IA segura mediante IA TRiSM:

Reducción de ciberataques y brechas de seguridad

Alumno: Natale, Ariel Alejandro

Tutoría Técnica: Ing. Romero, Raúl Oscar

Profesores de Trabajo Final:

Dra. Samela, Marcela

Mg. Ugarte, Carlos Alberto

Trabajo Final de Carrera presentado para obtener el título de

Licenciatura en Gestión de Tecnología Informática

Diciembre 2023

Resumen

Resulta evidente el aumento de la influencia de la Inteligencia Artificial en nuestra vida cotidiana, tanto en el ámbito empresarial como en el personal.

La seguridad en los modelos de Inteligencia Artificial se vuelve fundamental debido a la creciente amenaza de ciberataques que pueden afectar las tomas de decisiones.

El objetivo principal de este estudio se centra en evaluar la efectividad de técnicas de defensa en los sistemas de IA para mitigar la amenaza de ciberataques. Para ello se examinó el marco de trabajo AI TRiSM: AI Trust, Risk, and Security Management, y como resultado de este trabajo se brinda una guía de buenas prácticas para su aplicación.

En esta investigación se enfatizó la importancia de gestionar la confianza, el riesgo y la seguridad en la IA para abordar de manera proactiva cuestiones de gobernanza y cumplimiento. Además, se subraya la necesidad de brindar explicaciones claras sobre el funcionamiento de la IA y establecer controles de acceso para las herramientas de generación de IA.

A través del análisis de casos de uso empresariales se pudo evidenciar que AI TRiSM ofrece una solución integral para abordar riesgos relacionados con la confidencialidad de los datos, mantener un monitoreo constante y protegerse contra ataques, promoviendo sistemas de IA confiables, equitativos y seguros en un entorno de regulaciones en constante evolución.

Palabras clave: aprendizaje automático, ciberataques, inteligencia artificial, privacidad, seguridad

Abstract

It is evident the increasing influence of Artificial Intelligence in our daily lives, both in the business and personal realms.

Security in Artificial Intelligence models becomes crucial due to the growing threat of cyberattacks that can impact decision-making processes.

The primary objective of this study is centered on assessing the effectiveness of defense techniques within AI systems to mitigate the cyberattack threat. To achieve this, the AI TRiSM framework, which stands for AI Trust, Risk, and Security Management, was scrutinized. As a result of the investigation, a best practices guide is provided for its implementation.

This research emphasizes the importance of managing trust, risk, and security in AI to proactively address governance and compliance issues. Furthermore, it underscores the need for clear explanations about the functioning of AI and the establishment of access controls for AI generation tools.

Through the examination of business use cases, it was evident that AI TRiSM offers a comprehensive solution for addressing risks associated with data confidentiality, maintaining ongoing monitoring, and safeguarding against attacks. This promotes the development of dependable, equitable, and secure AI systems in a continually evolving regulatory environment.

Keywords: artificial intelligence, cyberattacks, machine learning, privacy, security

Dedicatoria

A mis amados abuelos Berta y Nicolás, quienes han sido pilares fundamentales en mi vida. Desde mi infancia, me han criado con amor, cuidado y sabiduría. Han sido mi guía y ejemplo, transmitiéndome los valores más importantes de la vida.

A mi mamá Patricia y mi hermana Milagros que, desde mis primeros pasos hasta este logro académico, han estado a mi lado, apoyándome incondicionalmente y brindándome su amor inquebrantable.

A mi querido tío Víctor, aunque desearía que estuviera físicamente aquí para presenciar este momento tan importante, sé que su espíritu está conmigo. Su recuerdo siempre vivirá en mi corazón y en cada logro que alcance.

A Gaspar, por ser el motor que me impulsó a retomar y finalizar esta carrera. Quiero expresar mi profundo agradecimiento por su constante apoyo y por creer en mí incluso cuando yo dudaba de mis propias habilidades.

A Segovia y a la memoria de Guille, mis primeros jefes. Con su confianza y motivación me impulsaron a esforzarme y a dar lo mejor de mí en cada paso de mi camino profesional.

Reconocimientos

A la Dra. Marcela Samela y Mr. Carlos Ugarte agradezco infinitamente por su constante motivación durante toda la cursada y en el proceso de redacción del Trabajo Final de Carrera. Clase a clase me enseñaron que con esfuerzo y perseverancia se puede lograr el objetivo.

Especial agradecimiento a mi tutor, Ing. Oscar Romero, quien ha brindado su tiempo y conocimientos de manera desinteresada para asegurarse de mi éxito durante todo nuestro tiempo juntos en mi carrera universitaria. Su orientación y aliento han sido inmensamente valiosos, sirviendo como una fuente constante de inspiración y motivación.

A los docentes de la Universidad Abierta Interamericana, así como a mis compañeros de la universidad y del trabajo, que me acompañaron durante todo este largo recorrido. A cada uno de ellos, les extiendo mi más sincero agradecimiento de todo corazón.

Índice General

Índice de Figuras

Índice de Tablas

Capítulo 1 – Introducción

1.1 Justificación de la Investigación

La presencia de la Inteligencia Artificial (IA) es cada vez más notable tanto en nuestra vida diaria como en el ámbito laboral. Las organizaciones han incorporado cientos o incluso miles de modelos de IA, lo que ha generado inquietudes entre los responsables de Tecnología de la Información, quienes a menudo se enfrentan a desafíos al explicar o interpretar su funcionamiento. La falta de conocimiento y comprensión puede tener consecuencias significativas, ya que un rendimiento deficiente de los modelos de IA puede resultar en vulnerabilidades de la privacidad, ocasionar pérdidas económicas, dañar la reputación e incluso conllevar riesgos personales. Por lo tanto, es fundamental que las organizaciones tomen medidas preventivas para garantizar la seguridad y eficacia de la IA en su uso (Perri, 2022).

En este estudio, nos enfocamos en el desafío de asegurar la confiabilidad de la IA. Las amenazas de ciberataques son cuestiones que demandan atención, ya que pueden erosionar la confianza de los usuarios, especialmente cuando se trata de información sensible.

Por ende, es imperativo abordar este desafío para garantizar que la IA sea confiable y bien recibida en la sociedad. Este problema es de relevancia global, ya que la IA se utiliza cada vez más en diferentes países y contextos.

Las consecuencias de un uso incorrecto o de fallos en la seguridad de la IA pueden ser devastadoras y socavar la confianza de las personas en estos sistemas. En consecuencia, se hace necesario encontrar una solución sólida y efectiva para abordar este desafío.

En el marco de esta investigación se llevará a cabo un estudio cualitativo con el propósito de analizar las distintas perspectivas y enfoques existentes en relación con las medidas necesarias para garantizar la integridad en el campo de la Inteligencia Artificial.

1.2 Antecedentes

En este campo de investigación, se nota la ausencia de estudios exhaustivos que se centren específicamente en la gestión rigurosa de la confianza, el riesgo y la seguridad en la Inteligencia Artificial.

A pesar de ello, hemos identificado un artículo relevante que servirá como punto de partida para nuestra investigación. El artículo, titulado "What It Takes to Make AI Safe and Effective" y escrito por Lori Perri (2022), examina detalladamente cómo las organizaciones que aplican el enfoque de IA TRiSM pueden lograr una implementación exitosa de estos modelos. En el artículo, la autora destaca la necesidad de adoptar un enfoque riguroso en la gestión de la confianza, el riesgo y la seguridad en los sistemas inteligentes. Este enfoque implica el uso de

soluciones, técnicas y procesos que permiten interpretar y explicar los modelos, así como la protección de la privacidad, las operaciones de los modelos y la resistencia a ataques adversarios.

A pesar de los avances significativos en el ámbito de la IA, todavía no se comprende completamente en qué medida la implementación de políticas de seguridad y privacidad en los modelos de IA pueden reducir significativamente el riesgo de ciberataques y vulnerabilidades. Esto se debe, en parte, al constante desarrollo del campo de la IA, lo que dificulta la comprensión total de sus implicancias en términos de defensa.

1.3 Hipótesis

La adopción de la guía y sugerencias presentadas en este estudio de investigación, orientadas a fortalecer la protección y confidencialidad en sistemas de Inteligencia Artificial, resultará en una considerable disminución del riesgo asociado a ataques cibernéticos y vulnerabilidades.

1.4 Objetivos

El objetivo principal de la presente investigación se enfoca en evaluar de qué manera la implementación de medidas apropiadas de protección y privacidad en los sistemas de IA puede reducir el riesgo de ciberataques y posibles vulnerabilidades.

Con el fin de alcanzar esta meta, se establecen los siguientes objetivos específicos:

- Identificar los principales riesgos y desafíos asociados a los sistemas de Inteligencia Artificial.
- Analizar el marco de trabajo AI TRiSM: AI Trust, Risk, and Security Management para la implementación de barreras de seguridad y privacidad en sistemas de Inteligencia Artificial.
- Proporcionar una guía para la implementación y buenas prácticas para mejorar la protección y confidencialidad en sistemas de IA a través del uso del marco de trabajo AI TRiSM: AI Trust, Risk, and Security Management.

1.5 Enfoque Metodológico

En este estudio se emplea un enfoque cualitativo que se centra en la investigación y el análisis de fuentes bibliográficas relacionadas con la protección de modelos de IA, especialmente en el contexto del marco de trabajo AI TRiSM. Se ha recopilado información pertinente de literatura académica mediante el examen de casos de uso y materiales bibliográficos. Se han explorado beneficios, riesgos, amenazas y enfoques de confianza y seguridad en profundidad.

Este enfoque metodológico ha permitido obtener una visión completa sobre la seguridad de los modelos de IA y ha servido como base para desarrollar recomendaciones prácticas. A partir del análisis bibliográfico, se ha elaborado un conjunto de buenas prácticas específicas para la implementación del marco AI TRiSM: AI Trust, Risk, and Security Management.

1.6 Estructura General del Trabajo Final de Carrera

El trabajo consta de cuatro capítulos que abordan los aspectos clave de la investigación. Finalmente se incluyen las conclusiones, las direcciones para futuras investigaciones, los acrónimos y las referencias bibliográficas.

Capítulo 1 – Introducción: este capítulo justifica la realización de este Trabajo Final de Carrera, resaltando la importancia de analizar el problema en cuestión. Además, se presenta la hipótesis y los objetivos, tanto generales como específicos.

Capítulo 2 – Marco Teórico: en este capítulo se exploran los antecedentes históricos y teóricos que respaldan el análisis del trabajo. Se documenta cómo esta investigación contribuye al conocimiento existente.

