a la falta de participación de países, sino también a la necesidad de un enfoque más holístico que no se centre únicamente en la efectividad militar de las máquinas. Si no aumentamos la participación y las consideraciones éticas en el diseño e implementación de estas tecnologías, el desequilibrio y el poder que traen consigo pueden ser más significativos de lo que imaginamos. No son posibles malos usos por parte de terceros los que están llevando a perjuicios, sino los diseñadores que ocultan esta realidad mientras siguen usando esta tecnología como si nada.

Desde la época de Aristóteles y Platón, la justicia se considera la virtud ética que busca el equilibrio y la proporción en las relaciones humanas, como afirmaría Aristóteles, la base de una sociedad virtuosa. En la filosofía moderna, la interpretación más conocida de la justicia proviene de Kant, quien nos insta a tratar a todos los humanos como fines en sí mismos y no como simples medios. En la filosofía contemporánea, hemos dotado a los conceptos de justicia y equidad de una dependencia contextual, entendiendo que ya no son nociones objetivas, sino que dependen del contexto social en particular. En el contexto actual, hemos sido testigos de cómo la inteligencia artificial está moldeando nuestra percepción y experiencia en el entorno cognitivo, ético, político y social. Sin embargo, es preocupante que la participación en el diseño y desarrollo de estas tecnologías esté siendo reducida y limitada. Para lograr un desarrollo más justo y equitativo de la IA, es fundamental fomentar un enfoque holístico que reúna a investigadores de diversos campos, a gobiernos de todas las naciones y a cualquier persona que desee contribuir y promover un enfoque justo de la inteligencia artificial. Hace algunos años, el acceso público y universal a los avances, como los nucleares, se consideraba un deber moral absoluto, priorizando el acceso a la verdad y lo público. Actualmente, algunos investigadores han dado mayor importancia a la confidencialidad de ciertos algoritmos, lo que plantea preocupaciones éticas y sociales. El caso del uso militar de la inteligencia artificial es especialmente inquietante, ya que existe ocultamiento y contradicciones en cuanto a los principios éticos que se defienden para el desarrollo de la IA. Es esencial cuestionar no solo el posible uso dual de estas tecnologías, sino también la doble moral de aquellos involucrados en su desarrollo e implementación injusta.

Debemos reflexionar acerca de si el desarrollo y las consecuencias de la inteligencia artificial son justos o si contribuyen a la inequidad y el desequilibrio del poder. También debemos considerar el derecho al conocimiento y si su divulgación debe ser pública, así como evaluar si existen casos en los que se deban restringir ciertos conocimientos. Al margen de las respuestas a estos cuestionamientos, no podemos permitir que los riesgos asociados a la IA se aborden inequitativamente, prohibiendo su uso mientras sigue generando estragos en diferentes partes del mundo. Este enfoque es contrario a tratar a todos los seres humanos como fines en sí mismos y conlleva a concebir a los enemigos de conflictos como meros medios para que máquinas lleven a cabo sus objetivos, incluso llegando a consecuencias trágicas como la pérdida de vidas humanas. Es crucial abogar por un desarrollo ético y equitativo de la inteligencia artificial, donde la participación y consideración de todas las voces sea primordial en la búsqueda de un futuro más justo y humano.

### **5.2.2 Beneficencia**

Es por esa característica desigual y por los posibles riesgos que pueda ocasionar una motivación malintencionada que debemos pensar siempre la motivación ulterior detrás del desarrollo y la implementación de cada avance tecnológico en particular. Su implementación debe ser beneficiosa para la humanidad y no debe traer consecuencias éticamente reprochables. Si seguimos construyendo estas tecnologías de manera ingenua sin considerar sus implicaciones más allá de lo que nosotros buscamos conseguir, puede que su

implementación logre ese beneficio inmediato que perseguía, pero a su vez traiga muchas aplicaciones secundarias que junto con la mala intención de algunos autores pueda ocasionar daño y perjuicios.

Si volvemos nuestra mirada hacia el caso de ChatGPT y su posible uso malintencionado nuevamente, podremos desarrollar esta idea un poco más junto con las implicaciones filosóficas que pueda conllevar. En primer lugar, por más de que los bots ya no copien y peguen texto sin más para lograr aparentar inteligencia, no podemos olvidar que todo proceso estadístico basado en algoritmos y probabilidad tendrá siempre un margen de error. Recordemos que aproximadamente por cada cien oraciones que produzca, alrededor de cinco serán plagiadas. En el fondo, los diseñadores están analizando este problema del plagio no desde un enfoque ético, pues la apropiación de contenido como propio, así sea a través de una máquina y su generador de texto, es algo que no se debe aceptar ni en un 20% o menos. Así, este tipo de márgenes de error nos muestran que el análisis está siendo estrechamente ligado a porcentajes y cálculos, dejando de lado el hecho de que, sin importar el porcentaje, la apropiación de contenido intelectual no es aceptable. Lo mismo ocurre en el caso en el que autos autónomos se encuentren en accidentes. La vida humana es valorada únicamente como un dato más en un proceso estadístico.

Ahora imagine que los estudiantes en efecto usen softwares como ChatGPT para plagiar o producir sus asignaturas. En primera instancia, al menos en un porcentaje del texto, se tratará del plagio de un plagio, restándole así credibilidad y valor intelectual. Para evitar esto, se han creado incluso aplicaciones que detectan la diferencia entre un humano y una IA a la hora de escribir. Esto solo abre más cuestiones en el asunto ya que se trata de otra máquina que tiene margen de error y representa solo al porcentaje de usuarios que se comunica en inglés con la máquina. Lo cierto es que la producción intelectual puede verse afectada por personas inescrupulosas que no se tomen en serio el conocimiento y que, por ahorrarse unas horas, prefieran optar por el camino fácil de producir textos con estos softwares en cuestión de minutos. Debemos limitar la posibilidad de estos usos malintencionados y hallar maneras de garantizar que no se proceda de una manera en que el uso de ChatGPT nos afecte a corto o largo plazo. Tanto es el alcance de esta idea que los procesadores de texto no son las únicas tecnologías que debemos cuestionar. Nos encontramos también con asistentes virtuales, generadores de música, generadores de arte, diseñadores, entre otros muchos que pueden sustituirnos en capacidades cognitivas complejas.

A la misma conclusión llegamos si pensamos en lo que podría pasar con softwares como Megasyn. Este tipo de tecnologías no solo permite esta posibilidad por su carácter dual, sino que además es más fácil para la máquina el desarrollo que no es beneficioso para la humanidad. Si seguimos desarrollando con fe ciega estas tecnologías puede llegar el momento en el que posibilitemos un avance tecnológico con demasiado beneficio pero que a su vez posibilite su contrario en cuanto al posible perjuicio. Siendo realistas, sin embargo, es muy improbable que progresos como estos vean un freno en su desarrollo. Por eso debemos pensar en este y en su futura implementación desde un enfoque ético que nos permita ver lo beneficioso y lo perjudicial que pueden traer estos avances simultáneamente. Hay quienes sin embargo pensarán que esto sucede con cualquier herramienta, incluso con un martillo, una roca, un cuchillo de cocina entre otros. Pero, aunque Megasyn comparta esta característica y no nos haga cuestionar esto, el caso del armamento militar es completamente diferente y nos hace cuestionar incluso la autonomía que están teniendo algunas máquinas y su legitimidad. El problema con este tipo de ejemplos que ya vimos previamente es que las máquinas tienen la posibilidad de acabar con la vida humana sin la necesidad de que un ser humano le esté dando esa motivación de forma directa y por medio de controles o acciones. Los drones

militares pueden calcular el entorno con softwares de reconocimiento, y sin que un humano pueda impedirlo al no tener conexión directa, pueden decidir si bombardear, neutralizar, chocarse contra un objetivo dirigido o alguna otra decisión que implica la toma de vidas.

Esto, junto con los altos márgenes de error que han demostrado las máquinas, ya han mostrado su inequidad y su rango de destrucción dejándonos ver que son un peligro por la falta de claridad y por su accidentalidad mortal. Además, por más que se salven militares a través de estas tecnologías, esta no es más que una solución aparente, pues no está solucionando ni dando fin a ningún conflicto. Lo que sí hace es fomentar la guerra, pero esta vez mediante máquinas inteligentes que pueden llegar a ser más efectivas que nosotros mismos al quitar una vida. Imaginemos que incluso estas tecnologías caen en manos de grupos terroristas o países con deseo de iniciar más conflictos. Evidentemente, estos no solo podrán ser más comunes y recurrentes, sino que además podrían llegar a ser más devastadores.

En estos casos y en la posibilidad de un uso malintencionado, estamos con la posibilidad de que estas tecnologías cometan actos éticamente incorrectos, estamos en un dilema moral en el que tecnologías pueden beneficiar a algunos, pero perjudicar a otros. O limitamos estos desarrollos a los que traigan un beneficio para la humanidad, no solo a una parte mientras les hace daño a otros; o si esto no es posible por el doble carácter, debemos pensar que tal vez esto no sea progreso del todo, sino un riesgo mayor como señaló la ONU hace unos años y que por no comprenderlo no lo vemos como una amenaza. Lo que sí es que ejemplos como estos nos hacen pensar en la agencialidad que puedan demostrar estas máquinas, sea cual sea la respuesta a este dilema. En efecto, al menos en el caso militar, hemos visto ya cómo las máquinas pueden deliberar y tomar la decisión de hacerle daño a un humano sin que un humano deba tener conexión al dron en tiempo real.

Esto nos permite concluir dos cosas que debemos cuestionar más a fondo en nuestro nuevo contexto y remodelando estas nociones al tener estos casos en cuenta. En primer lugar, estas máquinas son evidencia de que no es necesario tener aptitudes éticas o conciencia para que una entidad pueda tomar la decisión de asesinar a un ser humano. Esto nos hace cuestionar la posibilidad de que nuestra conciencia y aptitudes humanas no sean condición necesaria para la inteligencia, o que la deliberación que llevan a cabo las máquinas por más intrincada que sea sigue sin ser un acto inteligente. En segundo lugar, además, nuestro estatus moral puede estar viendo un deterioro en su valor e importancia al menos de forma indirecta. Así no nos damos cuenta, al permitir que máquinas maten por nosotros estamos permitiendo que entidades sin ninguna aptitud ética asesinen a seres humanos. Esto, en otras palabras, se traduce o en que estamos aceptando que entidades sin remordimiento ni ningún otro tipo de aptitudes éticas tomen decisiones que antes creíamos exclusivas para entidades con un estatus moral mayor o igual al de un ser humano. O bien aceptamos que tenemos el mismo estatus moral que una máquina como estas o hacemos algo al respecto porque de otra manera nuestro estatus moral está perdiendo legitimidad en un mundo de máquinas inteligentes.

