Una breve mirada al estado actual de la Inteligencia Artificial

Marcelo Arenas, Gabriela Arriagada, Marcelo Mendoza y Claudia Prieto

1. Una breve historia de la inteligencia artificial

La mecanización de procesos ha sido fundamental para el desarrollo de la humanidad. Los enormes desafíos que ha significado llevar a cabo esta mecanización nos ha llevado a estudiar sus límites. Con el desarrollo de la lógica matemática a finales del siglo XIX y principios del siglo XX se establecieron las bases de una formalización matemática de la idea de proceso mecánico o algoritmo. En particular, el trabajo de Alan Turing en los años 30 llevó a la formalización matemática de la noción de algoritmo a través de lo que hoy se conoce como la Máquina de Turing [T37], sentando las bases para la teoría de la computación. Tanto la lógica matemática como la teoría de la computación han sido fundamentales en el desarrollo de la inteligencia artificial, como se describe a continuación.

El primer trabajo reconocido generalmente como Inteligencia Artificial fue la neurona artificial de Warren McCulloch y Walter Pitts en 1943, que es una unidad de cálculo que intenta modelar el comportamiento de una neurona [MP43]. Esta neurona artificial es el componente esencial de las redes neuronales artificiales, las cuales son una de las tecnologías más importantes desarrolladas en inteligencia artificial.

La conferencia de Dartmouth en 1956 es considerada como el punto de partida de la inteligencia artificial moderna.[1] Esta conferencia reunió a investigadores de diversas disciplinas para discutir la posibilidad de enseñar a los computadores a pensar como seres humanos. Este evento sentó las bases para el campo de estudio de la inteligencia artificial, y de hecho en ella se acuñó el término inteligencia artificial.

Estimulada por las discusiones de la conferencia de Dartmouth, la investigación en inteligencia artificial de los años 50 comienza a desarrollarse en dos corrientes, basadas en los conceptos que describimos anteriormente, y que tenían visiones distintas, y en algunos sentidos contrapuestas, sobre cómo debía desarrollarse esta disciplina. Estas dos corrientes, que

[1] https://en.wikipedia.org/wiki/Dartmouth_workshop

describiremos a continuación, se siguen desarrollando hasta el día de hoy, y han sido fundamentales para el crecimiento de la inteligencia artificial.

Por un lado, la inteligencia artificial simbólica (symbolic AI) es una rama de la inteligencia artificial que se basa en el procesamiento y manipulación de símbolos y reglas lógicas para la representación del conocimiento y razonamiento sobre él. Se enfoca en la utilización de técnicas de lógica matemática, inferencia y búsqueda para resolver problemas complejos y realizar tareas cognitivas, como la planificación, toma de decisiones, razonamiento deductivo e interpretación de datos. La inteligencia artificial simbólica se basa en la representación y manipulación de conocimiento simbólico estructurado, y es utilizada en áreas como la robótica, la planificación de rutas y la medicina.

Por otro lado, el enfoque conexionista (connectionist approach) en inteligencia artificial se basa en el uso de redes neuronales artificiales para simular el funcionamiento del cerebro humano en la realización de tareas. Estas redes neuronales están compuestas por nodos interconectados (como la neurona artificial de McCulloch y Pitts [MP43]) que procesan y transmiten información a través de conexiones ponderadas. El aprendizaje se realiza ajustando los pesos de las conexiones en base a los datos de entrenamiento, permitiendo que la red aprenda patrones y representaciones a partir de datos. El enfoque conexionista ha demostrado ser efectivo en áreas como reconocimiento de patrones, procesamiento del lenguaje natural y visión por computadora, y se ha utilizado en aplicaciones como reconocimiento de imágenes, diagnóstico médico y sistemas de recomendación.

El aprendizaje de máquina (machine learning) juega un rol fundamental en el enfoque conexionista basado en redes neuronales artificiales. Un importante hito en esta área fue el desarrollo por Frank Rosenblatt en 1958 del primer algoritmo de aprendizaje supervisado, llamado Perceptrón [R58]. Este algoritmo de aprendizaje sentó las bases para el desarrollo posterior de algoritmos de aprendizaje de máquina, que han sido fundamentales para el desarrollo de la inteligencia artificial.

En los años 60 y 70 se creía que los enfoques simbólicos eventualmente lograrían crear una máquina con inteligencia artificial general. Sin embargo, los resultados no fueron los esperados y el interés en el área de la inteligencia artificial decayó. En los años 80, la investigación en inteligencia artificial fue revivida por el éxito comercial de los sistemas expertos, una forma de programa de inteligencia artificial que simulaba el conocimiento y las habilidades analíticas de expertos humanos. Aunque los sistemas expertos generaron mucho interés, nuevamente la atracción por la inteligencia artificial decayó por la complejidad y el alto costo de desarrollo de estos sistemas, y por las limitaciones tecnológicas de la época. Incluso se dice que la inteligencia artificial cayó en un segundo invierno.

A finales del siglo XX y principios del siglo XXI, la inteligencia artificial se concentró en desarrollar soluciones a problemas específicos, en lo cual jugó un rol fundamental el uso de

técnicas de otras disciplinas como estadística, optimización y economía. Este enfoque permitió producir resultados verificables que empezaron a ser ampliamente utilizados.

El aprendizaje de máquina tuvo un avance fundamental en la última década, en lo que es conocido como aprendizaje profundo. Esta forma de aprendizaje es una rama del aprendizaje automático que utiliza redes neuronales artificiales con múltiples capas [LBH15]. En particular, Alex Krizhevsky, Ilya Sutskever y Geoffrey Hinton desarrollaron AlexNet [KSE12], una red neuronal profunda que ganó en el año 2012 un importante desafío en la clasificación de imágenes. Esto abrió nuevas posibilidades en el campo de la visión por computadora y allanó el camino para aplicaciones de inteligencia como reconocimiento de imágenes y de voz.

Es importante destacar que la disponibilidad de mayor poder de cálculo, en particular en la forma de unidades de procesamiento gráfico o GPUs, y el acceso a grandes volúmenes de datos digitalizados permitió el avance del aprendizaje profundo, y nos ha llevado a avances fundamentales en el procesamiento de lenguaje natural, como el desarrollo de ChatGPT.

2. Aprendizaje de máquina y aprendizaje profundo

En esta sección, se presenta una breve introducción al aprendizaje de máquina dentro de inteligencia artificial, y en particular al aprendizaje profundo que ha tenido un enorme desarrollo en los últimos años.