Capítulo 3 – Desarrollo: este capítulo se centra en la implementación de medidas de seguridad y recomendaciones en el contexto de la IA utilizando el marco de trabajo IA TRiSM.

Capítulo 4 – Conclusiones y Futuras Líneas de Investigación: en este último capítulo, se presentan las conclusiones alcanzadas y se plantean posibles áreas para futuras investigaciones.

Capítulo 2 - Marco Teórico

2.1 Inteligencia Artificial

En el campo de la Inteligencia Artificial se han investigado diferentes enfoques, algunos orientados a emular el rendimiento humano, mientras que otros se centran en una definición abstracta de racionalidad, la cual implica la toma de decisiones coherentes con criterios correctos. El concepto de racionalidad se divide en dos aspectos, algunos lo consideran como una propiedad de los procesos internos de pensamiento y razonamiento, mientras que otros se enfocan en el comportamiento inteligente, es decir, su caracterización externa. A partir de estas dos dimensiones, humano vs. racional y pensamiento vs. comportamiento, existen cuatro posibles combinaciones que han sido objeto de investigaciones (Russell & Norvig, pág. 1).

El método de evaluación conocido como la prueba de Turing, llamado "Actuando como humanos", fue desarrollado por Alan Turing en 1950 para abordar la incertidumbre filosófica de la pregunta "¿Puede una máquina pensar?". Si una máquina logra superar este examen, un interrogador humano no puede distinguir si las respuestas que recibe provienen de una persona o de una computadora. En cuanto a las capacidades necesarias para que una computadora pase la prueba, se destacan el procesamiento del lenguaje natural para comunicarse en un lenguaje humano, la representación del conocimiento, el razonamiento automatizado y Machine Learning. A pesar de que Turing consideró innecesaria la simulación física de una persona para demostrar inteligencia, otros investigadores han propuesto una versión más completa de la prueba que implica interacción con el mundo real. Para superar la prueba de Turing total, un robot deberá contar con habilidades como la visión por computadora y el reconocimiento de voz para percibir el mundo, así como la capacidad de manipular objetos y desplazarse. Aunque estas seis disciplinas son esenciales en la mayoría de las investigaciones sobre IA, muchos investigadores no han dedicado suficiente atención a la prueba de Turing, dado que consideran prioritario investigar los principios fundamentales de la inteligencia (Russell & Norvig, 2020, pág. 2).

Russell y Norvig (2020), mencionan que el enfoque del modelo cognitivo, pensando de manera humana, busca comprender el pensamiento humano, y para lograrlo, emplea tres métodos: la introspección, los experimentos psicológicos y las imágenes cerebrales. Una vez que se ha desarrollado una teoría precisa de la mente humana, se puede expresar como un programa de computadora y evaluar si los mecanismos del programa podrían estar operando en los humanos. La ciencia cognitiva es un campo interdisciplinario que se basa en la investigación experimental de humanos o animales reales. Aunque se comentará ocasionalmente sobre las similitudes o diferencias entre las técnicas de IA y la cognición humana, la verdadera ciencia cognitiva se enfoca en la investigación experimental. En los primeros días de la IA, había

confusión entre los enfoques, pero actualmente se separan los dos tipos de afirmaciones para permitir el desarrollo de la IA y la ciencia cognitiva. Los dos campos se complementan, sobre todo en la visión por computadora, que incorpora evidencia neurofisiológica en modelos computacionales. Gracias a la combinación de métodos de neuroimagen y técnicas de Machine Learning, se está desarrollando la capacidad de leer mentes y determinar el contenido semántico de los pensamientos internos de una persona. Esta habilidad podría ayudar a identificar sobre cómo funciona la cognición humana (pág. 2).

En relación con el pensamiento racional, podemos enfocarnos en las Leyes del Pensamiento, las cuales buscan establecer procesos de razonamiento que sean irrefutables. El pensador Aristóteles fue uno de los primeros en tratar de codificar este tipo de pensamiento, lo que condujo al estudio de la lógica. Los expertos en lógica del siglo XIX desarrollaron una notación precisa para declaraciones acerca de objetos en el mundo y las conexiones entre ellos, lo que permitió la resolución de problemas mediante software. Aunque la lógica usualmente requiere un conocimiento seguro del mundo, lo cual es raramente alcanzado, la teoría de la probabilidad permite un razonamiento riguroso con información incierta, construyendo un modelo integral de pensamiento racional, que abarca desde la información perceptual bruta hasta predicciones acerca del futuro. No obstante, esto no es suficiente para generar conducta inteligente, lo cual requiere una teoría de la acción racional. Este enfoque implica que un agente logre el mejor resultado o, en situaciones inciertas, el mejor resultado esperado. Se basa en que los agentes deben ser capaces de operar autónomamente, adaptarse al cambio, tomar decisiones y perseguir objetivos. Este enfoque no se limita a inferencias correctas y se enfoca en lograr la racionalidad de manera general y matemáticamente bien definida. En cambio, se reconoce que la racionalidad perfecta no siempre es factible en entornos complejos, por lo que se hace necesario un refinamiento para considerar la racionalidad limitada (Russell & Norvig, pág. 3).

Gartner (2021) sostiene lo expuesto en párrafos anteriores, definiendo a la Inteligencia Artificial como una disciplina de ingeniería informática que se basa en técnicas matemáticas o lógicas para descubrir, capturar y codificar conocimientos, aprovechando mecanismos sofisticados e inteligentes para resolver problemas, haciendo una simulación de procesos cognitivos mediante programas informáticos con el fin de interpretar eventos, automatizar decisiones, y llevar a cabo acciones. Se reconoce que la tecnología de IA implica análisis probabilísticos que combinan la probabilidad y la lógica para asignar un valor a la incertidumbre, y esta definición se ajusta a la actual y emergente situación de las tecnologías y capacidades de IA. En un contexto empresarial, la IA puede aplicarse desde automatización en una plantilla de empleados hasta la detección de actividades fraudulentas, la identificación de oportunidades de

venta cruzada, haciendo promoción de productos o servicios complementarios a los ya adquiridos por el cliente, la optimización de recursos, e incluso robots inteligentes que realizan tareas en el lugar de trabajo.

Para concluir, Russell y Norvig (2020) señalan que se han producido importantes avances en varios campos. Por ejemplo, se han desarrollado sistemas expertos que capturan el conocimiento humano y lo aplican para solucionar problemas del mundo real. Además, se han desarrollado técnicas de razonamiento probabilístico que abordan la incertidumbre de manera sólida y se ha creado el Deep Learning, que es un componente vital de la informática contemporánea.

2.1.1 Inicio de la Inteligencia Artificial

La génesis de la Inteligencia Artificial se remonta a los años 1943-1956, cuando Warren McCulloch y Walter Pitts presentaron un modelo de neuronas artificiales, evidenciando la capacidad de una red neuronal para calcular cualquier función computable (Russell & Norvig, 2020, pág. 17).

En 1950, Marvin Minsky y Dean Edmonds construyeron la primera computadora de red neuronal y posteriormente Minsky estudió la computación universal en redes neuronales en Princeton. Alan Turing también tuvo una gran influencia en la IA con su visión sobre el aprendizaje automático, los algoritmos genéticos, el aprendizaje por refuerzo y la prueba de Turing. En 1956, se llevó a cabo un taller de dos meses en Dartmouth College que reunió a investigadores interesados en la IA, pero no se logró ningún avance significativo en ese momento. No obstante, Newell y Simon presentaron el trabajo más maduro con un sistema de demostración de teoremas llamado Logic Theorist (Russell & Norvig, 2020, pág. 17).

2.1.2 Entusiasmo temprano y grandes expectativas

Durante los años 50, la mayoría de la comunidad intelectual sostenía la opinión de que las máquinas eran incapaces de realizar algunas tareas. Sin embargo, los investigadores de Inteligencia Artificial desafiaron esta creencia, presentando pruebas con juegos, rompecabezas, matemáticas y pruebas de CI, que se consideraban indicativas de la inteligencia humana. Un ejemplo destacado de esta aproximación, pensar humanamente, fue el programa GPS creado por Newell y Simon, que buscaba imitar los protocolos de resolución de problemas humanos. Además, el equipo de Rochester en IBM produjo algunos de los primeros programas de IA, incluyendo el Geometry Theorem Prover de Gelernter (Russell & Norvig, pág. 18).

En 1956, Samuel demostró que las computadoras podían aprender a realizar tareas que no se les habían ordenado específicamente, a través de su programa de aprendizaje por refuerzo en televisión. En 1958, McCarthy propuso un sistema de IA basado en conocimiento y razonamiento, y definió el lenguaje de programación de alto nivel Lisp en el Instituto Tecnológico de Massachusetts, Minsky supervisó la creación de micromundos, que eran dominios limitados que requerían inteligencia para resolver problemas. Estos entornos virtuales incluyeron programas que se especializaban en la resolución de problemas de integración de cálculo, analogías geométricas y problemas de álgebra, respectivamente (Russell & Norvig, 2020, pág. 18).

2.1.3 Una dosis de realidad

A medida que avanzaba la investigación en la Inteligencia Artificial, surgieron nuevas perspectivas y enfoques para abordar los desafíos planteados por Herbert Simon. En 1957, aseguró la existencia de máquinas que podían pensar, aprender y crear, cuyas capacidades se incrementarían hasta igualar las del cerebro humano en el futuro inmediato. No obstante, la confianza excesiva de Simon se debía al rendimiento prometedor de los primeros sistemas de IA en tareas sencillas, y estos sistemas fracasaron en problemas más complejos debido a que se basaron mayormente en la introspección informada, en lugar de un análisis riguroso de la tarea en sí. La mayoría de los primeros sistemas de resolución de problemas funcionaron mediante la prueba de diversas combinaciones de pasos hasta encontrar la solución, lo que funcionó en principio en micromundos con pocos objetos y acciones posibles. No obstante, esta táctica no fue efectiva en problemas más grandes, debido a la complejidad de estos y a la falta de comprensión sobre la dificultad para encontrar soluciones (Russell & Norvig, pág. 21).

La explosión combinatoria fue uno de los principales desafíos de la IA, según el informe Lighthill. Esto llevó a una reconsideración de las políticas de apoyo gubernamental a la investigación en este campo, lo cual provocó que el gobierno británico pusiera fin al apoyo a la investigación en IA. También, se descubrió que los perceptrones, una forma sencilla de red neuronal, tenían limitaciones fundamentales en su capacidad para representar y aprender cosas nuevas, lo que provocó una reducción en la financiación de la investigación en redes neuronales. Los algoritmos de aprendizaje de retro propagación, que causaron un enorme resurgimiento en la investigación de redes neuronales en las décadas de 1980 y 2010, ya habían sido desarrollados en otros contextos en la década de 1960 (Russell & Norvig, 2020, pág. 21).