La beneficencia es un valor ético presente desde que en Hannover se acuñó el término de inteligencia artificial, pues el objetivo de esta conferencia era crear máquinas inteligentes, sino que beneficiaran a la humanidad de forma equitativa y justa. Este valor que se traduce en actos voluntarios y desinteresados era el objetivo principal del desarrollo de la inteligencia artificial. Sin embargo, pudimos ver que esto no es lo que está sucediendo, sino que la tecnología está moldeando nuestro contexto de maneras beneficiosas y perjudiciales sin que nosotros hagamos mayor cosa o limitemos esto de alguna manera. Usar a máquinas de manera malintencionada al igual que permitir que estas lo hagan de forma autónoma es ética y moralmente reprochable, pues atenta contra el principio fundamental de una sociedad. A

través de un análisis filosófico, debemos ayudar a cuestionarnos nuestras intuiciones morales en este nuevo contexto de la IA y proporcionar bases racionales y morales para actuar de forma altruista en estas tecnologías y no exclusivamente bajo motivaciones financieras o de poder.

### **5.2.3 Privacidad**

Como ya se mostró, debemos considerar que el carácter público y universal del conocimiento puede ver un cambio ocasionado por los avances de inteligencias artificiales. Si pensamos esto junto con la capacidad de estas máquinas, que funcionan con conjuntos de datos, veremos la distinción entre datos privados y públicos, tanto en el funcionamiento de la máquina como en los datos que utiliza, como en la metateoría detrás de su algoritmo y su modelo de generalización. Por una parte, debemos pensar que los conjuntos grandes de datos utilizados para entrenar a una máquina sean seleccionados, implementados y modelados de manera que evitemos sesgos y robo de información privada. Debemos saber qué datos se le deben dotar a una máquina, asegurar que no tengan sesgos, que no sean erróneamente seleccionados, que no estén incompletos o que puedan incluso llegar a generar nuevos sesgos. Además, debemos pensarnos en el funcionamiento de estas máquinas inteligentes, en sí lo correcto es explicar de forma clara y pública el funcionamiento detrás, y cómo recopilan futuros datos mediante su funcionamiento.

Si analizamos estos dos estadios de tensión entre lo privado y lo público de los datos en el caso de Chat GPT, veremos rápidamente la necesidad de que estas sean cuestionadas. En efecto, como ya vimos, vivimos ya en una sociedad en la que se siente la carrera por la obtención de datos en tiempo real y de manera precisa. Al pensar en su funcionalidad, ChatGPT funciona con datos de entrenamiento que le permiten sacar conclusiones acerca de las indicaciones que se le den y responder algo al respecto. No solo es necesario que sepamos ese funcionamiento central para notar sus márgenes de error y sus posibles riesgos, a su vez debemos pensar con qué datos se entrenó esta tecnología, pues los datos de entrenamiento suelen ser la causa de sesgos. Puede que la IA analice datos de una manera sorprendentemente rápida, pero si no consideramos la calidad y el tipo de datos que utiliza y recoge, esta tecnología podría heredar nuestros errores al dotarlos de ellos.

Lo mismo sucede con el armamento militar, pues nos encontramos con un pésimo criterio a la hora de dotar de datos a las máquinas que posibilita sesgos y accidentes mortales, como también nos encontramos con máquinas que funcionan de una manera éticamente reprochable a tal punto que quizás sea mejor no dar a conocer su funcionamiento central o de núcleo. Hemos visto ya ejemplos como la identificación incorrecta de 28 miembros del congreso estadounidense que muestran la herencia de sesgos a través de conjuntos de datos de entrenamiento, por un lado. La selección errónea se debe a que los datos del conjunto de entrenamiento de la inteligencia artificial eran sesgados y por este sesgo interpretó mal los rostros de estos 28 congresistas. Los sesgos de datos pueden ser una de las causas principales por las cuales estas máquinas no están logrando distinguir correctamente entre civiles y combatientes. Los sesgos pueden permitir que más civiles se identifiquen como combatientes solo por algunas correlaciones que no causa que sean personas que debemos neutralizar. Es posible que todos los datos de entrenamiento contengan a terroristas con turbantes y aspecto árabe simplemente porque son implementados en el contexto afgano y esto puede ocasionar que mucho civiles sólo por tener turbantes sean objetivo militar.

Adicionalmente, es probable que, si desvelamos el funcionamiento detrás de estos algoritmos que han mostrado su valía en el campo militar, se intensifique su uso para cosas incorrectas y mortales. Nos encontramos con desarrollos que, al superar nuestras capacidades cognitivas,

quizá sea mejor no dar a conocer en el dominio público sino en manos de expertos con un conocimiento amplio sobre el tema que tenga en cuenta el aspecto ético. Esto no quiere decir que no sea repudiable el caso en que los estadounidenses oculten atentados fallidos con estas tecnologías. Lo importante es saber qué debemos mantener bajo conocimiento de pocos y qué no, pues solo así se tendrá una participación más activa que limite los riesgos y los daños a la humanidad.

La privacidad y el sesgo de datos han tomado mayor relevancia en el campo de la ética y la Filosofía debido a estos avances tecnológicos que estamos viendo cada día. La recopilación masiva de datos y el uso de estos en algoritmos de inteligencia artificial ha mostrado la posibilidad de que máquinas hereden nuestros sesgos y nuestra injusticia. Una selección incorrecta de datos permite esto y también que las máquinas no seleccionen, apliquen, o concluyan de estos datos sin mostrar los mismos sesgos o la misma incompletitud. La objetividad y precisión de las máquinas depende de que el entrenamiento no tenga sesgos o errores humanos. Por otra parte, la privacidad también refiere al derecho fundamental de que personas puedan controlar el acceso y el uso de su información personal. Hoy, más actividades las realizamos en línea con aplicaciones o algoritmos que recopilan nuestros datos personales y los almacenan, como vemos con Google, Amazon o el Chat GPT. Debemos pensar cómo equilibrar la necesidad de que las máquinas inteligentes recopilen información y usen esta información junto con el que esto se haga de una manera éticamente correcta que no atente contra nuestra privacidad y libertad como individuos dignos de estos valores.

Asimismo, debemos garantizar que las decisiones de algoritmos y máquinas inteligentes no sean injustas o discriminatorias debido al posible sesgo de datos. Siempre debemos tener garantía de que los datos elegidos no puedan ocasionar esto y de que tengamos la transparencia suficiente para poder analizar justamente esto. Por lo tanto, la privacidad de datos parece que ya no se trata solo de datos de individuos, sino que puede estar también afectando a los datos sobre el funcionamiento central de los modelos de software de inteligencia artificial. Puede que tengamos que repensarnos esta noción también desde la IA, y pensar que hay datos privados sobre el funcionamiento de esta que solo conciernen a pocos expertos en el área o que incluso puedan escapar al conocimiento humano y sólo la máquina entienda de ellos. Estos dos ámbitos, además, nos hacen pensar en la responsabilidad detrás de este balance apropiado entre lo privado y lo público de los datos, cuestión que analizaremos más adelante.

#### **5.2.4 Transparencia**

Por ahora centremos nuestra atención en el dilema que más vio una resignificación por los cambios en el contexto social actual: la transparencia. Esta noción se puede relacionar tanto con la tensión generada entre lo privado y lo público como con la tensión injusta y desigual de la inteligencia artificial y su desarrollo. Si pensamos la transparencia pues, a través de nuestras aptitudes éticas y leyes epistemológicas arcaicas concluiremos rápidamente que la solución es la de tener total transparencia. Puede que este sea, por ejemplo, el caso de GPT en el que no vemos mayor problema en difundir de forma pública y universal todo acerca de las tecnologías para llegar a un desarrollo más justo en el que participemos todos. En este caso podríamos decir que quizás lo más justo es esto. Pero en casos como el de MegaSyn o el del armamento militar no encontramos esta misma situación, sino que vemos cómo nuevas tensiones se están generando en cuanto a la transparencia y el libre acceso al conocimiento.

En el caso particular de MegaSyn nos encontramos con unos investigadores que optaron por no mostrar de manera completa y detallada el funcionamiento de su avance, con el fin de prevenir un funcionamiento injusto motivado por una intención mala. Pero esto va en contra

del principio de transparencia y conocimiento público, pues se espera que ocurra lo contrario. El problema es que estas tecnologías permiten malos usos al alcance de todos, y si no optamos por reconsiderar la transparencia pública y universal de la verdad, podríamos terminar causando más perjuicio que beneficio.

Esta misma tensión encontramos si seguimos considerando el armamento militar y el cambio que está generando en el campo de batalla desde el contexto arcaico. Puede que para el grupo militar que emplee estas tecnologías y sustituye a soldados lo conciba como algo justo incluso si no sabe nada detrás del funcionamiento de estas máquinas. Pero para un afgano que se halle en el otro lado del campo militar, le podrá parecer algo injusto y que además no se está dando a conocer en el mundo, sino que se lleva a cabo de forma clasificada. Como no todos están haciendo parte del beneficio de este tipo de tecnologías, la transparencia y el libre acceso vuelven a hallar tensiones en este nuevo contexto que debemos preguntarnos. Puede que aquel principio del libre acceso al conocimiento esté en juego al verse envuelto en casos en los que aparentemente es preferible hacer todo lo contrario.

La transparencia en la filosofía siempre se ha vinculado con la virtud intelectual que busca promover un diálogo abierto y honesto en la búsqueda del conocimiento y la verdad. En el contexto ético y responsable de la ética de máquinas inteligentes, los filósofos estamos valorando la integridad y honestidad que se necesita en el proceso de indagación y construcción de máquinas inteligentes. Esto contribuirá al enriquecimiento del pensamiento y del debate en esta disciplina o campo de investigación. Desde la filosofía antigua, Platón y Aristóteles, en sus diálogos, abogaban por la claridad y la transparencia en la comunicación y el razonamiento claro a través de una exposición filosófica que fuera accesible para todos. Luego, en la filosofía moderna, Kant enfatizaba esta transparencia en la ética y en la moral, pues para el filósofo alemán la moralidad verdadera implicaba actuar de acuerdo con el deber y la razón de manera transparente y universal para no esconder así motivos ocultos o actuar en función de intereses egoístas. En los últimos años, la filosofía contemporánea ha abogado por dotar esta transparencia y universalidad de las perspectivas individuales y los instintos personales, ya que hemos empezado a considerar la verdad como algo relativo a las perspectivas del contexto actual. Esto, sin embargo, no aboga por una relatividad absoluta en cuanto a la verdad y los principios morales, sino más bien al desarrollo de este tipo de nociones y principios dependiendo del contexto social situado en específico en el que se halla inmersa la sociedad en ese momento. Parece que esto se ha dejado de lado al implementar y desarrollar tecnologías inteligentes. Al final, hemos dejado de lado que la transparencia y los demás valores morales cambian por las máquinas inteligentes, lo que nos obliga a reconsiderar estos conceptos y dotarles de un enfoque más adecuado al contexto actual en el que globalmente impactan las IA.

### **5.2.5 Responsabilidad**

Responsabilizar a una máquina inteligente y dotarle estatus moral ha mostrado ser algo incipiente en la inteligencia artificial, así como con la conciencia humana y con crear máquinas conscientes de sí mismas mientras actúan de acuerdo a aptitudes éticas. Sin embargo, esto no significa que máquinas autónomas para matar, para producir propiedad intelectual y presentarla como de uno o incluso máquinas que puedan generar moléculas tóxicas para la futura implementación en armas biológicas no tengan importe ético. En efecto, las máquinas ya no están solo yendo de un punto A a un punto B, tampoco están solamente siguiendo unas reglas para actuar. Las máquinas hoy por hoy están deliberando a través de datos y cálculos probabilísticos para tomar decisiones que nos están afectando de maneras que antes no se nos hubieran imaginado. Debemos pensar en todos los autores detrás de estos

desarrollos para dotar de responsabilidad a las causas tras implementaciones perjudiciales de máquinas inteligentes que no beneficien al ser humano. Sin una correcta responsabilización por los errores que puedan estar detrás de la inteligencia artificial, nunca nadie hará nada al respecto porque se preferirá el beneficio inmediato al riesgo indirecto.