El aprendizaje de máquina es un área de la inteligencia artificial en la cual se desarrollan algoritmos que pueden aprender modelos desde datos, con el objetivo de que estos modelos puedan ser usados en predicciones posteriores. Por ejemplo, podemos pensar en un modelo meteorológico para predecir si va a llover. Los datos en este caso corresponden a medidas de ciertas variables meteorológicas registradas en ciertas fechas, tales como temperatura, humedad del aire y presión atmosférica, junto con la información sobre si llovió en esas fechas. El algoritmo en este caso debe aprender un modelo que toma como entrada las variables meteorológicas consideradas, y entrega como salida un número que indica si va a llover o no, o en un modelo más general un número que representa la probabilidad de que llueva dados los valores de las variables meteorológicas.

Los componentes esenciales del aprendizaje de máquina son entonces:

1. los datos de entrenamiento utilizados por los algoritmos de aprendizaje para construir modelos;
2. los algoritmos de aprendizaje, que son métodos matemáticos utilizados para entrenar y mejorar los modelos;

3. los modelos, que son representaciones matemáticas de las relaciones entre los datos utilizados para hacer predicciones;
4. los métodos de evaluación y validación, que implican el uso de datos de prueba para medir el desempeño del modelo y su capacidad de generalización; y
5. los métodos de optimización y ajuste, que permiten la mejora continua de los modelos.

El aprendizaje profundo (deep learning) es una sub-área del aprendizaje de máquina en el cual los modelos generados por los algoritmos de aprendizaje son redes neuronales artificiales con múltiples capas [LBH15, GBC16]. El aprendizaje profundo tiene los mismos componentes esenciales que el aprendizaje de máquina, pero se pueden distinguir algunas características particulares:

1. para el aprendizaje de los modelos es necesario contar con grandes volúmenes de datos, dada la complejidad de estos modelos y los resultados precisos que se espera obtener;
2. los algoritmos de aprendizaje son esencialmente algoritmos de optimización, que se utilizan para entrenar los modelos de forma eficiente y ajustar los pesos de las conexiones entre neuronas,[2] que son denominados parámetros de la red neuronal; y
3. los modelos son redes neuronales profundas, que son arquitecturas con múltiples capas de neuronas artificiales para aprender características y patrones complejos de los datos.

Además, es importante considerar que dado el volumen de datos y la complejidad de los modelos a utilizar, en general se necesita de hardware especializado, como GPUs, para realizar el aprendizaje de los modelos.

3. Procesamiento de lenguaje natural

El procesamiento del lenguaje natural (PLN) es un área de la inteligencia artificial que se enfoca en la interacción entre humanos y máquinas mediante el uso del lenguaje. Estudia tareas de análisis léxico, morfológico, sintáctico y semántico. También aborda tareas relacionadas con la generación de lenguaje, como traducción de máquina y generación de texto. El procesamiento del lenguaje natural facilita el desarrollo de sistemas de traducción automática, generación de resúmenes de texto y chatbots, entre otras aplicaciones.

A continuación explicamos las razones y principales hitos por los cuales el PLN se ha transformado en tema central en el área de inteligencia artificial en la última década.

[2] El ejemplo más importante de este tipo de algoritmos es backpropagation [RHW86].

3.1 Aprendizaje de representaciones: desde word2vec a ELMO

El PLN ha experimentado avances significativos en la última década gracias al uso de word embeddings, que son representaciones de palabras que capturan su significado. Un word embedding es una representación que mapea cada palabra en un vector numérico denso de longitud fija. Este vector representa la semántica de la palabra, lo que significa que palabras similares se asignan a vectores similares. El siguiente es un ejemplo simple de un word embedding para algunas palabras en inglés:

- car: [0.25, 0.68, 0.42, -0.11]
- bus: [0.22, 0.71, 0.38, -0.08]
- train: [0.23, 0.67, 0.45, -0.07]
- plane: [0.13, 0.54, 0.62, 0.09]

Como se puede ver, los vectores para palabras relacionadas como "car", "bus" y "train" son similares entre sí, mientras que el vector para "plane" es distinto a los otros.

El primer modelo de word embeddings fue word2vec [MK13]. Este modelo usa redes neuronales para aprender representaciones de palabras a partir de grandes volúmenes de texto. Word2vec superó en su época a otros enfoques de PLN en tareas como analogías de palabras y clasificación de textos.

Posteriormente surgieron otros modelos más avanzados de word embeddings, como FastText [BO17]. FastText es un modelo basado en sub-palabras lo que le permite producir representaciones de palabras que no se encuentran en el corpus de entrenamiento. Otro avance introducido por FastText consistió en entrenar estos modelos en corpus multilingües, permitiendo desarrollar sistemas PLN en idiomas como el castellano.

ElMo (Embeddings for Language Models) fue introducido por Peters et al. [PE18] para generar representaciones de palabras dependientes del contexto en el cual se usan. El generador de embeddings de ElMO produce word embeddings condicionados a la oración en la cual la palabra se usa. ElMO fue capaz de superar a los modelos anteriores en tareas de índole semántico, como etiquetado de rol semántico. El etiquetado de rol semántico es una tarea de PLN en la que se asigna a cada palabra en una oración un papel semántico específico. Un ejemplo de etiquetado de roles semánticos sería el siguiente:

Oración: "Juan compró una manzana en la tienda"

Etiquetado de roles semánticos:

- "Juan" es el agente (el que realiza la acción).
- "compró" es el predicado (la acción que se realiza).
- "una manzana" es el objeto (sobre el cual se realiza la acción).

- "en la tienda" es el lugar (donde se realiza la acción).

3.2 Transformers y auto-atención

El surgimiento de nuevas arquitecturas de redes neuronales artificiales como la arquitectura Transformer [VA17] influyeron de forma decisiva en los avances más sustanciales del PLN. La arquitectura Transformer es una arquitectura de aprendizaje profundo que procesa en paralelo a través de múltiples capas las palabras de una oración. Cada una de estas capas realiza una transformación no lineal de la entrada. La arquitectura introduce un mecanismo de auto-atención que permite que la red preste atención simultáneamente a diferentes elementos de la entrada, capturando relaciones de dependencia de largo plazo en el texto. La arquitectura se usa como codificador-decodificador para resolver tareas simples, como la predicción de la siguiente palabra de un texto. Sin embargo, a partir de esta tarea tan simple, y debido a que el volumen de textos que se le muestra a la máquina es tan grande, se codifica en la máquina información de mayor nivel de abstracción como estructuras sintácticas y asociaciones entre conceptos. Por esta razón los Transformers se usan tanto como codificadores de texto, como en el caso de BERT (Bidirectional Encoder Representations from Transformers) [DE19] así como generadores de texto.