2.1.4 Sistemas expertos

Los enfoques para resolver problemas en la investigación en Inteligencia Artificial eran considerados débiles debido a su falta de escalabilidad. La alternativa era el uso de conocimientos específicos del dominio que permitieran razonamientos más grandes y manejara más fácilmente los casos típicos en áreas de experiencia estrechas. El programa DENDRAL, desarrollado en la Universidad de Stanford en 1969, fue un ejemplo temprano de este enfoque. Se usaron conocimientos especializados para inferir la estructura molecular a partir de la información proporcionada por un espectrómetro de masas. En 1971, los investigadores en Stanford iniciaron el proyecto Heuristic Programming Project (HPP) para investigar el método de sistemas expertos. El siguiente gran esfuerzo fue el sistema MYCIN para el diagnóstico de infecciones sanguíneas. A diferencia de las reglas DENDRAL, los expertos tuvieron que ser entrevistados para desarrollar las reglas de MYCIN (Russell & Norvig, 2020, pág. 22).

En 1982, se lanzó el primer sistema experto comercialmente exitoso, llamado R1, que revolucionó la configuración de pedidos de nuevos sistemas informáticos. Este avance destacó la importancia fundamental del conocimiento de dominio en el campo de la Inteligencia Artificial. Asimismo, se evidenció su relevancia en el área de la comprensión del lenguaje natural, impulsando el desarrollo de una amplia variedad de herramientas de representación y razonamiento que ampliaron aún más las capacidades de la IA en ese período. Para concluir, la IA en la década de 1969 a 1986 se centró en el uso de conocimientos especializados del dominio y el desarrollo de sistemas expertos para resolver problemas en áreas específicas (Russell & Norvig, 2020, pág. 22).

2.1.5 Retorno de las redes neuronales

En la década de los años 80, diversos colectivos investigativos redescubrieron el algoritmo de aprendizaje por retro propagación, aplicándolo con éxito a una multitud de dilemas de aprendizaje en los ámbitos de la informática y la psicología. La difusión de los resultados causó gran emoción y se les llamó modelos conexionistas. Estos modelos fueron vistos como competidores directos a los modelos simbólicos y lógicos promovidos por otros investigadores. Aunque los humanos parecen manipular símbolos, algunos argumentan que los modelos conexionistas pueden formar conceptos internos de manera más fluida e imprecisa, lo que los hace más adecuados para el mundo real. Adicionalmente, estos modelos tienen la capacidad de asimilar conocimiento mediante ejemplos y ajustar meticulosamente sus parámetros con el objetivo de perfeccionar su desempeño futuro (Russell & Norvig, 2020, pág. 24).

2.1.6 Razonamiento probabilístico y Machine Learning

Durante los últimos años de la década de 1980, la Inteligencia Artificial experimentó una evolución hacia un enfoque más científico, incorporando la probabilidad en lugar de la lógica booleana, el uso de Machine Learning en lugar de la programación manual y resultados experimentales en lugar de afirmaciones filosóficas. Se volvió más común construir sobre teorías existentes que proponer nuevas, basar afirmaciones en teoremas rigurosos o metodologías experimentales sólidas en lugar de la intuición, y demostrar relevancia en aplicaciones del mundo real en lugar de ejemplos de juguetes. Los avances positivos en áreas como el control y la estadística, que antes se veían como limitaciones, también fueron adoptados por la IA (Russell & Norvig, pág. 24).

En 1988, Judea Pearl y Rich Sutton hicieron importantes contribuciones que tuvieron un impacto significativo en el campo de la Inteligencia Artificial. Pearl desarrolló las redes Bayesianas, que permiten representar el conocimiento incierto y tienen algoritmos prácticos para el razonamiento probabilístico, mientras que Sutton estableció la conexión entre el aprendizaje por refuerzo con los procesos de decisión de Markov, lo que ha impulsado a una gran cantidad de investigaciones y aplicaciones en robótica y control de procesos. La Inteligencia Artificial se ha beneficiado de la reunificación de subcampos como la visión por computadora, la robótica, el reconocimiento del habla, los sistemas multiagentes y el procesamiento del lenguaje natural, la IA ha experimentado una mejora tanto en su comprensión teórica como en su aplicación práctica, especialmente en el desarrollo y despliegue de robots en diversos ámbitos (Russell & Norvig, 2020, pág. 24).

Gartner (2021) ofrece una descripción detallada de Machine Learning como una técnica fundamental que permite a la Inteligencia Artificial tener la capacidad de resolver problemas. A menudo hay una interpretación errónea sobre el aprendizaje de las máquinas, ya que en realidad su función consiste en almacenar y computar datos de manera cada vez más compleja. A través del entrenamiento, se aplican modelos matemáticos a los datos para extraer conocimiento y descubrir patrones que probablemente los humanos pasarían por alto. También, Machine Learning recomienda acciones, pero no dirige a los sistemas a tomar medidas sin intervención humana. Más específicamente, crea un algoritmo o fórmula estadística a la cual se denomina Modelo, que convierte una serie de datos en un resultado único. Los algoritmos de Machine Learning adquieren conocimiento a través de un proceso de entrenamiento, en el cual reconocen patrones y correlaciones en los datos, y utilizan esta información para generar nuevas ideas y predicciones, sin necesidad de ser programados explícitamente con dichas habilidades.

2.1.7 Big Data

En los últimos años, los avances sobre la informática y la creación de la World Wide Web, han permitido la generación de conjuntos de datos muy grandes, conocidos como Big Data. Estos conjuntos incluyen billones de palabras de texto, miles de millones de imágenes y horas de audio y video, así como también grandes cantidades de datos genómicos, de seguimiento de vehículos, de clics en sitios web, de redes sociales, entre otros. Esto ha llevado al desarrollo de algoritmos de aprendizaje especialmente diseñados para aprovechar al máximo estos grandes conjuntos de datos. A menudo, la gran mayoría de los ejemplos en estos conjuntos de datos no están etiquetados, pero con algoritmos de aprendizaje adecuados, se puede lograr una precisión del 96% o más en tareas de identificación de sentidos. El aumento en el tamaño de los conjuntos de datos ha demostrado ser más importante que la mejora de los algoritmos. Esto ha tenido un impacto en la visión artificial, donde llenar espacios en fotografías ha mejorado significativamente al tener bases de datos de millones de imágenes. Todo esto ha impulsado la comercialización de la IA y ha permitido el surgimiento de sistemas como Watson de IBM, que pudo vencer a humanos en un juego de preguntas y respuestas, lo que ha cambiado la percepción pública de la Inteligencia Artificial (Russell & Norvig, pág. 26).

2.1.8 Deep Learning

Deep Learning es una variante de los algoritmos de Machine Learning que utiliza varias capas de algoritmos para resolver problemas a partir de la extracción de conocimiento desde datos en bruto, transformándolos en cada nivel. Puede trabajar con datos complejos y de alta dimensión, como imágenes, voz y texto. En la mayoría de las empresas, no se ha incorporado aún Deep Learning en la planificación de productos, ya que los sistemas basados en reglas o Machine Learning tradicionales son suficientes para la mayoría de los casos de uso de IA en la actualidad. De todas maneras, su uso está aumentando rápidamente gracias a los avances en el procesamiento de datos y las técnicas computacionales. Al utilizar técnicas de Machine Learning e incluyendo Deep Learning para realizar predicciones, se puede automatizar la selección del resultado más favorable a través de un proceso impulsado por IA, lo que elimina la necesidad de la intervención humana en la toma de decisiones (Gartner, 2021).

Deep Learning emplea múltiples capas de componentes informáticos simples y adaptables en su técnica de aprendizaje automático. Aunque los estudios de estas redes se realizaron desde los años 70, y lograron cierto progreso en el reconocimiento de dígitos manuscritos en los años 90, no fue hasta el 2011 cuando se empezó a utilizar ampliamente este método de Deep Learning. Primero, en la identificación de voz y luego en el reconocimiento

visual de objetos. Durante la competencia de ImageNet en 2012, que consistía en la clasificación de imágenes en una de las mil categorías disponibles, un sistema de Deep Learning desarrollado en el grupo de Geoffrey Hinton en la Universidad de Toronto demostró una mejora significativa en comparación con los sistemas previos, que se basaban mayormente en características creadas manualmente. Desde entonces, los sistemas de Deep Learning han superado el rendimiento humano en algunas tareas de visión y se han quedado atrás en otras. Asimismo, se han registrado avances similares en el reconocimiento de voz, la traducción automática, el diagnóstico médico y los juegos (Russell & Norvig, pág. 26).

El éxito del Deep Learning ha renovado el interés en la Inteligencia Artificial, atrayendo la atención de estudiantes, empresas, inversores, gobiernos, medios de comunicación y el público en general. Se reportan constantemente nuevas aplicaciones de IA que están alcanzando o superando el rendimiento humano, lo que genera expectativas sobre su éxito en el futuro. Sin embargo, para lograr estos avances, Deep Learning requiere de hardware potente, ya que los algoritmos que se ejecutan en hardware especializado como GPU, TPU o FPGA, pueden realizar cientos de miles de millones de operaciones por segundo, principalmente en forma de operaciones matriciales y vectoriales altamente paralelizadas. Además, para lograr buenos resultados, se necesita una gran cantidad de datos de entrenamiento y el uso de algunos trucos algorítmicos (Russell & Norvig, 2020, pág. 26).

2.2 Beneficios y riesgos de la Inteligencia Artificial

El potencial de la IA y la robótica para liberar a la humanidad del trabajo repetitivo, y aumentar drásticamente la producción de bienes y servicios, podría presagiar una era de paz y abundancia. La capacidad de acelerar la investigación científica podría resultar en curas para enfermedades y soluciones para el cambio climático y la escasez de recursos (Russell & Norvig, 2020, pág. 31).

Por otro lado, Gartner (2021) afirma que las ventajas de la Inteligencia Artificial radican en su capacidad para encontrar formas más efectivas de realizar tareas, a través de un análisis avanzado de probabilidades en los resultados. Además, permite interactuar directamente con sistemas que llevan a cabo acciones, lo que reduce la necesidad de cálculos intensivos en humanos y pasos de integración. Los líderes de IT y análisis de datos ven una gran oportunidad en los beneficios de la Inteligencia Artificial, pero experimentan dificultades para implementarla. A pesar de ello, la Inteligencia Artificial tendrá un impacto significativo en la forma en que se desempeñan las tareas laborales a medida que reemplace algunas de las tareas tradicionalmente manuales realizadas por empleados y modifique la forma en que se toman decisiones cotidianas.