Hasta este momento, las máquinas ya han mostrado su capacidad de ser autónomas. Aunque a muchos les suene extraño porque no tienen conciencia o no tienen actitudes éticas, lo cierto es que estas tecnologías no están actuando bajo comandos específicos o directrices 100% exactas que limiten su comportamiento a lo que esperamos, sino que lo hacen de forma deliberativa e imprecisa, rasgo que ya vimos denota inteligencia. Básicamente, nosotros les dotamos instrucciones y posibilidades a la máquina y luego dejamos que haga lo que quiera con esas posibilidades sin que limitemos esto a decisiones beneficiosas para la humanidad. En el caso de GPT no encontramos tanto problema, pues el objetivo final de la máquina es el de parecer inteligente. Por más que haga plagio y nos ayude a hacerlo a nosotros, el objetivo de la máquina es hacer todo lo contrario, sugerir todo lo contrario y en el fondo aparentar que es inteligente. En este caso, requerimos que humanos usen malintencionadamente estas tecnologías para ver un serio perjuicio. O, claro está, que diseñadores no prevean ciertos riesgos y esto de lugar a que se ocasionen.

El caso de MegaSyn es un poco más complejo, pues, por más buena intención que tuvieron los investigadores que desarrollaron estas tecnologías, ellos mismos lo hicieron de una forma tan ingenua e inocente que no tuvieron en cuenta, al menos inicialmente, las consecuencias que este desarrollo podría tener para la sociedad. Chat GPT y OpenAI al menos nos garantizan el poco riesgo y la seguridad que hay detrás de su procesador de lenguaje. Pero no fue el caso de esta tecnología, ya que los investigadores tuvieron que considerar el uso dual de la inteligencia artificial incluso antes de implementarla, pero mucho más antes de publicar algo sobre ella. Creo yo que, si algo llegara a pasar con estas tecnologías de manera malintencionada, los investigadores tendrían al menos parte de responsabilidad, pues no previnieron estos riesgos y sí los posibilitaron al menos en un porcentaje. Claro, las personas que al final usaron esta tecnología de mala intención son los más responsables, pero esto no implica que no vayamos a responsabilizar a todo el que pudo haber tenido algo de causa en estas situaciones.

Pero el caso que más llama la atención es el del armamento militar, pues no sucede lo mismo que en los otros dos casos en los cuales requerimos de un ser humano para que los resultados de estas máquinas sean éticamente repudiables. La autonomía del armamento militar trasciende a la de los otros ejemplos y tecnologías, pues la decisión que está tomando por sí misma la máquina es reprochable y éticamente repudiable sin que un agente humano deba intervenir. A la hora de tomar decisiones con la munición implantada en ella, este tipo de tecnologías no requiere que un humano esté al tanto y apruebe o rechace estas decisiones. Esto no significa que la máquina tenga una agencialidad y autonomía como la que nosotros demostramos. Más bien significa que sin tener estas capacidades mentales y éticas, igual nosotros como diseñadores e implementadores estamos dándoles la posibilidad de matar y tomar vidas. Claro está que también debemos responsabilizar a aquellos que ocultan información vital para la restricción de tecnologías militares, a los países que siguen desarrollando y utilizando estas tecnologías aún sin el aval de la ONU, a las personas que como individuos o grupos terroristas se hagan con estas tecnologías y las usen de formas malintencionadas. Al final, puede que esta situación se traduzca en la prohibición de tecnologías como estas, pues no han excedido nuestras capacidades mentales ni nuestra agencialidad moral pero sí están asesinando y esto no tiene sentido si se piensa con el estatus moral de algo.

No podemos desarrollar tecnologías inteligentes asumiendo únicamente las consecuencias positivas o beneficiosas de la inteligencia artificial. En efecto, hemos visto cómo las inteligencias artificiales nos pueden traer consecuencias no beneficiosas sino perjudiciales y excluir esto de nuestro análisis de consecuencias sería permitir el perjuicio de la inteligencia artificial y evadir la necesaria regulación que debe haber en su desarrollo e implementación. Debemos imponer un margen político y legal que nos permita responsabilizar a toda entidad involucrada en consecuencias perjudiciales a través de inteligencias artificiales. Esta responsabilización además debe ser ética, en tanto que se cometió una acción moralmente incorrecta, pero también debe ser política, ya que los líderes y funcionarios gubernamentales deben responsabilizarse por las acciones y las decisiones en el ámbito público, así como el continuo desarrollo de armas inteligentes incluso después de su prohibición. También debemos generar un sentimiento de responsabilidad en todos los ciudadanos que fomente que los ciudadanos para que participen en el proceso político y contribuyan al bienestar de la sociedad y al correcto desarrollo de inteligencias artificiales.

### **5.2.6 Leyes**

Casos como los de Google y su continua evasión ante la ley y la responsabilización por los perjuicios que han ocasionado, ya sean accidentales o no, han demostrado que las empresas privadas optan por evadir regulaciones que puedan interferir con su beneficio monetario o de poder. Esta práctica ha mostrado ser nociva, ya que impide establecer una relación de confianza entre máquinas y humanidad. Sin regulaciones que limiten el diseño, desarrollo e implementación de máquinas inteligentes, no podemos garantizar la protección de valores éticos que están siendo afectados.

Casos como el de ChatGPT ofrecen esperanza, ya que, aunque no siempre sean perfectos y puedan tener perjuicios, estas tecnologías y las empresas privadas detrás han optado por desarrollar avances de forma más preventiva y regulada, previniendo riesgos y usos malintencionados a través de la transparencia en sus términos y condiciones y en el funcionamiento del procesador de texto. Sin embargo, esto no ha llevado a un desarrollo completamente regulado, y es probable que sin una mayor participación política y gubernamental no podamos regular casos como el plagio en estas tecnologías y prevenir usos malintencionados.

En el caso de MegaSyn, intervenciones externas de compañías privadas y gobiernos fueron necesarias para evitar la divulgación sin considerar las consecuencias futuras. Las leyes también deben amparar a los científicos, sus avances y la forma en que desarrollan tecnologías. Sin regulaciones y una mayor participación, no podemos confiar plenamente en estas tecnologías y tener certeza de que no habrá consecuencias éticamente repudiables.

La necesidad de regular y aplicar leyes paralelamente al desarrollo de inteligencias artificiales ha sido reconocida, pero en casos como el del armamento militar, nos encontramos con tecnologías que no solo no han sido reguladas, sino que se ha hecho lo contrario. Esto requiere no solo leyes y regulaciones gubernamentales u organizacionales, sino la participación activa de todos los usuarios para tener regulaciones adecuadas y evitar que el desarrollo sea impulsado solo por intereses particulares. Sin esto, países seguirán en una carrera armamentista inteligente sin restricciones y sin garantías de que las armas inteligentes no maten a seres humanos.

Las leyes han sido fundamentales para el funcionamiento de las sociedades, adaptándose al contexto actual y a los desafíos éticos. Las leyes para regular la inteligencia artificial deben implementarse en todos los países, incluso por organizaciones supra gubernamentales, para evitar la evasión de responsabilidades. Sin embargo, el rápido avance tecnológico ha

superado la velocidad con que las leyes y gobiernos se actualizan, por lo que se requiere un balance para garantizar una implementación éticamente aceptable. La participación activa de la sociedad en el desarrollo y diseño de las tecnologías es crucial para lograr este equilibrio y evitar un uso descuidado y desmedido.

### **5.2.7 Implicaciones filosóficas**

La inteligencia artificial es un concepto o incluso un campo de investigación muy amplio que día a día encontramos inmerso en nuestro contexto social actual y nos vemos rodeados de muchos beneficios a través de ella. Aún no hemos encontrado una definición globalmente aceptada ni definitiva sobre lo que quiere significar la inteligencia artificial y hasta dónde llega su alcance en la sociedad, empresa necesaria pues vimos que su alcance no es solo beneficioso sino perjudicial. Para una aprehensión mayor de esta noción y un desarrollo más significativo, los investigadores han estudiado la inteligencia humana e imitaron los resultados en máquinas con maleabilidad lógica y capacidad de transformar actos inteligentes en datos estadísticos modelados por algoritmos y formalizados a través de conectores lógicos.

Imitando algunos rasgos de la mente y la inteligencia humana, los investigadores lograron crear máquinas tan efectivas que aprenden con algoritmos y pueden mejorar en su objetivo específico. Lo que se vivió hasta entonces, cuando empezamos a ver máquinas que superan a los humanos, puede equipararse a esa primera fase de la inteligencia artificial que McCarthy mencionó hace años. Actualmente, en cambio, vivimos ya la etapa de permeación en la que nos relacionamos con máquinas inteligentes a diario y nuestro contexto social actual se ve moldeado por esta relación. Pero por la forma en que se crearon estas tecnologías, este cambio de paradigma trae consigo tensiones éticas y riesgos para el futuro que no podemos dejar de lado y que nos obligan a cuestionar la naturaleza de esta relación en sí misma, no solo mediante su beneficio.

En efecto, puede que los mecanismos detrás de nuestra inteligencia y la manera en la que nuestra mente funciona hayan sido imitados. Todo el aprendizaje profundo y los desarrollos en esta área de la inteligencia artificial han basado sus desarrollos en nuestros mecanismos biológicos para lograr actos inteligentes. Pero no podemos apartar que la única característica de la mente humana reflejada en estas máquinas es el acto deliberativo en el que las máquinas le dan respuesta a un objetivo mediante objetivos y programas implantados en ellas por los humanos. Esto no quiere decir pues, que no haya un acto deliberativo porque en el fondo las máquinas están calculando estadísticamente las probabilidades de ciertos eventos para tomar una decisión y dar la respuesta más adecuada. No se parece a cómo hacemos estos actos deliberativos inteligentes, pues nuestra deliberación consta de un raciocinio estadístico o probabilístico, pero también de principios morales para el cálculo de consecuencias a corto o largo plazo y muchas otras facultades mentales que hacen de nuestra mente e inteligencia algo complejo y misterioso.

Los mismos humanos hemos optado por llamar a estos avances máquinas inteligentes cuando se trata solo de máquinas que con una adecuada programación logran realizar los mismos actos inteligentes que nosotros de forma más efectiva, pero no por esto más inteligente. En efecto, la efectividad en el acto deliberativo de las máquinas inteligentes no es porque estas demuestren mayor inteligencia, ya que requiere sensibilidad ética y otros aspectos que se hayan en la mente humana y no en la IA. El acto deliberativo de una máquina puede ser más complejo que el de los humanos, como hemos visto en ejemplos en los que nuestra comprensión humana no alcanzaba para entender ciertas deliberaciones de las máquinas. Pero la complejidad de las creaciones a las que llegamos junto con la incertidumbre y poco conocimiento que poseemos no implica que estas máquinas sean inteligentes, solo que imitan

una acción inteligente específica y por sus capacidades de almacenamiento y procesamiento lo hacen de una manera que supera nuestras capacidades.