3.3 Generative pre-trained transformers

El decodificador de la arquitectura Transformer permitió el desarrollo de modelos de lenguaje generativos. Estos modelos se entrenan en base a grandes cantidades de texto. El tipo de entrenamiento de estos modelos es no supervisado y se basa en tareas de generación autoregresiva. La generación autoregresiva de texto es una técnica que consiste en predecir la siguiente palabra o caracteres en una secuencia de texto, basándose en las palabras o caracteres previos en la secuencia. Se trata de un modelo que aprende a producir un flujo continuo de texto que sigue una estructura gramatical y semántica coherente, basándose en un corpus de entrenamiento previamente establecido. Este enfoque se basa en el uso de modelos de lenguaje de secuencia, como los modelos de lenguaje basados en RNN (redes neuronales recurrentes) o los modelos de lenguaje basados en Transformers, que son capaces de aprender la probabilidad condicional de cada palabra en la secuencia, dada la historia previa de la secuencia. En la práctica, la generación autoregresiva se utiliza en una amplia gama de aplicaciones de NLP, como la generación de texto para chatbots, la generación de respuestas automáticas a correos electrónicos o la creación de resúmenes de texto.

El mecanismo de entrenamiento de los generadores de texto se basa en la proximidad que existe entre el texto sintético generado y el texto real. Otra característica de estos modelos es su tamaño. Los Generative pre-trained transformers (GPT) son entrenados sobre datos enormes como BookCrawl (corpus que contiene más de mil millones de palabras recopiladas de libros en

inglés disponibles en línea), WebText (colección de documentos web seleccionados al azar, que abarcan una amplia gama de temas, desde deportes y tecnología hasta ciencia y política) y Common Crawl (corpus que contiene miles de millones de páginas web rastreadas y descargadas de la World Wide Web). GPT-1 [ST20] usó una arquitectura Transformer con 12 capas y 117 millones de parámetros. GPT-2 usó una arquitectura de 48 capas y 1.5 billones de parámetros. GPT-3 usó una arquitectura de 96 capas y 175 billones de parámetros. El tamaño de estos modelos ha permitido que generen texto coherente condicionado a consultas escritas en lenguaje natural que se codifican en el espacio de representación del modelo GPT. Las capacidades de estos modelos les permiten obtener resultados sorprendentes en traducción automática de textos y construcción de resúmenes.

3.4 Chat GPT y chatbots

Sin duda el modelo generativo de texto más conocido hoy es Chat GPT. Chat GPT es una extensión de GPT-3 que fue entrenado para responder preguntas y generar respuestas en lenguaje natural. Una capacidad que tiene Chat GPT es la de incorporar al mecanismo generativo de texto una conversación, poniéndo énfasis en la retroalimentación que obtiene en la interacción con el humano. Esto le permite generar respuestas coherentes en el contexto de una conversación. A este tipo de interacción humano-máquina se le denomina chatbot.

Chat GPT incorporó un tipo de entrenamiento denominado aprendizaje reforzado para mejorar la calidad de sus respuestas. El aprendizaje reforzado incorpora retroalimentación durante el proceso de entrenamiento. Esta retroalimentación fue usada para producir la primera versión de Chat GPT, basada en GPT 3.5. También incorpora este mecanismo de refuerzo en producción, lo cual le permite mejorar sus respuestas. Otra característica fundamental de Chat GPT es que es multilingüe.

Una limitación que tiene Chat GPT es que requiere ser reentrenado para incorporar información actualizada. Esto limita su capacidad para interactuar en torno de temas de actualidad. Esta limitación podría ser abordada por GPT-4, cuyo lanzamiento fue anunciado el 13 de marzo de 2023 por OpenAI. GPT-4 anuncia disponer de una mayor base de conocimiento para interactuar en torno de temas de actualidad. También es capaz de manejar contextos más largos, de hasta 25,000 tokens, favoreciendo la creación de contenidos, conversaciones extensas y análisis automático de documentos. GPT-4 acepta imágenes en la entrada, lo cual le permite realizar tareas de análisis bimodal.

4. Aplicaciones de la Inteligencia Artificial

La disponibilidad de mayor poder de cálculo y el acceso a grandes volúmenes de datos digitalizados han permitido avances fundamentales en inteligencia artificial durante las últimas décadas, incluyendo adelantos a gran velocidad en las áreas de procesamiento de imágenes y procesamiento de lenguaje natural. Estos avances han permitido su aplicación en prácticamente todos los sectores productivos, servicios y áreas del saber. Ejemplos de estos sectores incluyen educación, salud, agricultura, transporte, finanzas, recursos humanos, ciberseguridad, redes sociales, entretenimiento, astronomía y robótica, entre muchos otros. Para ejemplificar el estado actual, potencial impacto y desafíos de las aplicaciones de la inteligencia artificial, en el siguiente apartado primero nos enfocamos en salud, para luego dar dos ejemplos de aplicaciones de inteligencia artificial en el ámbito legal desarrollados en Chile.

4.1 Inteligencia artificial en salud

La atención médica y el cuidado de la salud están bajo una gran presión debido al aumento de costos, la escasez de personal altamente calificado, el diagnóstico tardío de muchas enfermedades, el acceso limitado y desigual a servicios de salud, el envejecimiento creciente de la población y el aumento de enfermedades crónicas.

El desarrollo de nuevas tecnologías, incluyendo avances fundamentales en inteligencia artificial, son cada vez más importantes para hacer frente a estos retos. En particular se espera que la inteligencia artificial tenga un rol preponderante en:

- Generación y análisis de cantidades masivas de datos digitalizados generados por los pacientes, los que incluyen datos genéticos, pruebas de laboratorio, historias clínicas electrónicas, parámetros fisiológicos monitoreados por sensores portátiles, imágenes médicas, reportes médicos, parámetros fisiológicos monitoreados por dispositivos en el hogar a través del internet de las cosas (IoT), entre muchos otros.
- Aumento de la productividad de los profesionales de la salud mediante la automatización y simplificación de tareas, permitiendo así un aumento de capacidades y la concentración de los esfuerzos de los profesionales de la salud en la atención de pacientes.
- Reducción de costos mediante el desarrollo de procesos y tecnologías más eficientes y menos costosas (por ejemplo, adquisiciones más rápidas de imágenes médicas, dispositivos médicos menos costosos, etc.) y gracias a la toma de mejores decisiones relacionadas a la salud realizadas de manera más temprana (por ejemplo, diagnóstico

temprano de enfermedades que evitan intervenciones complejas y/o costosos tratamientos posteriores).