Los casos de uso más comunes de la IA incluyen la automatización y optimización, la generación de información y la creación de interacciones similares a las humanas, como los chatbots y los asistentes virtuales.

No obstante, en la actualidad el entusiasmo desmedido en torno a la IA puede ser excesivo, lo cual complica a algunas organizaciones el establecimiento de expectativas realistas con respecto a los resultados. Los líderes empresariales que tienen estas expectativas poco realistas culpan a la tecnología y la ciencia por no poder generar las transformaciones que esperaban. Resulta fundamental crear una estrategia sólida para el uso de la IA que permita identificar los casos de aplicación y las métricas de éxito desde el inicio. Entre las formas más comunes de evaluar los beneficios de la IA, se encuentran la disminución de riesgos, la aceleración de procesos, el incremento de ventas, el aumento de la satisfacción del cliente y la reducción de costos y necesidades de mano de obra. En muchos casos, los beneficios que aporta la IA se conforman por una combinación de aspectos tangibles e intangibles (Gartner, 2021).

En relación con los beneficios, Google (2022) menciona en primer lugar, que la IA destaca su capacidad de automatizar flujos de trabajo y procesos, como ocurre en la seguridad cibernética y en fábricas inteligentes. Esto se logra mediante el uso de robots y analítica en tiempo real. Además, la IA contribuye a la reducción de errores al eliminar los errores humanos presentes en el procesamiento de datos, el ensamblaje y otras tareas. Esta eliminación se lleva a cabo mediante la automatización y el uso de algoritmos precisos. Otro beneficio importante de la IA es su capacidad de eliminar tareas repetitivas. De esta manera, la IA puede realizar tareas aburridas o peligrosas, permitiendo así que el capital humano se libere y pueda abordar problemas de mayor impacto. La IA también se caracteriza por ser rápida y precisa en el procesamiento de información. En comparación con los humanos, la IA procesa información de manera más rápida, gracias a su capacidad para buscar patrones y descubrir relaciones en los datos. Asimismo, la IA ofrece una disponibilidad infinita, ya que no está sujeta a limitaciones de horarios. Esto permite que pueda trabajar de manera continua en las tareas asignadas, especialmente cuando se ejecuta en la nube. Por último, la IA acelera la investigación y el desarrollo en diversas áreas. Al facilitar el análisis de grandes cantidades de datos, la IA mejora la investigación y el desarrollo, especialmente en áreas como el modelado predictivo de tratamientos farmacéuticos y la cuantificación del genoma humano.

Russell y Norvig (2020), determinan que a medida que la Inteligencia Artificial juega un papel cada vez más importante en los ámbitos económico, social, científico, médico, financiero y militar, correremos riesgos por el mal uso de la IA. Algunos de estos ya son evidentes, mientras que otros parecen probables según las tendencias actuales.

Las armas autónomas letales son aquellas que, según las Naciones Unidas, poseen la capacidad de localizar, seleccionar y eliminar objetivos humanos sin necesidad de intervención humana. El riesgo principal de estas armas es su capacidad de escalar sin requerir supervisión humana, lo que significa que un grupo pequeño podría utilizar un gran número de armas contra objetivos humanos definidos según cualquier criterio de reconocimiento viable (Russell & Norvig, 2020, pág. 31).

Continuando con la exposición de los autores mencionados antes, la vigilancia y persuasión han sido temas que plantean interrogantes legales para el personal de seguridad, especialmente en lo que respecta al monitoreo de líneas telefónicas, cámaras de video, correos electrónicos y otros canales de comunicación. En cambio, la IA puede ser utilizada de manera escalable para realizar vigilancia masiva de individuos y detectar actividades de interés, mediante técnicas de reconocimiento de voz, visión por computadora y comprensión del lenguaje natural. A través de la adaptación del flujo de información a los usuarios a través de las redes sociales, basados en Machine Learning, se puede modificar y controlar el comportamiento político, lo que se ha convertido en una preocupación desde las elecciones de 2016 (pág. 31).

De acuerdo con la postura sostenida de Russell & Norvig (2020), la toma de decisiones sesgadas es una preocupación relevante en la situación donde se emplean algoritmos de Machine Learning para evaluar solicitudes de libertad condicional y préstamos, esto puede resultar en decisiones sesgadas basadas en la raza, el género y otras categorías protegidas. Esto se debe a que, con frecuencia, los datos mismos reflejan el sesgo generalizado en la sociedad. En el mismo sentido, el impacto en el empleo es un tema importante para considerar en el contexto de la automatización y la introducción de máquinas en diversas tareas que anteriormente realizaban los seres humanos. Aunque las máquinas pueden realizar tareas que antes hacían los humanos, también se ha observado que aumentan la productividad de los trabajadores, lo que hace que las empresas sean más rentables y puedan pagar salarios más altos. Además, pueden hacer económicamente viables algunas actividades que de otra manera serían imprácticas. Sin embargo, esto puede tener el efecto de desplazar la riqueza del trabajo al capital, exacerbando aún más el aumento de la desigualdad. Aunque los avances tecnológicos anteriores han causado graves interrupciones en el empleo, las personas eventualmente encuentran nuevos tipos de trabajo para hacer (pág. 31).

Continuando con los riesgos planteados por los autores Russell y Norvig (2020), las aplicaciones críticas de seguridad se han convertido en un ámbito cada vez más importante para las técnicas de IA, como la conducción de vehículos y la gestión del suministro de agua de las ciudades. Ya se han producido accidentes mortales que destacan la dificultad de la verificación

formal y el análisis estadístico de riesgos para sistemas desarrollados con técnicas de Machine Learning. Por lo tanto, el campo de la IA necesita desarrollar estándares técnicos y éticos al menos comparables a los prevalentes en otras disciplinas de ingeniería y atención médica donde están en juego vidas humanas (pág. 31).

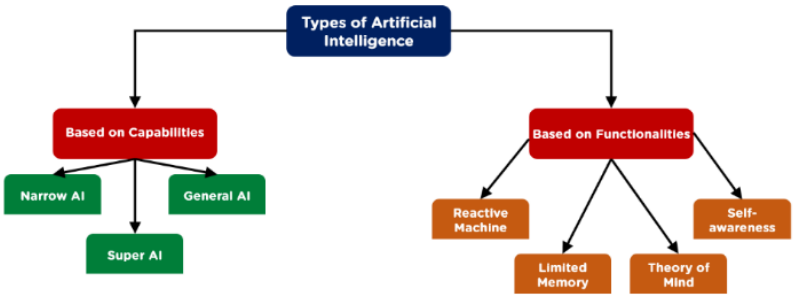
La ciberseguridad es un ámbito en el que las técnicas de IA juegan un papel relevante, son útiles para defenderse contra ciberataques, como en la detección de patrones de comportamiento inusuales. También pueden contribuir a la potencia, supervivencia y capacidad de proliferación del malware. Los métodos de aprendizaje por refuerzo se han utilizado para crear herramientas altamente efectivas para ataques de chantaje y phishing automatizados y personalizados (Russell & Norvig, 2020, pág. 31).

2.3 Tipos de Inteligencia Artificial

La Inteligencia Artificial se clasifica según diversos criterios. En cuanto a la capacidad, la IA se divide en tres tipos: Inteligencia Artificial Débil o Estrecha, Inteligencia Artificial Fuerte o General e Inteligencia Artificial Súper. En cuanto a la funcionalidad, hay cuatro tipos de IA: Máquinas Reactivas, Memoria Limitada, Teoría de la mente y Autoconciencia (Javatpoint, 2021). En la siguiente figura, se mencionan los tipos de IA junto a sus subtipos:

Figura 1

Clasificación de la Inteligencia Artificial



Nota. Esta ilustración presenta los tipos de IA clasificados en base a la capacidad y en funcionalidad. Adaptado de *7 Types of Artificial Intelligence That You Should Know in 2023* de Simplilearn, 2023, https://www.simplilearn.com/tutorials/artificial-intelligence-tutorial/types-of-artificial-intelligence

2.3.1 IA basada en Capacidad

La Inteligencia Artificial Débil, también conocida como Estrecha o Artificial Narrow Intelligence (ANI), engloba todos los sistemas de IA actualmente disponibles. Tiene la capacidad

de llevar a cabo una tarea particular con habilidades similares a las humanas. Estos sistemas presentan un alcance de competencia bastante limitado o estrecho. Por consiguiente, se clasifican como sistemas reactivos de IA con una capacidad de memoria restringida (Joshi, 2019). A pesar de sus limitaciones, estos sistemas tienen un uso extenso en el mundo actual. Los algoritmos utilizados por los motores de búsqueda son un ejemplo de la complejidad de la modelización predictiva en ANI. Aunque estos sistemas no pueden improvisar ni crecer, aún son muy útiles en tareas específicas (Techliance, 2023). Según Gamco (2021), en esta categoría se incluyen desde asistentes virtuales como Alexa o Siri hasta sistemas autónomos de vehículos, filtros de correo no deseado o recomendaciones publicitarias basadas en búsquedas. La Inteligencia Artificial Estrecha se fundamenta en la aplicación de una secuencia de acciones y directrices en el sistema informático.

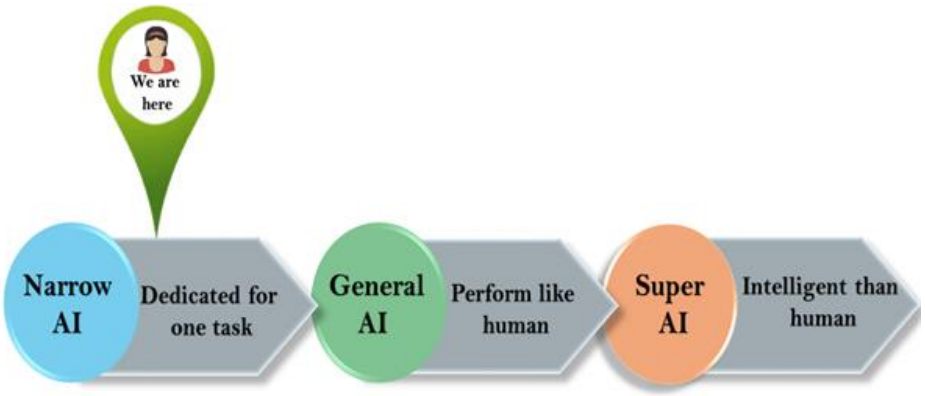
La Inteligencia Artificial Fuerte, General o también conocida como Artificial General Intelligence (AGI), engloba la habilidad de adquirir conocimiento, interpretar, comprender y operar de manera completa como un ser humano (Joshi, 2019). Javatpoint (2021), indica que la AGI es un tipo de inteligencia que podría realizar cualquier tarea intelectual con eficiencia, pero su desarrollo aún está en proceso. Lateef (2023), destaca que la AGI es el siguiente paso en la evolución de la Inteligencia Artificial, ya que permitiría a las máquinas pensar y tomar decisiones como los humanos. La AGI permitiría a un sistema de IA aplicar conocimientos y habilidades en diferentes contextos, aunque actualmente no existe un sistema que cumpla con estas características (Biswal, 2023).