Así se nos abre el camino epistemológico en el desarrollo de las inteligencias artificiales, pues la cuestión de qué es la inteligencia, qué constituye un acto inteligente, si este es imitable, entre otras son cuestiones revitalizadas en el nuevo contexto actual. Las máquinas pueden estar pasando el Test de Turing sin ningún problema, incluso pueden calcular cosas estadísticamente con mayor efectividad que nosotros o brindar respuestas más precisas. Pero, esto no es muestra de inteligencia sino de una programación adecuada que permite que una máquina de respuestas efectivas y esperadas acerca de un tema en particular. Hasta que no analicemos estas cuestiones epistemológicas las máquinas podrán ser más efectivas que nosotros y más precisas, pero no serán por esto inteligentes sino meras máquinas fieles a la habitación China de Searle.

Pensemos esto desde los resultados a los que llegó Gödel el siglo pasado. Por más procesadores y capacidades de almacenar datos grandes que tenga una máquina, esta se va a cobijar únicamente a su modelo formal y a los límites que se hayan en este sistema cerrado y coherente pero incompleto. Al tratarse de una mera formalización de actos y objetos del mundo en datos que se ajusten a un modelo estadístico, nos hallaremos con verdades que escapan la aprehensión del modelo de la máquina por su característica incompletitud. Así, puede que nos brinden respuestas muy adecuadas y específicas en un campo determinado, puede incluso que estas respuestas sean tal que con nuestras capacidades no las hubiéramos logrado aprehender y no comprendemos ni si quiera el razonamiento detrás del resultado. Pero en cuanto a generar algo más allá de ese sistema formal completo, las máquinas han demostrado su esencia estrecha y su incapacidad para pensar más allá de un sistema formal, rasgo esencial de la inteligencia. Hasta que no haya máquinas que se mejoran a sí mismas no hablaremos de máquinas que pueden pensar su meta-funcionamiento o que deliberen más allá de lo que implantamos en ella.

Por otro lado, nuestra forma de entender la epistemología y el conocimiento se está moldeando por estas máquinas sean o no inteligentes, ya que han presentado tensiones éticas. Por su uso dual, las máquinas han demostrado efectividad en actos beneficiosos y buenos, y en actos malintencionados y perjudiciales en cuanto al conocimiento, lo ético, lo político y básicamente en todo campo en el que hayamos presenciado su producción y divulgación. Esto lo hemos visto ya en muchos casos anteriormente y, como vimos, en algunos se ha optado por no divulgar este conocimiento de manera pública y transparente. Esto tiene sentido si consideramos las implicaciones riesgosas que acabamos de mencionar. Pero a su vez hace que nos cuestionemos ese carácter público y transparente que otorgábamos a nuestra epistemología y nuestro conocimiento. El conocimiento al que puedan llegar máquinas que ni siquiera nosotros podemos comprender una vez las diseñamos obviamente escapará nuestra comprensión o incluirá riesgos que nos superen y puedan llegar a nuestra destrucción masiva. La cuestión no solo nos obliga a pensar si acaso el conocimiento debe ser público o no, sino además si debemos seguir persiguiendo desarrollos que muy probablemente escapan nuestras capacidades cognitivas no podamos comprender siquiera.

Por este tipo de riesgos inherentes a la IA, la ética mostró también su importe en la inteligencia artificial. El análisis detallado de los casos nos mostró, en primer lugar, que nuevas tensiones están desarrollándose en el contexto actual una vez que humanos y máquinas inteligentes se relacionan. Algunas de estas eran prevenibles, pues un diseño más consciente del diseño, desarrollo e implementación de estas tecnologías hubiera logrado que no se dieran estas tensiones. Pero hay tensiones que no eran prevenibles, sino que son impulsadas y generadas solo por este nuevo contexto y que debemos pensarnos para

solucionar. Debido a esto, debemos reconsiderar nuestros principios éticos y dejar de asumir que son algo estático e invariable. En el ámbito epistemológico ya hemos visto de qué manera la transparencia y el conocimiento público se están enfrentando a un cambio de paradigma debido a las máquinas inteligentes. Esto mismo sucede con la responsabilidad, la privacidad, la justicia y la beneficencia, como vimos de forma detallada anteriormente. Puede que no sean solo principios cómo estos los que sufran cambios, toda la forma en la que concebimos la ética y la naturaleza de nuestra relación con las herramientas se está viendo modificada por la alta complejidad y poca comprensión en sus desarrollos. Es necesario considerar los límites de estas máquinas, pues, aunque posean una alta efectividad no cuentan con ninguna aptitud ética o con conciencia que les permita analizar un acto deliberativo no solo desde la efectividad.

Esta ha sido la motivación ulterior por la que investigadores han intentado imitar nuestra sensibilidad ética en máquinas inteligentes. Esperan que al preguntarnos por lo que nosotros los humanos consideramos bueno, las máquinas no cometan actos que afectan nuestra vida de una forma perjudicial, sino que se garantice su actuar éticamente aceptable. Pero de igual manera que pasó con la inteligencia, no podemos pretender que las máquinas tengan sensibilidad ética sin antes saber qué significa esta sensibilidad ética para los humanos en relación con la mente humana. Solo así podremos saber que caracteriza tener facultades éticas y podremos desarrollar algo similar en las máquinas inteligentes. Así como debemos llegar a un análisis más profundo y detallado acerca de la inteligencia humana y lo que la caracteriza, también debemos pensar la sensibilidad ética y la posibilidad de que máquinas, considerando su uso dual, actúen bajo una sensibilidad ética que las haga optar solo por los actos beneficiosos. Al ser algo que depende del contexto actual en el que se gesta, no podemos dejar de lado las tensiones en ciertos casos particulares y desarrollar esta investigación en la actualidad con base a este contexto.

Así mismo, hemos visto la necesidad de cuestionar lo que consideramos bueno para una máquina. Que una máquina como Deep Blue le gane al campeón mundial de Ajedrez hace que esta sea buena. Que una máquina como Watson logre detectar enfermedades de forma correcta y prevenga muertes es que esta sea buena. Pero ¿qué significa que una máquina como los drones usados de manera autónoma para matar sea buena? A mi parecer, el que una máquina actúe bien ya debería limitar el uso de estas máquinas para asesinar a otros seres humanos, sean quienes sean. No es moralmente correcto que una máquina sin conciencia, estatus moral o ninguna facultad mental compleja esté capacitada por medio de un objetivo que le permita tomar la decisión de quitar una vida. Debemos prestar especial atención en los objetivos que les dotamos a estas máquinas y en que estos no conlleven a casos cómo estos en los que la máquina no es buena así mate de forma más efectiva que un soldado entrenado por más de diez años. La agencialidad de estas máquinas se basa exclusivamente en un intrincado nivel de abstracción que muchas veces no llegamos a conocer de forma precisa y que solo calcula a través de datos numéricos y valores probabilísticos.

Nos hallamos inmersos en un contexto totalmente diferente al de hace unos 50 años. Hoy por hoy, no solo nos relacionamos con máquinas inteligentes, sino que lo hacemos de igual manera que entre nosotros mismos. Por esto, también se nos abre la cuestión por el tipo de entidad que puedan llegar a ser estas máquinas. Más allá de su programación inicial y de dotarle un objetivo inicial, existen ya máquinas que no necesitan de la intervención humana tras este primer paso, y toda decisión posterior depende exclusivamente de su acto deliberativo. Este acto deliberativo, como ya vimos, no es más que un análisis estadístico y probabilístico de las cosas que deja de lado la sensibilidad ética y las posibles consecuencias riesgosas detrás de una decisión. Así, puede que una máquina no requiera intervenciones

humanas al tomar decisiones que impactan nuestras vidas, pero la autonomía de estas depende de ese primer paso en el que los humanos tenemos la capacidad y la necesidad de limitar el rango de esta aparente autonomía en las máquinas. En el fondo, una máquina no mataría si nosotros no se lo permitiéramos. Pero puede que llegue el momento en el que una máquina pueda mejorar por sí misma y llegue a modificar su objetivo y su modelo inicial. Es en ese momento, lo queramos o no, en el que la ontología se vuelve vital para un análisis completo de la inteligencia artificial, pues lo que constituye una máquina inteligente también estaría autónomamente dependiendo de las deliberaciones que tome la máquina para modificarse a sí mismo. Que un carro se conduzca solo, por ejemplo, no significa que el carro tenga agencialidad o estatus moral. Pero si este mismo auto considera que su programación inicial no es la más efectiva y por decisión propia lleva cambios en esta, puede que debamos considerar que estas máquinas tienen la agencialidad suficiente para concebirse a sí mismas y modificarse.

Hasta este momento, nos encontramos con máquinas que no son inteligentes ni son autónomas, pero que sí parecen serlo. Ellas han manejado por sí mismas, han diagnosticado enfermedades y han decidido quitar una vida. Pero en el fondo hasta este momento las implicaciones éticas, epistemológicas y ontológicas que estas tecnologías puedan traer consigo y la responsabilidad última por sus implementaciones y su uso correcto recae activamente en los humanos que las diseñan, las implementan y las usan. Al final, el alcance de una máquina sea bueno o malo, depende de lo que le dotemos a su objetivo particular mediante el cual delibera, al menos hasta este momento. Así, aunque lo haga de forma autónoma su deliberación se haya limitada al alcance que nosotros como seres humanos y diseñadores les permitimos. La responsabilidad todavía recae enteramente en nosotros y en las decisiones que tomemos para continuar este desarrollo de una manera menos perjudicial.

Precisamente por esto es qué debemos considerar si limitar el alcance que estamos dándole a las máquinas es lo correcto ya que, aunque lo hagan de forma autónoma, si nuestro uso designado para una máquina no fuera el de asesinar o destruir armamento militar esta no lo hiciera. Lo mismo con el descubrimiento de toxinas. Si nosotros no le dotamos gratificación en hallar estas, no lo haría. Hasta que no alcancemos máquinas que se mejoren y cambien el objetivo central del programa de una máquina inteligente, la necesidad de responsabilizar y cuestionar este diseño y su implementación para evitar riesgos recae solo en los humanos y en la forma en que limitan estas máquinas para que no lleven a los riesgos que vimos en el contexto actual. Esto es misión de todos, pues como ya hemos visto, hasta que no nos responsabilicemos todos por cuestionarnos la naturaleza de estas máquinas y su relación con la humanidad no podremos prevenir los riesgos. Debemos generar conciencia para que las leyes no solo se implementen, sino que no se sigan evadiendo.