- Cambio de paradigma desde tratamiento centrado en la enfermedad hacia una gestión de la salud centrada en el paciente mediante avances hacia medicina predictiva, preventiva y personalizada, potenciados por el procesamiento de cantidades masivas de datos de pacientes y biobancos.
- Mejora de la accesibilidad a servicios de salud mediante la automatización y simplificación de tareas y la atención médica remota en tiempo real, permitiendo mayor accesibilidad a regiones remotas/áreas de la población que no cuentan con acceso a suficientes profesionales de la salud altamente calificados o a tecnología de punta de alto costo y/o de difícil operación.

Según Zion Market Research, el mercado mundial de la inteligencia artificial en salud fue de aproximadamente 1.400 millones de dólares en 2018 y se espera que alcance aproximadamente 17.800 millones de dólares en 2025 [DDF+23, AE23]. A modo de ejemplo, al 5 de octubre del 2022 existían 521 software y dispositivos médicos basados en inteligencia artificial aprobados por la FDA (lista generada por la FDA de acuerdo a información disponible públicamente)[3].

4.1.1 Áreas de aplicación de la inteligencia artificial en salud

Los desarrollos basados en inteligencia artificial tienen aplicaciones en una amplia gama de aspectos de la atención médica y cuidado de la salud, los cuales se pueden dividir de manera general en: gestión y administración, consulta y comunicación con el paciente, diagnóstico, tratamiento y pronóstico, y salud pública.

En el ámbito de la gestión y la administración, esto incluye la programación automatizada de citas médicas, la programación automatizada de citas de seguimiento, predicción de pacientes que con alta probabilidad no asistirán a su cita médica y mitigación de posibles ausencias, soluciones para mejorar la asignación de camas hospitalarias, mejora de la programación de quirófanos y salas de cirugías, mejora de la gestión de historiales médicos, mejora y automatización de la facturación de pagos, entre muchas otras.

En términos de consulta y comunicación con el paciente, esto incluye consultas médicas de manera remota, notas clínicas automatizadas, prescripción de medicamentos automatizada/asistida, chatbots para la interacción con el paciente, entre otros.

[3] https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices

En cuanto al diagnóstico, tratamiento y pronóstico, se incluye la detección y el diagnóstico precoz, preciso y eficiente, basados en innovaciones en radiología (imágenes médicas), patología, genómica, y otros campos; la generación de informes médicos automatizados; la monitorización remota de pacientes; la planificación y realización de cirugías/intervenciones asistidas de manera remota; el desarrollo de terapias y tratamiento con robótica médica; y el descubrimiento y desarrollo más eficiente de fármacos.

En el ámbito de la salud pública, esto implica listas de esperas, identificación de brotes de enfermedades, el apoyo a la respuesta ante brotes de enfermedades y la vigilancia de la salud pública, entre otros.

4.1.2 Desafíos para la adopción de la inteligencia artificial en salud

A pesar de los enormes avances en el ámbito de la investigación, la adopción de la inteligencia artificial en el área de la salud ha sido (y probablemente seguirá siendo) más lenta que en otros campos.

Hay un gran número de resultados de investigación recientes que sugieren que la inteligencia artificial puede rendir igual (o mejor) que los médicos en la toma de decisiones; los ejemplos van desde la medicina de precisión basada en genómica hasta el escaneado de retina y el análisis de imágenes radiológicas para detección temprana de enfermedades, incluyendo por ejemplo varios tipos de cáncer [LFK+19]. Así mismo, las empresas tecnológicas, como Google y Meta, entre varias otras, también están trabajando para construir modelos de predicción a partir de grandes volúmenes de datos y proporcionar un mejor apoyo en la toma de decisiones a los profesionales de la salud [DK19].

Sin embargo, todavía existen varios desafíos fundamentales que deben abordarse para permitir la adopción generalizada de soluciones basadas en inteligencia artificial en la práctica clínica. Esto porque en el área de la salud, los desarrollos de la inteligencia artificial deben guiarse por una constante preocupación en el impacto humano y clínico y, por lo tanto, se requiere prestar atención a las implicaciones éticas, legales y sociales de la inteligencia artificial, así como a los sesgos presentes en los datos y algoritmos; las preocupaciones relacionadas con la privacidad y la confidencialidad de los datos; las disparidades en salud; y las consecuencias sociales adversas y no intencionadas de la investigación y el desarrollo de la inteligencia artificial en salud.

Estos desafíos pueden clasificarse de manera general en:

- Disponibilidad y acceso a datos de diferentes modalidades (datos genéticos, imágenes, reportes médicos, listas de esperas, etc.) organizados y anotados/etiquetados que

permitan el entrenamiento adecuado de algoritmos y herramientas de inteligencia artificial.