Por último, la Superinteligencia Artificial (ASI) sobrepasa la capacidad intelectual humana y es capaz de ejecutar cualquier tarea de manera superior a un ser humano. La idea de la superinteligencia artificial considera que la IA evolucionada posee similitudes con los sentimientos y experiencias humanas, ya que no solo las comprende, sino que también es capaz de evocar emociones, necesidades, creencias y deseos propios. Aunque su existencia es aún hipotética, algunas de sus características clave incluyen la capacidad de pensar, resolver rompecabezas, y tomar decisiones por su cuenta (Biswal, 2023). El punto culminante de la investigación en IA probablemente será considerado el desarrollo de este tipo de IA, ya que se espera que será la forma de inteligencia más sobresaliente en la Tierra. Además de la reproducción de la inteligencia multifacética de los seres humanos, se espera que todas las capacidades de la ASI se destaquen excepcionalmente debido a su amplia memoria, su velocidad de procesamiento y análisis de datos superior, y su capacidad de toma de decisiones (Joshi, 2019). Según Lateef (2023), la ASI es actualmente una situación hipotética como se muestra en películas y libros de ciencia ficción, donde las máquinas han conquistado el mundo. La autora

cree que las máquinas no están muy lejos de alcanzar este nivel, teniendo en cuenta nuestro ritmo actual. A continuación, en la Figura 2 se puede visualizar en que etapa nos encontramos en la actualidad.

Figura 2

Etapa actual de la IA



Nota. Esta ilustración representa la etapa en la que nos encontramos en referencia al tipo de IA que estamos utilizando. Adaptado de *Types of Artificial Intelligence* de Javatpoint, 2021, https://www.javatpoint.com/types-of-artificial-intelligence

2.3.2 IA Basada en Funcionalidad

Las Máquinas Reactivas son sistemas de IA de los primeros tipos que existieron, y poseen una capacidad muy limitada. La habilidad de reacción a diferentes tipos de estímulos que posee la mente humana es emulada por estas máquinas. No se basa en la memoria, lo que implica que no pueden utilizar experiencias previas para guiar sus acciones actuales, es decir, carecen de la capacidad de aprendizaje. Estas máquinas se limitan a responder automáticamente a un conjunto específico o combinación limitada de entradas. No es posible utilizar la memoria para mejorar sus operaciones basadas en este aspecto (Joshi, 2019). Una máquina reactiva es la forma primaria de Inteligencia Artificial que no almacena recuerdos ni utiliza experiencias pasadas para determinar las acciones futuras, solo funciona con datos presentes, perciben el mundo y reaccionan ante él. A las máquinas reactivas se les proporcionan tareas específicas, y no tienen capacidades más allá de esas tareas. Por ejemplo, en un juego de ajedrez, la máquina observa los movimientos y toma la mejor decisión posible para ganar (Biswal, 2023).

Las IA de Memoria Limitada o Limited Memory Machines aprenden a partir de datos históricos para tomar decisiones. Los sistemas de IA actuales, como aquellos basados en el Deep Learning, son sometidos a entrenamiento mediante el procesamiento de extensos conjuntos de

datos de entrenamiento, los cuales son almacenados en su memoria con el propósito de establecer un modelo de referencia que les posibilite enfrentar problemas futuros (Joshi, 2019). Es una tecnología utilizada en vehículos autónomos, la cual entrena a partir de datos pasados para tomar decisiones, pero su memoria es limitada en el tiempo y no puede agregarla a una biblioteca de experiencias. Este tipo de AI observa cómo se mueven otros vehículos y la información se agrega a los datos estáticos del vehículo, como los marcadores de carril y las luces de tráfico, para decidir cuándo cambiar de carril o evitar chocar con otro vehículo. Mitsubishi Electric está trabajando en mejorar esta tecnología (Biswal, 2023).

La Teoría de la Mente, conocida como Theory of Mind, está siendo objeto de investigación como el próximo nivel de desarrollo de los sistemas de IA. Se espera que se logre un nivel de IA que permita una comprensión más profunda de las entidades con las que interactúa, reconociendo sus necesidades, emociones, creencias y procesos de pensamiento. Aunque la inteligencia emocional artificial ya se encuentra en desarrollo y es un área de interés para los principales investigadores en IA, alcanzar el nivel de Theory of Mind AI requerirá avances en otros campos de la IA. Esto se debe a que, para comprender realmente las necesidades humanas, los sistemas de IA deberán percibir a los seres humanos como individuos cuyas mentes son influenciadas por múltiples factores, logrando así una comprensión esencial de los seres humanos (Joshi, 2019). Esta IA representa una clase avanzada de tecnología y solo existe como concepto, requiere una comprensión profunda de que las personas y las cosas dentro de un entorno pueden alterar los sentimientos y comportamientos. Debe comprender las emociones, sentimientos y pensamientos de las personas (Biswal, 2023). Esta categoría de máquinas se especula que desempeñará un papel importante en la psicología, se centrará en la inteligencia emocional para que se pueda comprender mejor las creencias y pensamientos humanos. Theory of Mind aún no se ha desarrollado por completo, pero se está realizando una investigación rigurosa en esta área (Lateef, 2023).

Por último, la Autoconciencia o Self-awareness según el artículo de Forbes (Joshi, 2019), coincide en que este es el último nivel de desarrollo de la IA y que solo existe hipotéticamente. El objetivo final de la investigación en IA siempre ha sido crear una IA con autoconciencia similar a la del cerebro humano, capaz de entender y evocar emociones en otros y tener emociones, necesidades, creencias y deseos propios. Si bien esto podría impulsar el progreso humano, también es motivo de preocupación, ya que una IA con autoconciencia y la capacidad de autopreservación podría superar la capacidad intelectual de los humanos y tomar medidas para su propia supervivencia, lo que podría llevar a la catástrofe. Aunque no hay una IA autoconsciente en la actualidad, algunos científicos tienen la autoconciencia como horizonte. La

máquina tendría que entender que hay individuos con emociones y pensamientos propios, lo que sería un nivel de complejidad actualmente inalcanzable para los sistemas informáticos. Si se logra desarrollar una IA autoconsciente, las máquinas podrían aprender de nuestros comportamientos y deducir nuestros gustos, necesidades y deseos (Gamco, 2021). Lateef (2023), advierte sobre el peligro de una IA con autoconciencia y pide que no se llegue a ese punto, aunque reconoce que la superinteligencia podría ser posible en el futuro. Elon Musk y Stephen Hawking también han advertido sobre los riesgos de la IA autoconsciente.

2.4 Innovaciones en Inteligencia Artificial

En el campo de la Inteligencia Artificial existen cuatro categorías principales de innovaciones, que se espera tengan un gran impacto tanto dentro como fuera del ámbito empresarial, afectando tanto a las personas como a los procesos empresariales. Es importante que las partes interesadas, desde los líderes empresariales hasta los equipos de ingeniería, estén al tanto de estas categorías para poder implementar y gestionar sistemas de IA de manera efectiva (Wiles, 2022). A continuación, se detallan cada una de estas categorías.

2.4.1 Inteligencia Artificial centrada en datos

La Inteligencia Artificial orientada a los datos se centra en mejorar y enriquecer los conjuntos de datos empleados para el entrenamiento de los algoritmos. Esta nueva forma de abordar los datos específicos de la IA representa un cambio disruptivo en la gestión de datos tradicionales. No obstante, las organizaciones que implementan la IA a gran escala deben transformarse con el fin de mantener los principios fundamentales de gestión de datos, pero ajustándolos a las exigencias de la IA. La Inteligencia Artificial ofrece la oportunidad de fortalecer y potenciar aspectos fundamentales y duraderos como la gobernanza de datos, la persistencia de estos, su integración y calidad. Las innovaciones en IA basada en datos incluyen los datos sintéticos que se generan artificialmente en lugar de ser obtenidos a partir de observaciones directas del mundo real. En diferentes industrias, se observa un crecimiento en la adopción de datos sintéticos, los cuales se utilizan para entrenar modelos de Machine Learning sin necesidad de recurrir a información personalmente identificable. Además, reducen costos, ahorran tiempo en el desarrollo, y mejoran el rendimiento del Machine (Wiles, 2022).

2.4.2 Inteligencia Artificial centrada en el modelo

A pesar de la transición hacia una orientación basada en datos, es vital estar atentos a los modelos de Inteligencia Artificial para asegurar que los resultados sigan siendo beneficiosos para

la toma de decisiones. En este sentido, las últimas novedades en IA, como los modelos que se basan en física, IA causal, IA generativa, modelos base y aprendizaje profundo, resultan de gran importancia. La Inteligencia Artificial compuesta hace referencia a la fusión de múltiples técnicas de IA con el objetivo de optimizar el proceso de aprendizaje y extender el nivel de comprensión. Debido a que ninguna técnica de IA es suficiente por sí sola, la IA compuesta brinda una plataforma para abordar una amplia variedad de desafíos empresariales de manera más eficaz. Es probable que la adopción de la IA compuesta se expanda significativamente durante los próximos dos a cinco años, alterando la manera en que las empresas llevan a cabo sus operaciones en todos los ámbitos. Un ejemplo es que la IA compuesta puede brindar la capacidad de la IA a empresas que carecen de acceso a grandes cantidades de datos históricos o etiquetados, pero que cuentan con habilidades expertas considerables en personas (Wiles, 2022).

La IA causal, que emplea técnicas como gráficos y simulación causal, es útil para mejorar la toma de decisiones al descubrir relaciones causales. Se espera que esta tecnología se adopte ampliamente en los próximos 5 a 10 años, lo que tendrá un enorme impacto en las empresas al permitir la integración de nuevos procesos horizontales o verticales que reduzcan costos y aumenten ingresos. Los beneficios de la IA causal incluyen mejoras en la eficiencia, la toma de decisiones más autónoma y potenciada, una explicabilidad mejorada y una mayor robustez y adaptabilidad, así como una reducción del sesgo en los sistemas de IA (Wiles, 2022).