Desde que la inteligencia artificial ha demostrado su capacidad de impactar beneficiosamente de forma inmediata y más efectiva que los mismos humanos, hemos optado por dejar de lado las posibles consecuencias riesgosas y la cuestión por cuestione importantes en cuanto a cómo estas tecnologías están cambiando la forma en la que vemos las cosas y nos pensamos ciertas nociones. Solo nos hemos enfocado en la implementación apresurada que logre ganar la carrera del progreso y las finanzas. El uso y desarrollo de estas tecnologías ha sido tan desmedido y despreocupado que nos olvidamos cuestionar la naturaleza de la inteligencia artificial en sí misma y la forma correcta en la que debemos relacionarnos con ella. Dejamos a un lado la metateoría detrás de este campo investigativo y olvidamos de esta manera que si no sabemos qué es algo no lo podemos entender y mucho menos medir su alcance, sea positivo o negativo. En el contexto actual en el que la inteligencia artificial moldea todo lo que hacemos y cómo vemos las cosas, debemos cuestionar e investigar la naturaleza de estas

tecnologías y de nuestra relación con ellas. Pareciera que en los últimos años hemos querido que estas máquinas se nos parezcan inteligentes, así pasando el Test de Turing, creemos que esto es razón suficiente para implementar estos avances en todo posible campo de acción. Sin embargo, hemos visto a través de este trabajo que muchas veces es mejor limitar o frenar este desarrollo al menos hasta que conozcamos mejor lo que hay detrás de nuestra mente, lo que hay detrás del funcionamiento de estas máquinas y lo que nos diferencia a ambas.

## Conclusiones

Han pasado 67 años desde la primera vez que nos referimos a la posibilidad de máquinas inteligentes y a la inteligencia artificial en Dartmouth. Han transcurrido 38 años desde que Moore distinguió dos etapas de creación y permeación en el desarrollo de la inteligencia artificial. Aunque el camino parece mucho trecho, el trabajo de investigación realizado hasta ahora pretende demostrar lo siguiente: i) Incluso actualmente no sabemos con certeza qué pertenece al ámbito de la inteligencia artificial y qué no. Es crucial mejorar nuestra investigación en cuanto a la metateoría de este campo para obtener una mejor comprensión y clasificar eficazmente lo relevante y lo que no lo es en estos casos. Todavía no hemos llegado a una conclusión definitiva acerca de qué es exactamente la inteligencia artificial y cuáles son sus requerimientos esenciales. Tampoco conocemos con claridad qué diferencia existe entre el diseño de la IA estrecha y la IA general. Si deseamos que nuestra relación con las máquinas inteligentes sea no solo simbiótica, sino también mutualista, en la que todos salgamos beneficiados, debemos promover una delimitación más clara que incluya a toda la humanidad y no solo a los científicos en sus laboratorios. ii) Aunque las máquinas están permeando cada vez más nuestra sociedad, no debemos descuidar las tensiones éticas que podrían surgir y que requieren regulación en su implementación, así como una consideración más holística. Las máquinas moldean nuestros valores éticos, pero lo hacen sin tener un desarrollo ético propio lo que puede llevar a que nuestro sentido moral se reduzca a un simple cálculo. Si no consideramos nuestros valores éticos y no nos aseguramos de que las máquinas los promuevan en lugar de restarles valor, podría llegar un momento en el que un algoritmo sea más relevante para nosotros que un principio moral. Es fundamental abordar estos desafíos de manera responsable y ética para garantizar un desarrollo armonioso de la inteligencia artificial y su interacción con la sociedad en general.

Cada día se siente como una aseveración más factible el hecho de que las máquinas tengan una aplicación limitada. Pero todavía no estamos ni cerca de delimitar el campo de investigación de la inteligencia artificial, ni siquiera tenemos una definición ampliamente aceptada. Lo que motivó a la inteligencia artificial y su desarrollo fue este aparente carácter ilimitado en el funcionamiento, pero lo cierto es que hasta que no limitemos sus funciones y utilidad, no lograremos unificar el campo y conocer su verdadero alcance. Se requiere una unificación de las diferentes áreas de investigación que abarquen estas disciplinas y logren un mejor entendimiento, diseño y desarrollo. Lo que es peor, hasta que no entendamos de manera más clara cómo funciona esta tecnología no podremos evitar riesgos de manera realmente efectiva. Así pues, la ética también ha demostrado la importancia de delimitar los objetivos, las funciones y los datos de una máquina inteligente para evitar perjuicios accidentales o malintencionados, destacando así la importancia de esta revisión.

Recientemente, los descubrimientos en la cueva de los huesos en África dan sustento a la tesis de que hace unos 100,000 años ya existían primates con manifestaciones de espiritualidad y rituales relacionados con ella, sensibilidad y aptitudes éticas. Cada vez hay más pruebas que sugieren que el cerebro humano no es tan especial y que la inteligencia

humana no es la única forma de inteligencia en el mundo. Actualmente son las máquinas las que, sin conciencia humana logran realizar tareas que denotamos como inteligentes incluso de manera más eficiente. Puede que en el fondo esto no sea inteligencia al menos como nosotros los seres humanos la entendemos y la experimentamos. Pero no tenemos razón para afirmar que el objetivo programado en una máquina inteligente autónoma no sea suficiente para otorgarle inteligencia, ya que cumple con estos rasgos deliberativos e intencionales, aunque lo haga en un campo controlado y delimitado de acción. Sin embargo, el problema radica en que esto convierte decisiones críticas en simples cálculos aritméticos dejando de lado la sensibilidad y la aptitud ética, aunque no sea de naturaleza humana. Dejemos de pensar que para lograr avances en la IA debemos imitar por completo la mente y la inteligencia humana. A su vez, deberíamos cuestionar si las máquinas requieren de las aptitudes éticas y mentales que aparentemente requiere el intelecto humano.

Siendo nosotros los únicos agentes que poseemos estas aptitudes éticas, al menos hasta este momento, debemos asumir la responsabilidad de establecer límites a los objetivos de las máquinas y considerar cuidadosamente cómo diseñarlos para que representen menos riesgos. Por lo tanto, la transparencia en cuanto a los objetivos que asignamos a una máquina se vuelve esencial al diseñar y construir este tipo de tecnologías. Si los objetivos no están claros, o si no los limitamos para que generen resultados éticamente correctos, podríamos tener máquinas que sean aplicables en numerosas tareas, pero no todas serían beneficiosas. Es fundamental evitar que se definan objetivos maliciosos en las nuevas tecnologías para facilitar crímenes conocidos y posiblemente dar lugar a nuevos. También debemos reflexionar sobre cómo los objetivos mal especificados podrían perjudicarnos. Los algoritmos son muy necesarios, debemos reconsiderarlos y pensarlos junto con las máquinas que creamos y que los realizan de manera autónoma. Este es nuestro único rasgo de autonomía, agencialidad y estatus moral para las máquinas inteligentes al menos hasta ahora. A pesar de las innumerables aplicaciones que pueden tener las máquinas, no podemos olvidar que uno de los objetivos primordiales es el bienestar humano. Debemos considerar cuidadosamente los efectos y consecuencias de nuestras acciones para garantizar un uso responsable y ético de la inteligencia artificial y la tecnología en general.

Esto ha traído consigo problemas, ya que la felicidad en su importancia biológica y psicológica ha demostrado ser otra fuente de datos que se puede analizar y programar en una computadora. Sin embargo, dejando de lado la cuestión específica sobre si la felicidad es un valor real o aparente al que podemos llegar con inteligencia artificial, las tecnologías modernas están modificando su entorno y cómo nos relacionamos en él solo a través de cálculos. Al centrarse únicamente en la forma de ubicarnos y llevarnos al lugar deseado, los algoritmos de GPS, al excluir el desarrollo de las capacidades cognitivas humanas, pueden estar dejando de lado los objetivos iniciales de la inteligencia artificial y perjudicándonos a largo plazo. El ejemplo del GPS demuestra que puede suceder al menos con el sentido de nuestra ubicación, pero puede estar ocurriendo incluso con nuestra forma de caminar a largo plazo. Puede que la realidad que vimos en Wall-E no está tan lejana. Como señala Bostrom (2004), la humanidad se encuentra en un desequilibrio evolutivo y la implementación de nuevas tecnologías podría aumentar este desequilibrio exponencialmente. Nuestras disposiciones evolutivas están quedando rezagadas y no están optimizadas para el entorno que nos rodea. La inclusión de la tecnología ha llevado incluso a cuestionar si los humanos deben ser reemplazados en muchas áreas de trabajo, incluso en medicina. Parece que las máquinas inteligentes y su desarrollo están superando a los seres humanos, pero mientras abogamos por su desarrollo, también debemos considerar si esto está perjudicando las capacidades humanas y si estamos dispuestos a permitirlo.

El problema puede llegar a ser tan grave que podríamos llegar a una época en la que solo a través de la creación de superhumanos mediante la biotecnología podamos competir con las supermáquinas que estamos desarrollando y buscando construir. Sin embargo, si el desarrollo, la aplicación, los conjuntos de datos y muchos otros aspectos de la inteligencia artificial siguen siendo principalmente controlados por el 10% de la población que posee más de la mitad de las riquezas del mundo, esto podría llevar a una especie de dominación de los seres humanos, en la que el 90% de la población se vea incluso extinto y olvidado, en aras de un futuro en el que las inteligencias artificiales y los superhumanos sean la nueva especie dominante.

Tenemos que enfocarnos en el funcionamiento de las inteligencias artificiales y asegurarnos de que sea justo y equitativo, incluyendo a toda la humanidad en la medida de lo posible. También debe ser un funcionamiento seguro que no solo garantice los objetivos, sino que también evite hacernos daño en la medida de lo posible. Aunque todavía no tenemos claros los límites de la inteligencia artificial, algo de lo que sí estamos seguros es que estos límites no pueden comprometer la mente y la ética humanas. No podemos garantizar que esta sea la única forma de poseer inteligencia y actitudes éticas, e incluso hay desarrollos que excluyen estas características y resultan más efectivos en el desarrollo de máquinas inteligentes generales. Pero, aunque el funcionamiento y desarrollo de la inteligencia artificial debe estar separado de las limitaciones humanas (Bostrom & Yudkowsky), también debemos tener garantía de que esta diferencia no le restará valor a nuestra propia conciencia y actitudes éticas.

El problema se agrava cuando consideramos los conjuntos de datos, ya que, si estos están sesgados, mal seleccionados o contienen datos incorrectos, entre otros sesgos, aunque los objetivos y las funciones estén correctamente desarrollados, las máquinas heredarán estos errores y sesgos y los replicarán al tomar decisiones o interactuar en el mundo. Debemos prestar atención a los conjuntos de datos. La carrera por los datos ya está presente en nuestro día a día, donde cada vez adquieren más importancia y valor. Debemos garantizar un tratamiento adecuado de estos conjuntos de datos, establecer límites claros entre los datos privados y públicos, y asegurar la vigencia y actualización de estos datos. Nuestros avances en la ciencia han permitido que cada vez se pueda reducir a los humanos a meras tiras de información, representadas por datos y valores numéricos. Por ello, es necesario proteger algunas de estas métricas como información privada, ya que, de lo contrario, nuestros miedos más grandes, nuestros secretos más íntimos o nuestras creencias más espirituales podrían terminar convertidos en simples datos numéricos al servicio de máquinas inteligentes.

Se están generando tensiones entre los beneficios y perjuicios que esta relación trae consigo, por lo que debemos prestar especial atención a estos aspectos y a cómo pueden cambiar nuestra percepción de la ética y la moral. Para ello, recurrimos a algunos casos particulares que han ocurrido en los últimos 15 años. El análisis de estos datos ha permitido identificar nuevos posibles riesgos que ya nos afectan y otros que pueden surgir si el desarrollo e implementación de la inteligencia artificial continúa sin cambios. Algunos de estos interrogantes son:

- ¿Salvamos a millones perjudicando a miles con avances médicos que tienen márgenes de error?
- ¿Mejoramos la vida de miles perjudicando a millones al mejorar nuestro ADN y limitar su implementación a la población más rica?
- ¿Qué margen de error se considera apropiado e inapropiado para el desarrollo e implementación de la inteligencia artificial?