- Generalización adecuada de las herramientas y algoritmos de inteligencia artificial a diferentes poblaciones y entornos. Es común escuchar del éxito de herramientas entrenadas en datos de un cierto centro médico, en una población específica o un lenguaje dado, que son exitosos cuando se evalúan en datos similares a los de entrenamiento, sin embargo, fallan o tienen resultados menos exitosos cuando se generalizan a nuevos datos. Por lo tanto, se necesitan importantes esfuerzos en la generalización de estas tecnologías entre diferentes centros médicos, cohortes de pacientes, características demográficas (edad/raza/género), etc.
- Generación de confianza, por ejemplo, mediante el desarrollo de algoritmos de control de calidad, y el desarrollo de modelos explicables que los profesionales de la salud puedan entender y en los que puedan confiar, y el desarrollo de modelos de inteligencia híbrida donde máquina y humanos trabajan juntos con el objetivo no de reemplazar, sino que de aumentar las capacidades humanas.
- Equidad y mitigación de sesgos no intencionados presentes en datos de entrenamiento y en los algoritmos.
- Privacidad y protección de pacientes y datos, por ejemplo, mediante el aprendizaje federado, un paradigma de aprendizaje que trata de resolver el problema de la gobernanza y la privacidad de los datos entrenando algoritmos en colaboración sin la necesidad de compartir los datos [RHL20].
- Formación de profesionales de la salud para la adopción de herramientas basadas en inteligencia artificial e integración de la inteligencia artificial en los planes de estudios de las carreras de medicina y afines [CUB22].
- Marco normativo para la aprobación y adopción de soluciones basadas en inteligencia artificial en la práctica clínica [BDM20, FDA AI Software as Medical Device Action Plan 2021[4]], así como la imputación de responsabilidades.
- Aspectos éticos para el uso de la inteligencia artificial en salud. La Organización Mundial de la Salud ha esbozado en 2021 seis principios básicos que deben cumplir las herramientas de inteligencia artificial en salud[5], a saber:
    1. Proteger la autonomía humana (lo que significa que los seres humanos deben conservar el pleno control de los sistemas sanitarios y las decisiones médicas);
    2. Promover el bienestar y la seguridad de las personas y el interés público; 3) Garantizar la transparencia, la claridad y la explicabilidad;
    3. Fomentar la responsabilidad e imputación de responsabilidad (rendición de cuentas);
    4. Garantizar la inclusión y la equidad;
    5. Promover herramientas de inteligencia artificial que sean receptivas y sostenibles.

[4] https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device

[5] https://www.who.int/publications/i/item/9789240029200

La inteligencia artificial es un campo que evoluciona a una velocidad tremenda, por lo que abordar estos desafíos es fundamental para poder aprovechar y beneficiarse de las muchas promesas de la inteligencia artificial en el área de la salud. Para lograrlo, será fundamental el esfuerzo y trabajo interdisciplinario de todas las partes involucradas, incluyendo médicos, pacientes, gestores de la salud, investigadores, industria y reguladores.

4.2 Aplicaciones de inteligencia artificial en Chile

En Chile, las implementaciones de sistemas de inteligencia artificial han ido avanzando e integrándose en diversas áreas de la sociedad. Esto refleja el desarrollo tecnológico del país y el avance en el uso de la IA como herramienta de apoyo.

Algunos ejemplos de estas aplicaciones en nuestro país incluyen un sistema de identificación de transacciones anómalas e irregularidades implementado por el Servicio de Impuestos Internos (SII), y una aplicación con IA para sistematizar audiencias de control de detención, implementada por la Defensoría Penal Pública del Estado.

El SII debe fiscalizar el cumplimiento de las normas tributarias y, para ello, dentro de sus procesos se incluye la identificación de anomalías en el comportamiento de empresas o personas naturales que se asocian a ciertas acciones ilegales, e.g., el aprovechamiento de instrumentos como las facturas para disminuir el pago de impuestos. Esta tarea se realizaba de forma manual, significando cuantioso tiempo invertido en cada estudio de fiscalización. El sistema de IA, apoya en esta gestión ya que automatiza la identificación de posibles comportamientos anómalos para agilizar la detección de fraudes, potenciando la efectividad en la detección de posibles fraudes al fisco por medio de transacciones irregulares.

El sistema usa datos extraídos de la plataforma del SII sobre las transacciones de cada contribuyente y el comportamiento de estos (por ejemplo, fechas en las que realiza el pago de impuestos) que presentan. Los resultados que genera son la identificación de posibles contribuyentes (de acuerdo a un porcentaje de probabilidad a partir de los datos analizados) que se encuentren realizando acciones irregulares a nivel tributario, los cuales son informados a las entidades fiscalizadoras correspondientes.

Por otra parte, la Defensoría Penal Pública, previo a las audiencias de control de detención, requiere que abogados y jueces realicen una investigación para la obtención y análisis de los datos históricos del imputado, como lo son detenciones u otros delitos cometidos. Esto termina aumentando la duración de cada audiencia significando una lentitud en las demás audiencias o procesos pendientes, lo que hace más lento el sistema completo. Por esto, se crea esta plataforma accesible vía web para categorizar el tipo de delito y obtener información del imputado. El objetivo es agilizar la obtención de información penal y judicial para aportar a la defensa de imputados.

La aplicación se basa en algoritmos predictivos y se usan datos disponibles dentro del sistema judicial del país. A partir de esto se genera una visualización de los antecedentes del imputado

o acusado a consultar, analizando también la información comparándola con delitos asociados al cual se le acusa (por ejemplo, si se tiene una acusación de microtráfico, analizar posibles delitos asociados como lo son la violencia intrafamiliar).

5. Aspectos éticos de la inteligencia artificial

La disciplina de la ética de la inteligencia artificial, a grandes rasgos, se preocupa del comportamiento moral de los seres humanos a la hora de diseñar, fabricar, utilizar y tratar los sistemas artificialmente inteligentes, así como las consecuencias de implementar estos sistemas. En los últimos siete años, la ética se encuentra en el centro de las preocupaciones académicas y sociales de la IA, contribuyendo a la creación de principios éticos y líneas de investigación que se hacen cargo fundamentalmente de los riesgos y el impacto socio tecnológico de esta tecnología.

5.1 Breve introducción a la ética de la IA

A diferencia del desarrollo mismo de esta tecnología, menciones sobre ética de IA comienzan a florecer en el ámbito de la investigación recién en el año 2016. Anteriormente, las menciones sobre ética en trabajos de investigación sobre IA eran esporádicas. Un grupo de investigadores sistematizó este fenómeno realizando un recuento de citas de Google Scholar para identificar las tendencias históricas [BO21]. Los resultados muestran que es desde el 2016 en adelante que los problemas éticos de la IA se vuelven parte central y decisiva del desarrollo de la disciplina.

Hay diferentes factores atribuibles a este cambio exponencial en el interés y formalización de las investigaciones sobre ética de IA. Desde la academia, uno de los hitos fundamentales fue el artículo de Floridi y Taddeo titulado: “What is data ethics?” (¿Qué es la ética de datos?) [FT16]. Allí, los autores expresan su ambición de presentar aspectos fundacionales para sentar las bases de la discusión de la ética de datos, como una nueva rama de estudio que busca evaluar problemas morales relacionados a amplias implementaciones, incluyendo la generación, el registro, la conservación, el procesamiento, la difusión, el intercambio y uso de algoritmos. Estas temáticas incluyen la inteligencia artificial, los agentes artificiales, el aprendizaje automático y los robots, además de las prácticas correspondientes como la innovación responsable, la programación, la piratería informática y los códigos profesionales, con el fin de formular y apoyar soluciones moralmente buenas (por ejemplo, conductas correctas o valores correctos a ser integrados en el desarrollo de la tecnología).