2.4.3 Inteligencia Artificial centrada en aplicaciones

Dentro de este campo, se han generado diversas innovaciones como la ingeniería de Inteligencia Artificial, los sistemas operativos de IA, el ModelOps, los servicios de IA basados en la nube, los robots inteligentes, el procesamiento del lenguaje natural, los vehículos autónomos, las aplicaciones inteligentes y la visión por computadora. Se anticipa que la inteligencia de decisiones y la Inteligencia Artificial serán ampliamente adoptadas en el perímetro en los próximos dos a cinco años, generando impactos transformadores en las empresas. La inteligencia de decisiones es una práctica utilizada para mejorar la toma de decisiones a través del conocimiento y el diseño explícito de cómo se toman y evalúan las decisiones, y cómo se potencian los resultados mediante el feedback. Los beneficios de la inteligencia de decisiones incluyen la reducción de la deuda técnica, el aumento de la visibilidad, la mejora de los procesos empresariales, así como la transparencia y auditabilidad de las decisiones. La Inteligencia Artificial en el perímetro implica el uso de técnicas de IA integradas en los terminales, portales y servidores perimetrales del Internet de las cosas, para aplicaciones que abarcan desde los automóviles sin conductor hasta el análisis en tiempo real de transmisiones

de datos. Algunos de sus beneficios comerciales incluyen la mejora de la eficiencia operativa, la experiencia del cliente, la reducción de la latencia en la toma de decisiones y en el costo de conectividad, además de la disponibilidad permanente de soluciones independientes de la conectividad de red (Wiles, 2022).

2.4.4 Inteligencia Artificial centrada en las personas

Cuando la Inteligencia Artificial toma decisiones en lugar de los humanos, se busca alcanzar resultados positivos y abordar los desafíos relacionados con aspectos como el valor agregado y la disposición a asumir riesgos. En este contexto, los dilemas se refieren a los problemas éticos, morales o prácticos que puedan surgir al delegar decisiones a sistemas automatizados, incluyendo cuestiones sobre ética, responsabilidad y posibles consecuencias no deseadas (Wiles, 2022).

Es necesario considerar aspectos como la implementación ética de la IA, el valor empresarial y social, la confianza, la transparencia, la justicia, la reducción de sesgos, la explicabilidad, la responsabilidad, la seguridad, la privacidad y el cumplimiento normativo. En cuanto a la adopción generalizada de la IA responsable, se prevé que ocurra en los próximos 5 a 10 años, mientras que la ética digital, como menciona el mismo autor, es una tendencia a corto plazo, con un horizonte de 2 a 5 años. La ética digital se refiere a los valores y principios morales de conducta en las interacciones electrónicas, y plantea preocupaciones relevantes en relación con la privacidad y el sesgo. En este sentido, las organizaciones están implementando medidas para mitigar los riesgos asociados con la gestión y protección de datos personales, y los gobiernos están estableciendo leyes más estrictas en este ámbito (Wiles, 2022).

A pesar de estas consideraciones, es importante destacar que muchas empresas aún no consideran la ética digital como relevante para su sector. No obstante, Wiles (2022), señala que se espera que para el año 2024, el 30% de las principales organizaciones incluyan la voz de la sociedad como un indicador para abordar las preocupaciones sociales y evaluar el impacto de su desempeño comercial. En este contexto, será necesario que las organizaciones integren la ética digital en sus estrategias de IA con el fin de fortalecer su influencia y reputación en todos los ámbitos pertinentes.

2.5 Áreas de aplicación

La aplicación de la Inteligencia Artificial se extiende a través de varios sectores, organizaciones y funciones, manifestando su potencial de maneras diversas. En su análisis, se destaca como los chatbots, vehículos autónomos y robots inteligentes son impulsados por

Machine Learning en el área de las comunicaciones. Adicionalmente, el aprendizaje profundo ofrece alternativas biométricas a través de la identificación facial, de voz y el uso de redes neuronales, adaptando el contenido en función de la extracción de datos y la detección de patrones en amplios conjuntos de información (Gartner, 2021). En la Figura 3, Gartner proyecta el tiempo en el que la aplicación de IA alcanzará la meseta de productividad.

Figura 3

Proyección de la meseta de productividad de IA

Nota. La figura muestra las expectativas de adopción de IA en los próximos años. Adaptado de *Novedades del Hype Cycle de Gartner para la Inteligencia Artificial de 2022* de Gartner, 2022, https://www.gartner.es/es/articulos/novedades-del-hype-cycle-de-gartner-para-la-inteligencia-artificial-2022

Gartner (2021) menciona que, en las operaciones de IT y el servicio de Mesa de Ayuda, se brinda apoyo en la asignación de tickets y la extracción de información de fuentes de conocimiento mediante los agentes de soporte virtual. La gestión de la cadena de suministro se beneficia de la IA en áreas como el mantenimiento predictivo, gestión de riesgos, adquisiciones y planificación. Además, en las ventas y el marketing, se identifican potenciales clientes, se mejora la ejecución de las ventas y se permite la personalización en tiempo real, optimización de contenido y orquestación de campañas. En el servicio al cliente, se estiman las necesidades de los usuarios y se ofrecen opciones de autoservicio y asistencia las 24 horas a través de

asistentes virtuales. En el área de Recursos Humanos, facilita el reclutamiento, mejora la búsqueda y coincidencia de candidatos, y se utiliza el procesamiento del lenguaje natural para establecer sistemas de clasificación consistentes de habilidades y empleos. Por último, en el campo legal, se utilizan aplicaciones como el ensamblaje, la negociación, investigación exhaustiva y la gestión del ciclo de vida de los contratos mediante la IA. También se aplica en la investigación electrónica, la clasificación de documentos y facturas, así como también la gestión de gastos.

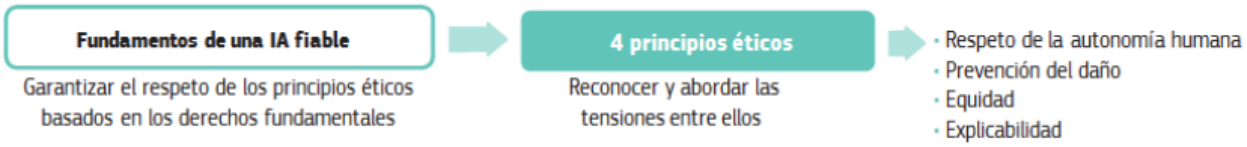
Estos son solo algunos ejemplos de cómo la IA se integra en el mundo empresarial, transformando y optimizando diversas áreas y funciones. Su adopción sigue en aumento, y el potencial de esta tecnología promete impulsar la eficiencia, la toma de decisiones y el crecimiento en un amplio espectro de sectores. La IA es una tecnología nueva que todavía no ha sido aprovechada al máximo en cuanto a sus impactos y beneficios. La innovación en IA es una de las fuerzas que está cambiando los mercados existentes y permitiendo nuevas iniciativas de negocios digitales. Además, la IA se está utilizando en diferentes industrias, organizaciones y funciones (Gartner, 2021).

2.6 Fundamentos de una Inteligencia Artificial fiable

Con el propósito de mejorar el bienestar individual y colectivo, se han establecido cuatro principios éticos arraigados en los derechos fundamentales, los cuales deben ser cumplidos para asegurar la fiabilidad de los sistemas de IA. Estos principios éticos son considerados como imperativos éticos que deben ser observados en todo momento por los profesionales de la IA (Comisión Europea, 2019). En la Figura 4, se describe la estructura de los Fundamentos de una IA fiable.

Figura 4

Fundamentos de una IA fiable y sus cuatro principios éticos



Nota. La figura presenta una representación visual de los fundamentos de una IA fiable, destacando los cuatro principios éticos clave: respeto de la autonomía humana, prevención del daño, equidad y explicabilidad. Adaptado de *Directrices éticas para una IA fiable* de la Comisión Europea, 2019, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60423

2.6.1 Principios éticos en el contexto de los sistemas de Inteligencia Artificial

Los cuatro principios éticos para una IA fiable consisten en el respeto de la autonomía humana, la prevención del daño, la equidad y la explicabilidad. Estos principios proveen una sólida base ética para el desarrollo y el uso responsable de los sistemas de IA, asegurando su capacidad de adaptación dinámica a medida que la sociedad evoluciona (Comisión Europea, 2019).

El principio del respeto de la autonomía humana se deriva de los derechos fundamentales respaldados por la Unión Europea, los cuales se enfocan en garantizar el respeto a la libertad y autonomía de los seres humanos. Es esencial que las personas que interactúan con sistemas de IA puedan mantener un control total y efectivo sobre sí mismas y participar en el proceso democrático. Es inapropiado emplear los sistemas de IA de forma injustificada con el propósito de subyugar, forzar, engañar, manipular, condicionar o controlar a los individuos. En cambio, se busca que estos sistemas sean diseñados de tal manera que mejoren, complementen y potencien las habilidades cognitivas, sociales y culturales de las personas. Los principios de diseño centrados en las personas deben guiar la distribución de responsabilidades entre los seres humanos y los sistemas de IA, permitiendo amplias oportunidades para la elección humana. Esto implica asegurar que los procesos de trabajo de los sistemas de IA estén supervisados y controlados por los seres humanos. Además, se reconoce el potencial transformador de los sistemas de IA en el ámbito laboral y se espera que brinden apoyo a las personas en su entorno laboral, con el objetivo de crear empleos valiosos y útiles (pág. 14).

Como menciona la Comisión Europea (2019), el principio de prevención del daño establece que los seres humanos no deben ser perjudicados de ninguna manera por los sistemas de IA, evitando causar daños adicionales o agravar los existentes. Esto implica la salvaguardia de la integridad física y mental, así como la protección de la dignidad humana. Para garantizar la seguridad, todos los entornos y sistemas de IA deben ser seguros y técnicamente sólidos, evitando su uso con intenciones maliciosas. Es importante prestar una atención especial a las personas vulnerables, permitiéndoles participar en el desarrollo y despliegue de los sistemas de IA. Además, se debe prestar especial atención a situaciones en las que los sistemas de IA pueden generar efectos adversos o empeorar la situación existente debido a desequilibrios de poder o información, como en las relaciones entre empleadores y trabajadores, empresas y consumidores, o gobiernos y ciudadanos. La prevención del daño también implica considerar el entorno natural y proteger a todos los seres vivos (pág. 15).