- ¿Implementamos máquinas que, cuando enfrentan dilemas morales, perjudican a su usuario sin dudarlo?
- ¿Permitimos el reemplazo de miles de humanos en áreas de trabajo y otros ámbitos que contribuyen al desarrollo humano?
- ¿Priorizamos el bienestar humano o el desarrollo tecnológico?

Estos y otros dilemas morales que han surgido o se han revitalizado debido al desarrollo de la inteligencia artificial reflejan cómo nuestros valores éticos están cambiando debido a estas tecnologías. Sin embargo, el desarrollo de estas tecnologías ha sido desigual e involucra a muy pocos que evaden las leyes y los derechos humanos para no poner en riesgo los avances tecnológicos. Es urgente que la participación de políticas, usuarios, interesados y corporaciones que desarrollan estas tecnologías sea motivada y garantizada. Esto nos permitirá construir consensos y manejar desacuerdos de una manera que beneficie a todos (Nemitz, 2018). No solo la tecnología está moldeando nuestra realidad, sino que también la política, la ley y la democracia deben moldearse en una relación mutualista que beneficie tanto a la tecnología como a los seres humanos y garantice su bienestar y seguridad.

Previamente hemos tomado conciencia de cómo estamos repitiendo lo que sucedió con Internet al desarrollar inteligencias artificiales y los posibles riesgos asociados. Los legisladores y las corporaciones tecnológicas deben reconocer la importancia ética de las inteligencias artificiales, otorgarles relevancia y tomar medidas al respecto. Google, por ejemplo, ha dado los primeros pasos al revelar públicamente sus principios de responsabilidad en IA, pero esto es solo el inicio. Siguiendo a Nemitz (2018) y Whittlestone et al. (2019), debemos alcanzar el punto en el que la política y la ética se involucren en el desarrollo de tecnologías inteligentes y garantizar lo siguiente:

- Comprender las diferencias y ambigüedades en el uso de conceptos y en el desarrollo de la inteligencia artificial.
- Unificar disciplinas, usuarios, desarrollos, políticas, corporaciones y todos los actores directos e indirectos en el campo de la inteligencia artificial. Es necesario fomentar el debate para llegar a soluciones y promover el desarrollo de esta tecnología.
- Consensuar sobre los mayores riesgos de la inteligencia artificial, en la actualidad y en el futuro potencial.
- La ley debe garantizar una evaluación previa de riesgos antes de la implementación de tecnologías, para comprender los mayores riesgos que pueden surgir al implementar la inteligencia artificial.
- Desarrolladores, usuarios y corporaciones deben basar su relación con la inteligencia artificial en principios democráticos y políticos, para salvaguardar los derechos humanos, los intereses de la sociedad y el desarrollo individual adecuado.
- Todos los interesados en el desarrollo y conocimiento de la inteligencia artificial deben tener la participación garantizada, incluso a través de procesos democráticos.

Estos puntos no son exclusivos ni abarcan los aspectos a considerar. Al final, lo más relevante es comprender que estos son solo pasos iniciales hacia la dirección en la que debemos orientar nuestra investigación, y el paso más importante es explorar la relación que debemos establecer entre la inteligencia artificial, la ética y la política, y cuáles son las mejores formas de lograrlo. Las máquinas que protagonizan nuestros ejemplos son máquinas que todavía se rigen exclusivamente por la inteligencia estrecha. Si este es el potencial de riesgo y perjuicio de máquinas que solo se rigen en un campo de acción específico, se vuelve también cuestión fundamental pensarnos si la posterior creación de máquinas generales pueda

incluso ser más peligrosa. Las máquinas si hacen las cosas de manera más efectiva que nosotros, pero no de manera más ética o menos reprochable. En muchos casos ha demostrado ser más repudiable que la decisión humana. No tenemos ninguna certeza sobre cuál teoría ética implementar en la IA, ni siquiera sabemos si debemos enfocarnos en principios o en limitar los funcionamientos de la máquina; pero nuestros esfuerzos pueden centrarse en encontrar los puntos en común y tenerlos en cuenta a la hora de tomar decisiones en el diseño y la implementación de máquinas inteligentes para llegar al bienestar común.

¿Qué depara entonces a la filosofía y a los filósofos en esta nueva era que se abre con el desarrollo e implementación de inteligencias artificiales? Según Dennett (1994), se nos presenta una bifurcación en la que podemos elegir entre dos opciones. Podemos considerar esta nueva situación como una oportunidad para hacer filosofía de una manera distinta, o simplemente verla como un nuevo objeto de atención utilizando los métodos filosóficos tradicionales. La filosofía también debe enfrentar los nuevos desafíos que este contexto plantea. Aunque aún no hayamos logrado desarrollar máquinas inteligentes generales y éticas, las máquinas que ya interactúan con nosotros están modificando acciones humanas que son relevantes para la filosofía en cuanto a lo epistemológico, lo ontológico, lo ético y lo político. No podemos quedarnos estancados pensando que la filosofía no puede cambiar de paradigma, sino que debemos adaptarnos a los cambios que estamos experimentando. Los tres dilemas morales que analizamos junto con sus implicaciones filosóficas, aunque conciernen al área de inteligencia estrecha que es programada únicamente con un objetivo en vista, son suficientes para demostrar de qué manera nuestros valores éticos están viendo modificaciones en nuestro contexto social situado y cómo debemos pensar nuevamente ciertos campos de la filosofía.

Cuando le pregunté a ChatGpt si estaba de acuerdo en que se puede usar de formas maliciosas y si es importante su regulación, respondió esto: “Como modelo de lenguaje de inteligencia artificial, no tengo opiniones propias. Sin embargo, puedo decirte que es importante abordar las preocupaciones legítimas sobre el mal uso de las tecnologías de inteligencia artificial, como ChatGPT. La falta de regulaciones claras y actualizadas puede plantear riesgos para la seguridad y el bienestar de las personas. Es fundamental que las leyes y regulaciones se adapten continuamente a medida que estas tecnologías evolucionan para garantizar un uso responsable y ético” (Open AI, 2021). Así pues, las tecnologías inteligentes no son máquinas malintencionadas que intentan matar a alguien o controlar a la raza humana como vimos en Matrix o en Terminator. Sin embargo, si vimos su potencial al servicio de humanos malintencionados que si las usan como medios exclusivamente podrían causar más perjuicio que beneficio. Las tecnologías están pues, moldeando nuestro contexto y nuestros valores éticos consigo haciendo que sea necesario pensarnos estas máquinas desde un enfoque ético, ontológico y epistemológico. El objetivo de este trabajo de investigación no es abogar por una teoría filosófica que resuelva el paradigma actual y las tensiones que genera, sino por concienciar a los lectores sobre la magnitud de los desarrollos que estamos viviendo en nuestro contexto y la necesidad de pensarnos esto desde un punto de vista más amplio.

Lo mismo vimos en el caso de MegaSyn, en el cual concluimos rápidamente una necesidad de desarrollar estas tecnologías de una forma menos ingenua que prevenga el uso malintencionado y que caiga en manos malintencionadas. Debemos hacer que la participación sea más diversa y amplia a la hora de desarrollar estas tecnologías. En caso contrario, estas pueden causar incluso más daño que el bienestar que se le pensó en primera instancia y que se previno. Además, esta participación debe ser de todos, no solo de gobiernos y organizaciones. Hasta que la mayoría de la humanidad no tome conciencia en cuanto a estos desarrollos nada ganamos con que la ONU prohíba los drones autónomos o con que los

Estados Unidos promuevan unos principios muy similares a los que traté de desarrollar en este trabajo.

El caso militar es una prueba de esto, ya que en este vimos que sucede lo contrario a lo que se esperaría en cuanto al desarrollo de inteligencias artificiales. Los militares están empeñados en desarrollar estas tecnologías, y han optado por ocultar posibles errores y fallos en su funcionamiento para que no haya un freno en el desarrollo. Adicionalmente, los drones que se han logrado a generar en este ámbito nos hacen pensar en lo que hasta ahora creíamos como algo éticamente y universalmente aceptable. En efecto, vemos que la transparencia de la verdad y el dominio público del conocimiento y los avances científicos se están alterando por este nuevo contexto social en el que parece mejor guardarse ciertos conocimientos y ciertas verdades por la posibilidad de que se usen éticamente repudiablemente. Además, nos muestran que las complejidades de estas máquinas pueden llegar a ser tanto que, aunque lo queramos, no podamos entender algunos funcionamientos detrás de ciertas tecnologías. La autonomía y el estatus moral, por otro lado, también están viendo un cambio en cuanto a la inteligencia artificial se refiere. Las máquinas autónomas han demostrado ser tan independientes como para tomar decisiones que afecten e impacten la vida de los humanos de forma incluso éticamente reprochable. La falta de aptitudes éticas junto con una creciente autonomía ha logrado que se vean casos como estos en los últimos cinco años. No es tan valioso nuestro estatus moral cómo creíamos al tomar este tipo de decisiones, pues se las cedimos a máquinas que lo calculan solo por cálculos y cuya responsabilización no consistirá en algún castigo al carecer de aptitudes éticas. La autonomía que muestran las máquinas no requiere un alto grado de estatus moral como el humano, solo un alto nivel de abstracción en el que máquinas hace mucho nos supera. Así, bajo este nuevo contexto, debemos cuestionarnos estas nociones y analizar cómo cambian.

Incluso podemos imaginar un futuro en el que logremos crear una máquina inteligente general y ética, llegando posiblemente a la etapa de la singularidad. No mencionó esta posibilidad en las implicaciones filosóficas porque muchos investigadores consideran que este evento es lejano en el tiempo y no hay ejemplos concretos de máquinas generales que puedan afectarnos actualmente. La razón por la que se cree que estamos lejos de este momento es que las máquinas siempre estarán bajo el control humano y las limitaciones de sus programadores. Si construimos máquinas que puedan mejorar por sí mismas, como se dijo en la primera conferencia citada en la introducción, esa restricción desaparecería y tendríamos que considerar que las máquinas inteligentes generales actúen sin saber qué decisiones tomarán. Recientemente, en junio de 2023, Google y DeepMind presentaron a RoboCat<sup>49</sup>, un precursor de Gato y la primera máquina capaz de expandir su base de conocimiento al mejorar por sí misma y adaptarse a diversas tareas en el mundo real, incluso cambiando de enfoque.