Paralelo a esto, desde el periodismo investigativo de ProPublica [AN16], se cuestiona el uso de un algoritmo de predicción llamado COMPAS en diferentes estados de EE.UU. Este algoritmo es utilizado para predecir el riesgo de reincidencia y, así, apoyar los procesos de juicio de

diferentes imputados. Este controversial caso se ha convertido en un referente para instalar problemas éticos en la implementación de algoritmos en ámbitos de alta connotación pública y con profundos impactos sociales, poniendo temas como la disparidad de raza y género, la discriminación sistemática, las definiciones de justicia y la influencia de los sesgos en la palestra. Asimismo, publicaciones de divulgación científica, como "Weapons of Math Destruction" (Armas de destrucción matemática) de Cathy O'Neil [O16], fueron instrumentales para la concientización generalizada sobre los efectos sociales de estas tecnologías, ejemplificando cómo afectan el día a día.

Entre 2017 y 2018, la discusión sobre ética de IA se extendió a la esfera pública, en gran parte debido a una serie de escandalosos casos que ejemplifican las faltas de integración de la ética en los procesos de desarrollo de las tecnologías. En el reporte del año 2018 del Instituto AI Now, liderado por la Universidad de Nueva York, se destacan hitos como los problemas relacionados a faltas de privacidad (e.g., brechas de seguridad y exposición de datos de Facebook), sesgos y discriminación (e.g., algoritmo que sugiere masivas deportaciones en Reino Unido), daño físico (e.g., accidentes de Tesla y Uber) y daño moral (e.g., Cambridge Analytica y su influencia en elecciones de Brexit en Reino Unido y presidenciales de EE.UU.). En paralelo a estas controversias, surgen las primeras iniciativas legales y normativas de directrices que guían el quehacer tecnológico, como la GDPR europea (Reglamento General de Protección de Datos), la ley de privacidad del consumidor en California, EE.UU., o las peticiones de Microsoft para regular el reconocimiento facial.

La alta demanda y la diversificación de los tipos de problemas éticos que surgían de la mano de la implementación de la IA en la sociedad, llevó a que surgieran sub-áreas relacionadas a la ética de la IA. Tal y como lo planteaban Floridi y Taddeo inicialmente, hay una serie de elementos que, si bien, son parte de una ética de datos o una ética de inteligencia artificial más general, se ocupan de ámbitos e implementaciones que requieren atención específica, ya sea por el tipo de problema que evocan o las ramificaciones que estos conllevan. Por esto es que al hablar de ética de la IA, se deben tener en cuenta diferentes sub-disciplinas que se han ido orgánicamente gestionando para hacer frente a las preocupaciones que el desarrollo de la IA trae consigo. Entre ellas se pueden identificar:

- Ética de datos: se ocupa de cómo se usan y manipulan los datos que alimentan sistemas de IA. Preocupaciones comunes: protección de privacidad, gobernanza, sesgos, y responsabilidad.
- Ética de máquinas: se ocupa de añadir o garantizar comportamientos morales a máquinas que usan IA. Preocupaciones comunes: estatus moral de máquinas e IA, creación de agentes morales artificiales, e.g., sistemas de armas, predicciones, y decisiones automatizadas.
- Ética y riesgo de singularidad: se ocupa de problemas futuros y de control como el desarrollo de una superinteligencia y el riesgo existencial que supone para la humanidad.

- Interacción humano robot (IHR): se ocupa de las responsabilidades asociadas al desarrollo de robots y su relación con humanos, además de problematizar la necesidad y posibilidad de derechos para los robots.
- Ética de algoritmos: se ocupa de cuestionar el proceso de desarrollo de modelos de IA y ML, para verificar una implementación ética. Preocupaciones comunes: transparencia, explicabilidad, sesgos, replicabilidad, interpretabilidad.

Lo que tienen en común todas estas sub-disciplinas o áreas de estudio, es que se basan en una serie de principios fundamentales que han permitido mayor estructuración en la integración de la ética en IA, a nivel metodológico y normativo.

5.2 Principios éticos y gobernanza en IA

Debido a esta creciente preocupación por el impacto social y las consecuencias del desarrollo e implementación de la IA en distintos ámbitos de la sociedad, se han propuesto variados principios para guiar las prácticas profesionales, institucionales, y gubernamentales que sustentan la IA. Estos principios pueden encontrarse en diversos documentos, como códigos de conducta, declaraciones, lineamientos, o reportes, que son a su vez elaborados por diversas instituciones, incluyendo empresas privadas, gobiernos, organizaciones gubernamentales y legislativas, informes académicos, ONGs, etc. A grandes rasgos, la mayoría de estos documentos tienen como finalidad básica y fundacional el presentar una visión sobre una gobernanza responsable de la IA, con distintos niveles de profundidad y alcance, apuntando así a variadas audiencias.

Se destacan aquí 4 principios éticos principales en base a la sistematización de cuatro influyentes artículos de investigación [JIV19, FJ20, H20, KH22] quienes han analizado comparativamente los diferentes principios propuestos. Esto incluye, además de reportes académicos, directrices usadas a nivel global como las de la OCDE, la UE, estrategias nacionales como las de Reino Unido, Chile, México, Alemania, Singapur, entre otros, y propuestas institucionales como las de Microsoft, IBM, Google, y Tesla.

1. Transparencia: Este es el principio más recurrente. Transparencia se entiende en términos generales desde una respuesta a uno de los mayores desafíos que plantea la IA desde el punto de vista de gobernanza y funcionalidad: la opacidad. Es difícil entender y saber a nivel técnico qué ocurre en cada proceso del funcionamiento de diferentes modelos algorítmicos, lo que dificulta la capacidad de explicar y transparentar qué se está haciendo con los datos de los involucrados y cómo se está llevando a cabo. Esto hace que el principio de transparencia se relacione también con preocupaciones sobre código abierto (open source), y el derecho a la información (y explicación). La transparencia, como principio, responde a visibilizar qué ocurre con el sistema y las

decisiones asociadas al proceso, desencadenando demandas sobre interpretabilidad y explicabilidad, para poder generar sistemas de IA confiables y fiables, mejorando las prácticas detrás del desarrollo de estos.