El principio de equidad se fundamenta en el desarrollo, despliegue y uso equitativo de los sistemas de IA. Se busca garantizar una distribución justa y equitativa de beneficios y costos,

así como prevenir sesgos injustos, discriminación y estigmatización hacia personas y grupos. Si se logra evitar sesgos injustos, los sistemas de IA podrían incluso promover la equidad social. Es importante fomentar la igualdad de oportunidades en términos de acceso a la educación, bienes, servicios y tecnología. Es fundamental que el uso de sistemas de Inteligencia Artificial nunca conduzca a engañar a los usuarios finales ni a limitar su libertad de elección. El principio de equidad exige que los expertos en IA respeten el principio de proporcionalidad entre los medios y los fines, y que consideren de manera cuidadosa cómo lograr un equilibrio entre diferentes intereses y objetivos en conflicto. La faceta procesal de la equidad conlleva la habilidad de cuestionar las decisiones adoptadas por los sistemas de IA y las personas encargadas de su administración, además de buscar una compensación adecuada en respuesta a dichas decisiones. Con este propósito, es importante poder identificar a la entidad responsable de la decisión y explicar los procesos de toma de decisiones (Comisión Europea, 2019, pág. 15).

El principio de explicabilidad enfatiza la importancia de generar confianza en los usuarios y mantenerla en los sistemas de IA. Para lograrlo, es crucial que los procesos sean transparentes y se comunique abiertamente sobre las capacidades y el propósito de los sistemas de IA. Las decisiones deben poder explicarse, en la medida de lo posible, a todas las partes afectadas directa o indirectamente por ellas. Sin esta información, impugnar adecuadamente una decisión se vuelve imposible. En algunos casos, donde los modelos generan resultados o decisiones de manera inexplicable, también conocidos como algoritmos de caja negra, se requiere una atención especial. Bajo estas condiciones, puede resultar necesario adoptar medidas adicionales en relación con la explicabilidad, como la trazabilidad, la auditabilidad y la comunicación transparente sobre el rendimiento del sistema, siempre y cuando se salvaguarden los derechos fundamentales en su totalidad. La importancia de la explicabilidad varía según el contexto y la gravedad de las consecuencias que puedan surgir debido a resultados incorrectos o inadecuados (Comisión Europea, 2019, pág. 16).

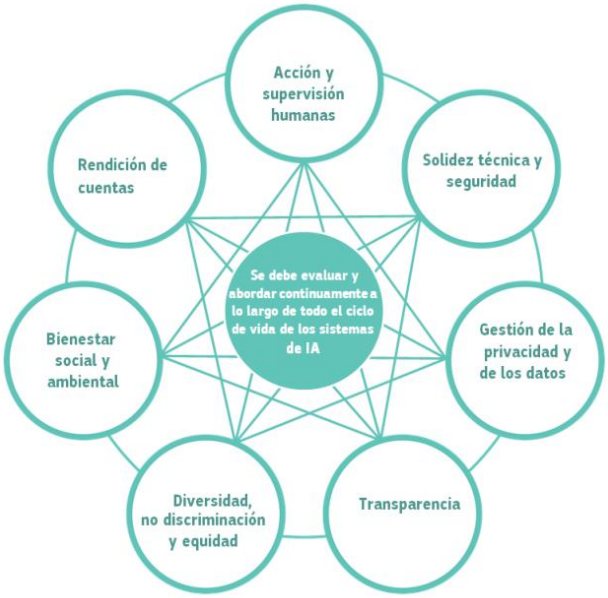
2.6.2 Requisitos de una Inteligencia Artificial fiable

Es esencial convertir los principios éticos en requisitos tangibles. Estos requisitos son aplicables a todas las partes involucradas en el ciclo de vida de los sistemas de IA. Los desarrolladores, aquellos que se dedican a la investigación, diseño o desarrollo de sistemas de IA, deben incorporar y aplicar estos requisitos en sus procesos. Por otro lado, los responsables del despliegue son las organizaciones que utilizan estos sistemas en sus procesos internos y para ofrecer productos y servicios a otros actores. Tienen la responsabilidad de garantizar que los sistemas que utilizan y los productos y servicios que ofrecen cumplan con los requisitos

establecidos. Los usuarios finales, a su vez, son aquellos que interactúan directa o indirectamente con los sistemas de IA. La sociedad en su conjunto abarca a todas las personas y entidades afectadas directa o indirectamente por los sistemas de IA. En general, deben estar informados sobre estos requisitos y tener la capacidad de exigir su cumplimiento (Comisión Europea, 2019, pág. 17). En el siguiente bloque, se presenta la Figura 5 que enumera los requisitos de una Inteligencia Artificial confiable.

Figura 5

Interrelaciones existentes entre los siete requisitos



Nota. La figura ilustra las conexiones entre los siete requisitos fundamentales para lograr una Inteligencia Artificial confiable. Adaptado de *Directrices éticas para una IA fiable* de la Comisión Europea, 2019, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60423

Aunque todos los requisitos son igualmente importantes, es necesario considerar el contexto y las posibles tensiones entre ellos al aplicarlos en diferentes sectores y ámbitos. Estos requisitos deben cumplirse durante todo el ciclo de vida de un sistema de Inteligencia Artificial, adaptándose a su aplicación específica. Aunque la mayoría de ellos son aplicables a todos los sistemas de IA, se presta especial atención a aquellos que afectan directa o indirectamente a las personas, aunque puedan ser menos relevantes en ciertas aplicaciones industriales (Comisión Europea, 2019, pág. 19). En los próximos párrafos se examinan minuciosamente cada uno de los requisitos, abordando su análisis detallado y comprensivo.

2.6.2.1 Acción y supervisión humanas

Los sistemas de IA deben respaldar la autonomía y la toma de decisiones de las personas, promoviendo los derechos fundamentales y permitiendo la supervisión humana. Es esencial evaluar el impacto en los derechos antes de desarrollar sistemas de IA y establecer mecanismos para considerar opiniones externas sobre los sistemas que podrían vulnerar los derechos fundamentales. Los usuarios deben tener la capacidad de tomar decisiones autónomas e interactuar de manera informada con los sistemas de IA. En cambio, se debe tener precaución con los sistemas que intenten influir en el comportamiento humano de manera injusta o manipuladora, ya que esto amenaza la autonomía individual. La presencia de supervisión humana asegura que los sistemas de IA no comprometan la autonomía de los seres humanos ni produzcan efectos negativos. Esto se logra a través de mecanismos de gobernanza, como la participación, el control y el mando humano. Es necesario que los responsables públicos puedan ejercer la supervisión de acuerdo con sus mandatos. La implementación de mecanismos de supervisión dependerá del ámbito y el riesgo del sistema de IA. Cuanto menor sea la supervisión humana, más rigurosas deben ser las verificaciones y la gobernanza necesarias (Comisión Europea, 2019, pág. 19).

En complemento a lo detallado por la Comisión Europea, la Institución White House Office of Science and Technology Policy (2022) indica que se pueden rechazar sistemas de IA cuando sea apropiado, basado en expectativas razonables y priorizando la accesibilidad y protección del público. En ocasiones, la ley puede exigir alternativas humanas. También se debe garantizar acceso oportuno a consideración y soluciones humanas en casos de fallas o deseos de apelación. Estas alternativas deben ser accesibles, equitativas, efectivas y sostenibles, con formación adecuada para los operadores, evitando cargas excesivas. En ámbitos sensibles, como justicia penal, empleo, educación y salud, se requiere un enfoque adicional que incluya supervisión, capacitación y consideración de decisiones adversas. Debe hacerse pública información sobre estos procesos, incluyendo evaluaciones de prontitud, accesibilidad, resultados y efectividad (pág. 7).

2.6.2.2 Solidez técnica y seguridad

La solidez técnica de los sistemas de IA es crucial para prevenir daños y asegurar su correcto funcionamiento. Esto implica desarrollarlos de manera preventiva, minimizando los daños involuntarios e imprevistos, incluso en entornos cambiantes o con la interacción de otros agentes. También se debe garantizar la integridad física y mental de las personas. Los sistemas de IA deben protegerse contra posibles vulnerabilidades que puedan ser explotadas por agentes

malintencionados, como piratas informáticos. Los ataques pueden dirigirse a los datos, el modelo o la infraestructura subyacente. Si un sistema de IA es atacado, puede alterarse su comportamiento o incluso desconectarse. También existe el riesgo de corrupción de datos o procesos inesperados. Por tanto, se deben tomar medidas de seguridad para prevenir y mitigar estos riesgos. Es esencial que los sistemas de IA estén provistos de medidas de protección que posibiliten la implementación de un plan de contingencia en situaciones problemáticas. Esto puede implicar cambiar de un enfoque basado en estadísticas a uno basado en reglas o requerir la intervención de un operador humano (Comisión Europea, 2019, pág. 20).

Es fundamental garantizar que el sistema se comporte de manera esperada y no cause daños. Se deben establecer procesos para evaluar los riesgos asociados al uso de la IA en diferentes áreas. El nivel de seguridad necesario depende del riesgo y las capacidades del sistema. La precisión se refiere a la capacidad de un sistema de IA para realizar juicios correctos, clasificar información adecuadamente, hacer predicciones, recomendar o tomar decisiones correctas basadas en datos o modelos. Esto es especialmente crucial en situaciones que afectan directamente a la vida humana. La reproducibilidad se refiere a que un experimento con IA muestre el mismo comportamiento al repetirse en las mismas condiciones. Esto facilita la descripción precisa de los sistemas de IA por parte de científicos y responsables políticos (Comisión Europea, 2019, pág. 20).

2.6.2.3 Gestión de la privacidad y los datos

La privacidad, un derecho fundamental afectado por los sistemas de IA, requiere una gestión adecuada de los datos. Esto implica garantizar la calidad e integridad de los datos utilizados, su relevancia en relación con el contexto de despliegue de los sistemas de IA, los protocolos de acceso y la capacidad de procesar datos sin violar la privacidad. Los sistemas de IA deben salvaguardar la intimidad y los datos a lo largo de su ciclo de vida. Esto abarca la información inicial proporcionada por el usuario y la información generada durante su interacción con el sistema. Los registros digitales del comportamiento humano pueden revelar preferencias, así como detalles personales como orientación sexual, edad, género y opiniones políticas y religiosas. Es esencial generar confianza al asegurar que la información recopilada no se utilizará para discriminar de manera injusta o ilegal (Comisión Europea, pág. 21).

La calidad de los conjuntos de datos utilizados es crucial ya que pueden contener sesgos, imprecisiones y errores, los cuales deben abordarse antes de la formación del sistema. Además, es necesario garantizar la integridad de los datos, ya que la inclusión de datos malintencionados puede alterar el comportamiento de los sistemas, especialmente en aquellos

con capacidad de autoaprendizaje. En cualquier organización que maneje datos personales, se deben establecer protocolos para regular el acceso a dichos datos, independientemente de si alguien es usuario del sistema o no. Estos protocolos deben definir quién puede acceder a los datos y en qué circunstancias. Solo el personal cualificado y competente que requiera acceder a la información pertinente debe tener acceso a los datos personales (Comisión Europea, 2019, pág. 21).