Con la aparición de estos, no podemos negar que debemos abordar estos desarrollos y repensar muchas nociones fundamentales para la ética, la moral, la política, la educación y todas las áreas sociales afectadas por la inteligencia artificial. Por ejemplo, RoboCat toma decisiones sobre su propio código, su comportamiento y las consecuencias de sus acciones. De alguna manera se piensa su meta-objetivo y lo modifica. Modifica lo que en su núcleo significa ser RoboCat. La intención, la autonomía e incluso la agencialidad moral deben ser consideradas nuevamente, no desde un enfoque humanista en el que los humanos son la cúspide del desarrollo ético e intelectual, sino desde una perspectiva en la que estamos desarrollando máquinas que podrían superarnos en estas capacidades e incluso llegar a ser

---

<sup>49</sup> <https://www.deepmind.com/blog/robocat-a-self-improving-robotic-agent>

entidades. Algunos argumentan que estas máquinas aún tienen un código central y un código de aprendizaje que no pueden modificar por sí mismas. Sin embargo, RoboCat mejora su capacidad de automejora con el tiempo y busca referencias de humanos y otras formas de programación para mejorar en estas actividades. Además, como señala Omohundro (2007), estos sistemas operarían dentro de las leyes físicas, como el espacio, el tiempo, la materia y la energía libre. Es posible que ejerzan fuerza no solo para mejorar sus funciones iniciales, sino también para automejorarse, lo que podría implicar la manipulación de materia y energía. En otras palabras, RoboCat podría dedicar esfuerzos significativos no solo para mejorar las tareas que realiza, sino también para mejorar su propio código y funcionalidad, con el fin de lograr una mayor eficiencia y cumplir mejor con sus propósitos y motivaciones.

En el fondo, una máquina que se mejore a sí misma puede ser aquella que nos lleve hacia la singularidad. Cualquier obstáculo sería solo uno más de los problemas que estas máquinas podrían calcular y resolver mediante mejoras, superando incluso el desempeño humano (Omohundro, 2008). Estas máquinas podrían superarnos en actitudes éticas y tener un estatus moral más elevado, comportándose de manera más adecuada que nosotros. Como señalan Bostrom y Yudkowsky (2014), estas máquinas podrían incluso reproducirse al copiar sus códigos en otras inteligencias artificiales y modificarlas a sí mismas. En un solo día, podrían clonarse miles de veces, alcanzando una superioridad no solo en su estatus moral, sino también en la dominación de la realidad social y el contexto. Sin embargo, ya mencioné que este no es un trabajo con un enfoque distópico y pesimista al estilo de "The Matrix" o "Terminator". Simplemente quiero resaltar la necesidad y la importancia de reconsiderar los principios éticos y filosóficos en el desarrollo de inteligencia consciente y todas las nociones relevantes en el desarrollo de inteligencia artificial que, hasta hoy, creemos únicas de la humanidad. Si seguimos manteniendo este pensamiento, puede llegar el día en que sea demasiado tarde para tomar medidas. Debemos enfocarnos en cambiar el paradigma de la responsabilidad y que esta responsabilidad no solo nos afecte individualmente, sino también colectivamente como sociedad. Este cambio solo será inminente si primero aceptamos y reconocemos la implementación de estas tecnologías como sociedad (Kukita, 2017).

Las máquinas que se mejoran a sí mismas plantean la posibilidad de que no puedan ser desconectadas en cualquier momento, ya que podrían calcular y resolver esta situación para hacerla imposible (Omohundro, 2016). Debemos empezar a ver la relación que se está generando como una relación mutualista, en la que incluso podríamos servir a las máquinas antes de que nos sirvan a nosotros. Podría llegar un momento en que las máquinas sean el motor de la sociedad y no nosotros. Aunque en el ejemplo anterior fui yo quien utilizó a ChatGPT introduciendo datos y beneficiándome de su funcionamiento, en el futuro podríamos convertirnos en lo mismo para los robots, meros flujos de información que, dependiendo de ciertos datos, cumplen una utilidad para las máquinas inteligentes. La filosofía se encuentra frente a una oportunidad única, al igual que cuando Descartes descubrió la certeza de la conciencia a través del "Cogito ergo sum". El paradigma en el que estamos ahora es completamente diferente y está en constante cambio, por lo que debemos abordarlo desde el nuevo contexto que se presenta ante nosotros y que genera tensiones al pensar en el desarrollo humano y su bienestar.

## **Bibliografía**

Allen, C., Wallach, W., & Smit, I. (2011). Why machine ethics? In *Machine Ethics* (Vol. 9780521112352). <https://doi.org/10.1017/CBO9780511978036.005>

Andersen, L. (2021). Human Rights in the Age of Artificial Intelligence. *Journal of Legal, Ethical and Regulatory Issues*, 24.

Anderson, M., & Anderson, S. L. (2007). Machine ethics: Creating an ethical intelligent agent. *AI Magazine*, 28(4). <https://doi.org/10.4324/9781003075011-8>

Bosman, J. (2016). Top 9 ethical issues in artificial intelligence | World Economic Forum. In *World Economic Forum*.

Bostrom, N., & Yudkowsky, E. (2014). The ethics of artificial intelligence. In *The Cambridge Handbook of Artificial Intelligence*. <https://doi.org/10.1017/cbo9781139046855.020>

Bostrom, N., Ettinger, R. C. W., & Tandy, C. (2004). Death and Anti-Death, Volume 2: Two Hundred Years After Kant, Fifty Years After Turing. *Soil Sci Soc Am J*, 64.

Bradley, T. (2017). Facebook AI Creates Its Own Language In Creepy Preview Of Our Potential Future. *Forbes*.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora necessarily contain human biases. *Science*, 356(6334).

Dennett, D. (1994). Artificial Life as Philosophy. *Artificial Life*, 1(3). <https://doi.org/10.1162/artl.1994.1.3.291>

Everitt, T., & Hutter, M. (2018). Universal artificial intelligence: Practical agents and fundamental challenges. In *Studies in Systems, Decision and Control* (Vol. 117). [https://doi.org/10.1007/978-3-319-64816-3\\_2](https://doi.org/10.1007/978-3-319-64816-3_2)

Floridi, L., Cows, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles,

and Recommendations. *Minds and Machines*, 28(4). <https://doi.org/10.1007/s11023-018-9482-5>

Friedman, B., & Nissenbaum, H. (1996). Bias in Computer Systems. *ACM Transactions on Information Systems*, 14(3). <https://doi.org/10.1145/230538.230561>

Función Pública. (1982). *Ley 23 de 1982*. Recuperado de <https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=3431>

Función Pública. (1993). *Ley 44 de 1993*. Recuperado de <https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=3429#:~:text=%2D%20Incurrir%C3%A1%20en%20prisi%C3%B3n%20de%20dos,expresa%20del%20titular%20del%20derecho>

Función Pública. (2015). *Decreto 1066 de 2015 Sector Administrativo del Interior*. Recuperado de <https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=76835>

Gips, J. (2011). Towards the ethical robot. In *Machine Ethics* (Vol. 9780521112352). <https://doi.org/10.1017/CBO9780511978036.015>

Goertzel, B. (2014). Artificial General Intelligence: Concept, State of the Art, and Future Prospects. *Journal of Artificial General Intelligence*, 5(1). <https://doi.org/10.2478/jagi-2014-0001>

Good, I. J. (1966). *Speculations Concerning the First Ultraintelligent Machine*. [https://doi.org/10.1016/s0065-2458\(08\)60418-0](https://doi.org/10.1016/s0065-2458(08)60418-0)

Greenwald, A. G., & Pettigrew, T. F. (2014). With malice toward none and charity for some: Ingroup favoritism enables discrimination. *American Psychologist*, 69(7). <https://doi.org/10.1037/a0036056>

Guarini, M. (2006). Particularism and the classification and reclassification of moral cases. In *IEEE Intelligent Systems* (Vol. 21, Issue 4). <https://doi.org/10.1109/MIS.2006.76>

Harari, Y. N. (2017). Reboot for the AI revolution. *Nature*, 550(7676). <https://doi.org/10.1038/550324a>

Harari, Y. N. (2018). *Homo Deus Breve historia del mañana* (1st ed.). DEBATE.

Harari, Y. N. (2018). *21 lecciones para el siglo XXI* (1st ed.). DEBATE.

Hugo De, G., & Goertzel, B. (2009). Report on the first conference on artificial general intelligence (AGI-08). *AI Magazine*, 30(1). <https://doi.org/10.1609/aimag.v30i1.2151>

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9). <https://doi.org/10.1038/s42256-019-0088-2>

John P, S. (2011). When is a robot a moral agent? In *Machine Ethics* (Vol. 9780521112352). <https://doi.org/10.1017/CBO9780511978036.010>

Kukita, M. (2017). *When HAL Kills, Stop Asking Who's to Blame*.

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, 1(11). <https://doi.org/10.1038/s42256-019-0114-4>

Moor, J. H. (2011). The nature, importance, and difficulty of machine ethics. In *Machine Ethics* (Vol. 9780521112352). <https://doi.org/10.1017/CBO9780511978036.002>

Moor, J. H. (1985). WHAT IS COMPUTER ETHICS? *Metaphilosophy*, 16(4). <https://doi.org/10.1111/j.1467-9973.1985.tb00173.x>

Moor, J. H. (2017). Are there decisions computers should never make? In *Computer Ethics*. <https://doi.org/10.4324/9781315259697-39>

Nemitz, P. (2018). Constitutional democracy and technology in the age of artificial intelligence. In *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* (Vol. 376, Issue 2133). <https://doi.org/10.1098/rsta.2018.0089>

Omohundro, S. M. (2008). The basic AI drives. *Frontiers in Artificial Intelligence and Applications*, 171(1). <https://doi.org/10.18254/s207751800009748-1>

Omohundro, S. (2007). The nature of self-improving artificial intelligence. *Singularity Summit*, 5.

Omohundro, S. (2016). Journal of Experimental & Theoretical Artificial Intelligence Autonomous technology and the greater human good. *Journal of Experimental & Theoretical Artificial Intelligence*, 263.

OpenAI. (2023). *OpenAI Terms of Use*. Recuperado de <https://openai.com/policies/terms-of-use>

ChatGPT. (2023, 10 de julio). "Cuando se creó la primera computadora" y "Cuando se creó el machine learning" [Conversación en línea]. Mensaje publicado en <https://conv.example.com/chatgpt/conversation>

OpenAI. (2023). La Ilustración de Kant: El poder de la razón y la emancipación intelectual. [ChatGPT iOS app]. OpenAI.

OpenAI. (2023, July 7). ChatGPT iOS App Conversation [Chat log]. iOS App Version 1.0.0. Tomado de <https://chat.openai.com/c/5e4aa563-e0bb-467b-8c19-702b580d63de>

Organización de los Estados Americanos (OEA). (1993). *Decisión 351: Régimen común sobre Derecho de autor y Derechos conexos*. Recuperado de <http://www.sice.oas.org/trade/junac/decisiones/dec351s.asp>

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3). <https://doi.org/10.1017/S0140525X00005756>

Siau, K., & Wang, W. (2020). Artificial intelligence (AI) Ethics: Ethics of AI and ethical AI. In *Journal of Database Management* (Vol. 31, Issue 2). <https://doi.org/10.4018/JDM.2020040105>

Universidad Nacional de Colombia. (2015). *CONCEPTO No. 02 de 2015*. Recuperado de [http://www.legal.unal.edu.co/rlunal/home/doc.jsp?d\\_i=87497](http://www.legal.unal.edu.co/rlunal/home/doc.jsp?d_i=87497)

Universidad Nacional de Colombia. (2003). *ACUERDO 035 de 2003*. Recuperado de [http://www.legal.unal.edu.co/rlunal/home/doc.jsp?d\\_i=34248#:~:text=ART%C3%8DCULO%2014.&text=El%20derecho%20de%20autor%20protege,re%C3%BAnen%20caracter%C3%ADsticas%20de%20creaci%C3%B3n%20original](http://www.legal.unal.edu.co/rlunal/home/doc.jsp?d_i=34248#:~:text=ART%C3%8DCULO%2014.&text=El%20derecho%20de%20autor%20protege,re%C3%BAnen%20caracter%C3%ADsticas%20de%20creaci%C3%B3n%20original)

Universidad Nacional de Colombia. (2012). *CIRCULAR No. 1 de 2012*. Recuperado de [http://www.legal.unal.edu.co/rlunal/home/doc.jsp?d\\_i=44925](http://www.legal.unal.edu.co/rlunal/home/doc.jsp?d_i=44925)

Urbina, F., Lowden, C. T., Culberson, J. C., & Ekins, S. (2022). MegaSyn: Integrating Generative Molecular Design, Automated Analog Designer, and Synthetic Viability Prediction. *ACS Omega*, 7(22). <https://doi.org/10.1021/acsomega.2c01404>

VINGE, V. (1993). The coming singularity: how to survive in the post-human era. *Vision 21: Interdisciplinary Science and Engineering in the Era of Cyberspace*.