2. Justicia o no-discriminación (*fairness*): Este es el segundo principio más recurrente, pero usualmente el más representado (ya que está presente en todos los reportes o declaraciones consideradas), y se basa fundamentalmente en la prevención de sesgos algorítmicos. Hay muchas maneras en las cuales diferentes sesgos pueden entrar en los sistemas de IA, influyendo en entrenamientos no representativos, resultados alterados por proxies imperfectos que alteran predicciones, o inclusive por sesgos cognitivos que traen los humanos en el bucle (*humans in the loop*). Este principio llama a prácticas representativas, como el uso de datos de alta calidad y diversidad, mantener equidad, establecer inclusividad en el impacto social y en el diseño, así como también en los equipos de trabajo que los desarrollan. Esto hace que se busque evitar la discriminación arbitraria realizada por sistemas de decisión automatizada o por quienes diseñan sistemas de clasificación o predicción, considerando el impacto que tiene en temas de dignidad, autonomía, y derechos humanos.

3. Privacidad: La privacidad se presenta como un valor a ser integrado en el desarrollo de la IA, y como un derecho que debe ser respetado. Por lo tanto, este principio se relaciona a prácticas de protección de datos y medidas de seguridad. La privacidad, sin embargo, también concierne aspectos relacionados a vigilancia, publicidad y manipulación, así como la toma de decisiones a gran escala (como por ejemplo durante la pasada pandemia del COVID-19). Este principio, a diferencia de otros, se vale de la institucionalidad que tiene a nivel legislativo en varios países, esclareciendo su rol como protector de un derecho humano básico. Así con este principio, se busca garantizar la privacidad de datos a través del ciclo de desarrollo de la IA, así como también proveer el derecho de control de la información a los ciudadanos.

4. Responsabilidad (*responsibility and accountability*): El concepto de responsabilidad se relaciona con aspectos legales, morales, personales, colectivos, y ecosistémicos. Por una parte, la responsabilidad como principio llama a asegurar la prevención de riesgos y daños causados tanto a individuos, como a ecosistemas (relacionandolo con el principio de sustentabilidad). Esto implica minimizar acciones y decisiones que no tengan asociadas responsabilidades directas, tanto en procesos de desarrollo como en implementación y decisiones operacionales. Esto facilita procesos de auditoría y rendición de cuentas. Por otra parte, se considera la responsabilidad moral en un sentido de integridad. Para esto la responsabilidad actúa también como un “meta-principio”, al buscar garantizar que se promuevan buenas prácticas, se transparenten las decisiones y acciones tomadas, y se hagan evaluaciones de riesgo e impacto que aseguren una implementación ética de los proyectos más allá de los requisitos legales básicos.

6. Reflexiones finales

Como mencionamos anteriormente, el desarrollo de aplicaciones como Chat GPT está basado en el desarrollo de las redes neuronales profundas. Estas redes están formadas por una gran cantidad de neuronas artificiales, que son unidades sencillas de cálculo, agrupadas en capas y unidas a través de conexiones ponderadas. Como tal, las redes neuronales son un modelo de computación, cuya capacidad de cálculo está acotada por la capacidad de la máquina de Turing, que es hasta el día de hoy nuestra mejor formalización de la idea de algoritmo o proceso mecánico. De hecho, las redes neuronales usuales deben agregar mecanismos adicionales, como el ser recurrentes (uso de datos secuenciales) para lograr el mismo poder de computación que las máquinas de Turing [SS92]. En este sentido, es importante dimensionar las limitaciones y los beneficios que esta tecnología tiene. Por un lado, hay problemas computacionales que las redes neuronales, y la inteligencia artificial en general, no van a ser capaces de solucionar. Y, por otro lado, resulta fundamental entender la capacidad de estas tecnologías, para poder identificar qué tipo de problemas pueden ser naturalmente solucionados con el uso de redes neuronales.

Desde una perspectiva técnica, una de las conclusiones que ha emergido del uso de Chat GPT es que el procesamiento del lenguaje natural por los humanos presenta regularidades que pueden ser identificadas por las redes neuronales profundas [W23]. Tales regularidades pueden estar presentes en otro tipo de problemas, aunque podrían necesitar de capacidades de computación distintas a las que pueden ser modeladas de manera natural en una red neuronal. Es por esto que hay un gran interés en la integración de modelos de lenguaje natural, como Chat GPT, con lenguajes formales usados para el manejo y análisis de datos, como Wolfram Language [W23] y SQL.[6] Esto refleja un camino abierto de posibilidades que puede diversificar ampliamente el uso de esta interacción a otros lenguajes, intensificando el impacto de la tecnología en diversas disciplinas.

El acelerado avance de la IA en los últimos cinco años ha revolucionado el quehacer científico y también las diferentes profesiones. Es por esto que resulta no sólo necesario, sino que fundamental, incluir el conocimiento sobre este tipo de tecnologías en el aula universitaria. Este conocimiento, sin embargo, no debe limitarse a la mera aplicación o uso de diferentes herramientas de IA, ya sea por parte de los docentes como por el mismo estudiantado. La inclusión de la IA a las aulas debe integrar una visión crítica y reflexiva, que permita dimensionar las limitaciones y posibles riesgos que la tecnología conlleva.

Las aprensiones que los cambios tecnológicos conllevan son esperables, han sido parte histórica de los grandes avances en ciencia y tecnología. Estos cambios forman parte de procesos que tienen sus propias dinámicas, impulsados por la creatividad humana y el enorme ingenio que se produce a partir de la colaboración entre personas. La historia nos muestra que

[6] https://blog.langchain.dev/llms-and-sql

estas transformaciones han producido beneficiosos efectos en la sociedad y, a la vez, han planteado nuevos desafíos.

Así, a pesar de que muchas veces la tecnología ha sido usada con fines distintos para los cuales fueron diseñados, esto no ha logrado detener el avance de la ciencia y la tecnología. Las reflexiones y aprensiones sobre el desarrollo tecnológico deben guiar y enfocar las discusiones normativas, pero en ningún caso tener como objetivo detener el avance tecnológico. En esta línea, uno de los enfoques principales que sustentan el debate sobre los avances de la IA es la perspectiva sociotécnica, que llama a impulsar el avance desde la innovación basada en principios éticos sólidos, considerando contextos de aplicación, promoviendo la responsabilidad y la preocupación por el impacto que estos cambios tecnológicos tienen en la sociedad actual y sus futuras generaciones.

Creemos, por lo tanto, que es parte de nuestra tarea como comunidad universitaria el incorporar estas tecnologías, entenderlas, mejorarlas y plantear las preguntas y cuestionamientos necesarios para abordar nuestras inquietudes, sin intentar bloquear el cambio a través de la prohibición de su uso o la limitación de sus alcances, sino más bien desde una mirada reflexiva y propositiva que incorpore en nuestro quehacer estas tecnologías para extraer desde ellas el máximo beneficio. Desde esta mirada crítica la colaboración interdisciplinaria resulta fundamental. Aspectos propios de las metodologías científicas y los análisis socio-técnicos de las humanidades y ciencias sociales, son elementos básicos para enfrentar este cambio de paradigma. Avanzar en la integración de la IA en la comunidad universitaria requiere tener una línea de acción y alfabetización clara, que promueva el diálogo interdisciplinario y colaborativo, y abra el espacio epistémico crítico.

Referencias

[T37] Turing, A. M. On computable numbers, with an application to the Entscheidungsproblem. Proc. London Math. Soc. s2-42(1): 230–265, 1937.

[MP43] McCulloch, W. S., and Pitts, W.. A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics, 5, 115-133, 1943.

[R58] Rosenblatt, F. The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review, 65(6), 386, 1958.

[LBH15] LeCun, Y., Bengio, Y., and Hinton, G. E. Deep learning. Nature 521(7553): 436–444, 2015.

[KSE12] Krizhevsky, A., Sutskever, I., and Hinton G. E. ImageNet Classification with Deep Convolutional Neural Networks. NIPS 2012: 1106-1114.

[GBC16] Goodfellow, I., Bengio, Y., and Courville, A. Deep learning. MIT press, 2016.

[RHW86] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. Learning representations by back-propagating errors. Nature, 323(6088): 533–536, 1986.

[MK13] Mikolov, T.; Sutskever, I., Chen, K., Corrado, G., Dean, J. Distributed Representations of Words and Phrases and their Compositionality. NIPS 2013: 3111-3119

[Bo17] Bojanowski, P., Grave, E., Joulin, A., Mikolov, T. Enriching Word Vectors with Subword Information. Trans. Assoc. Comput. Linguistics 5: 135-146 (2017)

[PE18] Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L. Deep Contextualized Word Representations. NAACL-HLT 2018: 2227-2237

[VA17] Vaswani,A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., Polosukhin, I. Attention is All you Need. NIPS 2017: 5998-6008

[DE19] Devlin, J., Chang, M., Lee, K., Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT (1) 2019: 4171-4186

[ST20] Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., Christiano, P. Learning to summarize with human feedback. NeurIPS 2020

[DDF+23] Dicuonzo, G., Donofrio,F., Fusco, A., Shini, M. Healthcare system: Moving forward with artificial intelligence, Technovation, 120:102510, 2023.

[AE23] Apell, P., and Eriksson, H. Artificial intelligence (AI) healthcare technology innovations: the current state and challenges from a life science industry perspective, Technology Analysis & Strategic Management, 35:2, 179-193, 2023 DOI: 10.1080/09537325.2021.1971188.

[LFK+19] Liu, X., Faes, L., Kale, AU., Wagner, SK., Fu, DJ., Bruynseels, A. et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis. Lancet Digital Health. 1:6, 2019

[DK19] Davenport, T., and Kalakota, R. The potential for artificial intelligence in healthcare. Future Healthc J. 6(2):94-98, 2019. doi: 10.7861/futurehosp.6-2-94. PMID: 31363513; PMCID: PMC6616181.

[RHL20] Rieke, N., Hancox, J., Li, W. et al. The future of digital health with federated learning. npj Digit. Med. 3, 119, 2020. https://doi.org/10.1038/s41746-020-00323-1

[BDM20] Benjamens, S., Dhunnoo, P. & Meskó, B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. npj Digit. Med. 3, 118, 2020. https://doi.org/10.1038/s41746-020-00324-0

[CUB22] Civaner, M.M., Uncu, Y., Bulut, F. et al. Artificial intelligence in medical education: a cross-sectional needs assessment. BMC Med Educ 22, 772, 2022. https://doi.org/10.1186/s12909-022-03852-3

WHO guidance on Ethics & Governance of Artificial Intelligence for Health, 2021. ISBN: 9789240029200.

[A18] AI Now Institute. AI Now 2018 Symposium. https://www.youtube.com/watch?v=NmdAtfcmTNg

[AN16] Angwin, J., Larson, J., Mattu, S., & Kirchner, L. Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[BO22] Borenstein, J., Grodzinsky, F. S., Howard, A., Miller, K. W., & Wolf, M. J. AI Ethics: A Long History and a Recent Burst of Attention. Computer, 54(01), 96–102, 2021. https://doi.org/10.1109/MC.2020.3034950

[FJ20] Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI, 2020. (SSRN Scholarly Paper No. 3518482). Social Science Research Network. https://doi.org/10.2139/ssrn.3518482

[FT16] Floridi Luciano and Taddeo Mariarosaria. What is data ethics? Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2083), 2016. https://doi.org/10.1098/rsta.2016.0360

[H20] Hagendorff, T. The Ethics of AI Ethics: An Evaluation of Guidelines. Minds and Machines, 30(1), 99–120, 2020. https://doi.org/10.1007/s11023-020-09517-8

[JIV19] Jobin, A., Ienca, M., & Vayena, E. The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), Article 9, 2019. https://doi.org/10.1038/s42256-019-0088-2

[KH22] Khan, A. A., Badshah, S., Liang, P., Waseem, M., Khan, B., Ahmad, A., Fahmideh, M., Niazi, M., & Akbar, M. A. Ethics of AI: A Systematic Literature Review of Principles and Challenges. Proceedings of the International Conference on Evaluation and Assessment in Software Engineering 2022, 383–39, 2022. https://doi.org/10.1145/3530019.3531329

[O16] O’Neil, C. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown Publishing Group, 2016.

[W23] Wolfram, S. What Is ChatGPT Doing ... and Why Does It Work? Wolfram Research, Inc., 2023.

[SS92] Siegelmann, H. T., and Sontag, E. D. On the computational power of neural nets. In Proceedings of the fifth annual workshop on Computational learning theory, pages 440–449, 1992.