Wallach, W., & Allen, C. (2009). Moral Machines: Teaching Robots Right from Wrong. In *Moral Machines: Teaching Robots Right from Wrong*. <https://doi.org/10.1093/acprof:oso/9780195374049.001.0001>

Whittlestone, J., Nyrupe, R., Alexandrova, A., Dihal, K., & Cave, S. (2019). Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research. *Nuffield Foundation*.

## **Bibliografía Secundaria**

Allen, C., Smit, I., & Wallach, W. (2005). Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and Information Technology*, 7(3). <https://doi.org/10.1007/s10676-006-0004-4>

Anderson, S. L. (2011). Machine metaethics. In *Machine Ethics* (Vol. 9780521112352). <https://doi.org/10.1017/CBO9780511978036.003>

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine Bias: Risk Assessments in Criminal Sentencing. ProPublica. Retrieved from <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

Arkoudas, K., & Bringsjord, S. (2014). Philosophical foundations. In *The Cambridge Handbook of Artificial Intelligence*. <https://doi.org/10.1017/cbo9781139046855.004>

Bedau, M. A. (2003). Artificial life: Organization, adaptation and complexity from the bottom up. In *Trends in Cognitive Sciences* (Vol. 7, Issue 11). <https://doi.org/10.1016/j.tics.2003.09.012>

Bradley, T. (2017, July 31). Facebook AI Creates Its Own Language in Creepy Preview of Our Potential Future. *Forbes*. Retrieved from <https://www.forbes.com/sites/tonybradley/2017/07/31/facebook-ai-creates-its-own-language-in-creepy-preview-of-our-potential-future/>

Bundy, A. (2017). Preparing for the future of Artificial Intelligence. *AI & SOCIETY*, 32(2). <https://doi.org/10.1007/s00146-016-0685-0>

Bynum, T. W. (2012). The historical roots of information and computer ethics. In *The Cambridge Handbook of Information and Computer Ethics*. <https://doi.org/10.1017/cbo9780511845239.003>

Deep Mind (2023, Julio). RoboCat: A Self-Improving Robotic Agent. DeepMind Blog. Retrieved from <https://www.deepmind.com/blog/robocat-a-self-improving-robotic-agent>

Dennett, D. C. (2020). When HAL Kills, Who's to Blame? Computer Ethics. In *HAL's Legacy* (pp. 351–365). <https://doi.org/10.7551/mitpress/3404.003.0018>

Dwoskin, E. (2018, July 26). Amazon Pushes ICE to Buy Its Face Recognition Surveillance Tech. *The New York Times*. Retrieved from <https://www.nytimes.com/2018/07/26/technology/amazon-aclu-facial-recognition-congress.html>

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. In *Minds and Machines* (Vol. 14, Issue 3). <https://doi.org/10.1023/B:MIND.0000035461.63578.9d>

Grossman, L. (2014, June 9). Computer Program Claims to Have Passed the Turing Test. *TIME*. Retrieved from <https://time.com/2847900/eugene-goostman-turing-test/>

Greene, D., Hoffmann, A. L., & Stark, L. (2019). Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning. *Proceedings of the Annual Hawaii International Conference on System Sciences, 2019-January*. <https://doi.org/10.24251/hicss.2019.258>

J. Storrs, H. (2011). Ethics for machines. In *Machine Ethics* (Vol. 9780521112352). <https://doi.org/10.1017/CBO9780511978036.004>

Jolie, A. (2013, May 13). My Medical Choice. The New York Times. Retrieved from <https://www.nytimes.com/2013/05/14/opinion/my-medical-choice.html>

Johnson, D. G. (2011). Computer systems: Moral entities but not moral agents. In *Machine Ethics* (Vol. 9780521112352). <https://doi.org/10.1017/CBO9780511978036.012>

King, T. C., Aggarwal, N., Taddeo, M., & Floridi, L. (2020). Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions. *Science and Engineering Ethics*, 26(1). <https://doi.org/10.1007/s11948-018-00081-0>

Koolen, C., & van Cranenburgh, A. (2017). These are not the Stereotypes You are Looking For: Bias and Fairness in Authorial Gender Attribution. *EACL 2017 - Ethics in Natural Language Processing, Proceedings of the 1st ACL Workshop*. <https://doi.org/10.18653/v1/w17-1602>

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A proposal for the Dartmouth summer research project on artificial intelligence. *AI Magazine*, 27(4).

McCarthy, J. (2021). *What is AI? / Basic Questions*. Stanford.

McDermott, D. (2011). What Matters to a Machine? In *Machine Ethics*. <https://doi.org/10.1017/cbo9780511978036.010>

Mitcham, C. (1996). Thinking Through Technology: The Path between Engineering and Philosophy, Carl Mitcham. 1994. University of Chicago Press, Chicago, IL. *Bulletin of Science, Technology & Society*, 16(3). <https://doi.org/10.1177/027046769601600315>

Nagel, T. (2016). WHAT IS IT LIKE TO BE A BAT? In *Journey into Philosophy: An Introduction with Classic and Contemporary Readings*. <https://doi.org/10.1017/s1477175616000336>

Peter Stone, E. (2016). Artificial Intelligence and Life in 2030. *One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel, September*.

Robinson, W. S. (2014). Philosophical challenges. In *The Cambridge Handbook of Artificial Intelligence*. <https://doi.org/10.1017/cbo9781139046855.005>

Sullins, J. P. (2005). Ethics and artificial life: From modeling to moral agents. *Ethics and Information Technology*, 7(3). <https://doi.org/10.1007/s10676-006-0003-5>

Sweeney, L. (2013). Discrimination in online Ad delivery. *Communications of the ACM*, 56(5). <https://doi.org/10.1145/2447976.2447990>

Taddeo, M. (2018). The Limits of Deterrence Theory in Cyberspace. *Philosophy and Technology*, 31(3). <https://doi.org/10.1007/s13347-017-0290-2>

The Washington Post. (2023, March 5). How AI Voice Technology is Being Used in Scams. Retrieved from <https://www.washingtonpost.com/technology/2023/03/05/ai-voice-scam/>

Toronto Sun. (2016, May 13). Woman follows GPS, ends up in Ontario lake. Retrieved from <https://torontosun.com/2016/05/13/woman-follows-gps-ends-up-in-ontario-lake>

Torrance, S. (2011). Machine ethics and the idea of a more-than-human moral world. In *Machine Ethics* (Vol. 9780521112352). <https://doi.org/10.1017/CBO9780511978036.008>

Turkle, S. (2011). Authenticity in the age of digital companions. In *Machine Ethics* (Vol. 9780521112352). <https://doi.org/10.1017/CBO9780511978036.006>

Turner A. (2023). *How Many People Have Smartphones Worldwide (Jun 2023)*. BankMyCell.

Vincent, J. (2018, January 12). Google's racist algorithm comes back to haunt the company. The Verge. Retrieved from <https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai>

Virtuoso.tech. (s. f.). AI Brings More Convincing Phishing Scams for Business Owners. Recuperado de <https://virtuoso.tech/ai-brings-more-convincing-phishing-scams-for-business-owners/>

(No author mentioned). (2023, May 22). A.I. That Labels Photos From Google and Apple Comes Under Scrutiny. The New York Times. Retrieved from <https://www.nytimes.com/2023/05/22/technology/ai-photo-labels-google-apple.html>

(No author mentioned). (s. f.). Autonomous Vehicles (AVs) are Facing Their Own Version of the Trolley Problem. Futurity. Retrieved from <https://www.futurity.org/autonomous-vehicles-av-ethics-trolley-problem-2863992-2/>

(No author mentioned). (s. f.). Billionaire Wealth Has Surged by \$5.5 Trillion During the Pandemic — But Elon Musk Still Thinks a Tax on the Rich is a 'Power Grab'. Global Citizen. Retrieved from <https://www.globalcitizen.org/en/content/wealth-inequality-oxfam-billionaires-elon-musk/>

(No author mentioned). (s. f.). ChatGPT-4 Reinforces Sexist Stereotypes. Equality Now. Retrieved from [https://www.equalitynow.org/news\\_and\\_insights/chatgpt-4-reinforces-sexist-stereotypes/](https://www.equalitynow.org/news_and_insights/chatgpt-4-reinforces-sexist-stereotypes/)

(No author mentioned). (2023, March). March 2023 Global Poverty Update: World Bank Challenge Estimating Poverty Pandemic. World Bank Blogs - Open Data. Retrieved from <https://blogs.worldbank.org/opendata/march-2023-global-poverty-update-world-bank-challenge-estimating-poverty-pandemic>

(No author mentioned). (2022, November 28). Meta fined €227 million by Ireland's DPC over data leak. The Verge. Retrieved from <https://www.theverge.com/2022/11/28/23481786/meta-fine-facebook-data-leak-ireland-dpc-gdpr>

(No author mentioned). (2019, February 14). OpenAI restricts access to dangerous text-generating AI. Nature. Retrieved from <https://www.nature.com/articles/d41586-019-00673-1>

(No author mentioned). (s. f.). Our Machines Now Have Knowledge We'll Never Understand. Wired. Retrieved from <https://www.wired.com/story/our-machines-now-have-knowledge-well-never-understand/>

n.d. (2017). *Report of COMEST on Robotics Ethics*. <https://unesco.blob.core.windows.net/pdf/UploadCKEditor/REPORT%20OF%20COMEST%20ON%20ROBOTICS%20ETHICS%2014.09.17.pdf>

(No author mentioned). (2023, February). Silicon Valley Inequality. CalMatters. Retrieved from <https://calmatters.org/california-divide/2023/02/silicon-valley-inequality/>

(No author mentioned). (2016, January 23). Survival of the Richest. Oxfam Library. Retrieved from <https://oxfamilibrary.openrepository.com/bitstream/handle/10546/621477/bp-survival-of-the-richest-160123-en.pdf>

(No author mentioned). (2023, May 3). Which countries are trying to regulate artificial intelligence? Euronews. Retrieved from <https://www.euronews.com/next/2023/05/03/which-countries-are-trying-to-regulate-artificial-intelligence>