White House Office of Science and Technology Policy (2022), enfatiza que los individuos deben estar protegidos de prácticas abusivas de datos y tener control sobre cómo se utiliza su información personal. Esto implica asegurar que las protecciones de privacidad estén incorporadas por defecto en los sistemas y que la recopilación de datos cumpla con expectativas razonables y sea limitada a lo estrictamente necesario. Los diseñadores y desarrolladores deben respetar tus decisiones y buscar su permiso en relación con la recopilación, uso, acceso, transferencia y eliminación de tus datos. En caso de no ser posible obtener el consentimiento, se deben emplear salvaguardias alternativas basadas en la privacidad por diseño (pág. 6).

Además, la institución menciona que se debe evitar el uso de decisiones de diseño que oculten la elección del usuario o impongan por defecto opciones invasivas para la privacidad. El consentimiento solo debe usarse cuando se pueda otorgar de manera apropiada y significativa, y las solicitudes de consentimiento deben ser breves, comprensibles y brindarte control sobre la recopilación de datos y su uso específico. Las protecciones y restricciones deben priorizar la privacidad en áreas sensibles, como salud, trabajo, educación, justicia penal, finanzas y datos relacionados con los jóvenes. En estos ámbitos, los datos solo deben utilizarse para funciones necesarias y estar protegidos mediante una revisión ética y prohibiciones de uso (White House Office of Science and Technology Policy, 2022, pág. 6).

2.6.2.4 Transparencia

Este requisito de transparencia está estrechamente relacionado con el principio de explicabilidad y abarca la divulgación de los elementos relevantes de un sistema de IA, como los datos, el sistema mismo y los modelos de negocio. Los procesos y conjuntos de datos que influyen en las decisiones de un sistema de IA, incluyendo la recopilación y etiquetado de datos y los algoritmos utilizados, deben ser documentados de manera exhaustiva para permitir la trazabilidad y aumentar la transparencia. La trazabilidad facilita la auditoría y la explicación, ya que ayuda a identificar las razones detrás de una decisión errónea y contribuye a prevenir futuros errores (Comisión Europea, 2019, pág. 22).

La explicabilidad se refiere a la capacidad de explicar tanto los procesos técnicos de un sistema de IA como las decisiones tomadas por seres humanos. La explicabilidad técnica implica que las decisiones tomadas por el sistema de IA sean comprensibles para las personas y que estas puedan rastrearlas. Cuando un sistema de IA afecta de manera significativa la vida de las personas, es necesario que se pueda solicitar una explicación adecuada sobre el proceso de toma de decisiones del sistema, adaptada al nivel de comprensión de las partes involucradas. Además, se debe proporcionar información sobre cómo el sistema de IA influye en las decisiones de la organización, en el diseño del sistema y en la lógica subyacente a su implementación, lo que garantiza la transparencia del modelo de negocio (Comisión Europea, 2019, pág. 22).

White House Office of Science and Technology Policy (2022), menciona que los implementadores de estos sistemas deben proporcionar documentación accesible en lenguaje claro, que incluya descripciones precisas del funcionamiento general del sistema, así como notificaciones de que se utilizan dichos sistemas, la entidad responsable del sistema y explicaciones de los resultados de manera clara, oportuna y accesible. Esta notificación debe mantenerse actualizada y las personas que utilizan el sistema deben ser informadas de cambios significativos. Deben saber cómo y por qué se determinó un resultado que afecta mediante un sistema automatizado, incluso cuando el sistema automatizado no sea la única fuente que determina el resultado. Estos sistemas deben proporcionar explicaciones técnicamente válidas, significativas y útiles tanto para el individuo como para los operadores u otras personas que necesiten comprender el sistema, y deben ajustarse al nivel de riesgo según el contexto. Se debe hacer público, siempre que sea posible, un informe que incluya información sobre estos sistemas en lenguaje claro y evaluaciones sobre la claridad y calidad de las notificaciones y explicaciones (pág. 6).

Los sistemas de IA no deben presentarse a sí mismos como seres humanos frente a los usuarios, estos tienen el derecho de saber que están interactuando con un sistema de IA. En consecuencia, es necesario que los sistemas de IA sean reconocibles como sistemas de IA. En situaciones que ameriten, se debe otorgar a los usuarios la oportunidad de elegir si desean entablar interacción con un sistema de inteligencia artificial o con un individuo humano genuino, con el propósito de asegurar el cabal cumplimiento de los derechos fundamentales. Además de esto, se debe proporcionar información sobre las capacidades y limitaciones del sistema de IA a profesionales y usuarios finales, de acuerdo con el caso de uso específico, incluyendo detalles sobre la precisión del sistema de IA y sus limitaciones (Comisión Europea, 2019, pág. 22).

2.6.2.5 Diversidad, no discriminación y equidad

Para lograr la confiabilidad de la IA, es esencial asegurar la inclusión y la diversidad en todas las etapas de los sistemas de Inteligencia Artificial. Además de considerar a todas las personas afectadas y garantizar su participación en todo el proceso, también es necesario garantizar la igualdad de acceso mediante enfoques de diseño inclusivos, sin descuidar la igualdad de trato, requisito estrechamente relacionado con el principio de equidad. Los conjuntos de datos utilizados por los sistemas de IA, tanto para su entrenamiento como para su funcionamiento, pueden contener sesgos históricos inadvertidos, vacíos o modelos de gestión incorrectos. Mantener estos sesgos podría resultar en prejuicios y discriminación directa o indirecta hacia ciertos grupos o personas, lo que podría agravar los estereotipos y la exclusión. Cuando sea posible, se deben eliminar los sesgos identificables y discriminatorios en la etapa de recopilación de datos. Asimismo, los métodos empleados en el desarrollo de sistemas de IA, como la programación de algoritmos, pueden generar sesgos injustos en sí mismos. Esto se puede lograr mediante procesos de supervisión que permitan analizar y abordar de manera clara y transparente el propósito, las restricciones, los requisitos y las decisiones del sistema. Contratar a personas con diversos antecedentes, culturas y disciplinas puede garantizar la diversidad de opiniones con el fin de evitar el sesgo (Comisión Europea, 2019, pág. 23).

En el contexto de las relaciones entre empresas y consumidores, los sistemas deben estar centrados en el usuario y diseñados de manera que todas las personas puedan utilizar los productos o servicios de IA, independientemente de su edad, género, capacidades o características. La accesibilidad de esta tecnología para las personas con discapacidad, que están presentes en todos los segmentos de la sociedad, es de particular importancia. Es fundamental que los sistemas de IA sean flexibles y consideren los principios del Diseño Universal, con el objetivo de atender a la mayor cantidad de usuarios posible y cumplir con las normativas de accesibilidad pertinentes. Esto permitirá un acceso equitativo y la participación de todas las personas en las actividades humanas informatizadas existentes y emergentes, así como en lo que respecta a las tecnologías de asistencia. Para desarrollar sistemas de IA confiables, es recomendable consultar regularmente a las partes interesadas que pueden verse afectadas directa o indirectamente por el sistema a lo largo de su ciclo de vida. Es fundamental buscar retroalimentación incluso después de implementar los sistemas de IA y establecer mecanismos para la participación continua de las partes interesadas, como garantizar la disponibilidad de información, consultar e involucrar a los empleados en todas las etapas de implementación de dichos sistemas en las organizaciones (Comisión Europea, 2019, pág. 23).

En complemento a lo expuesto por la Comisión Europea, el marco de trabajo creado por la institución White House Office of Science and Technology Policy (2022), también describe que las protecciones contra la discriminación algorítmica garantizan un uso equitativo de los sistemas y evitan la discriminación basada en características protegidas por la ley, como raza, género, religión y discapacidad. La discriminación algorítmica ocurre cuando se generan tratos injustificados o consecuencias negativas para las personas por parte de los sistemas automatizados. Para prevenir esta discriminación, los diseñadores, desarrolladores e implementadores de sistemas automatizados deben tomar medidas proactivas, incluyendo evaluaciones de equidad, uso de datos representativos y protección contra sesgos demográficos. Además, se requiere una supervisión clara y evaluaciones independientes del impacto algorítmico, cuyos resultados deben hacerse públicos para confirmar estas protecciones.

2.6.2.6 Bienestar social y ambiental

En concordancia con los principios de igualdad y prevención de daños, se debe tener en cuenta a la sociedad en su totalidad, así como a otros seres sensibles y al entorno natural los como actores involucrados a lo largo de todas las etapas del desarrollo y utilización de la IA. Es importante promover la sostenibilidad y la responsabilidad ambiental de los sistemas de IA, así como fomentar la investigación de soluciones de Inteligencia Artificial que aborden preocupaciones a nivel global, como los Objetivos de Desarrollo Sostenible. Es fundamental que la IA se utilice en beneficio de todas las personas, incluyendo a las generaciones futuras. Los sistemas de IA tienen el potencial de abordar problemas sociales urgentes, a pesar de eso, es crucial garantizar que lo hagan de manera respetuosa con el medio ambiente. En este sentido, es necesario evaluar integralmente el proceso de desarrollo, implementación y uso de los sistemas de IA, así como toda la cadena de suministro, a través de un examen crítico del uso de recursos y consumo de energía durante la fase de entrenamiento, priorizando opciones menos perjudiciales. Se deben promover medidas que aseguren el respeto ambiental en todos los eslabones de la cadena de suministro (Comisión Europea, 2019, pág. 23).

La omnipresencia de los sistemas de IA en todas las áreas de nuestra vida social puede influir en nuestra percepción de la acción social y afectar nuestras relaciones y vínculos sociales. Aunque los sistemas de IA tienen el potencial de mejorar las habilidades sociales, también pueden contribuir a su deterioro, lo que puede tener un impacto negativo en el bienestar físico y mental de las personas, por lo tanto, es necesario considerar y monitorear exhaustivamente los efectos de dichos sistemas. Además de evaluar el impacto del desarrollo, implementación y uso de un sistema de IA en las personas, también se deben evaluar sus repercusiones desde una

perspectiva social, teniendo en cuenta sus efectos en las instituciones, la democracia y la sociedad en su conjunto. El uso de sistemas de IA debe ser objeto de un estudio detallado, especialmente en situaciones relacionadas con el proceso democrático, no solo en el ámbito de la toma de decisiones políticas, sino también en contextos electorales (Comisión Europea, 2019, pág. 23).

2.6.2.7 Rendición de cuentas

Además de los requisitos mencionados anteriormente, es importante garantizar la rendición de cuentas, que está estrechamente relacionada con el principio de igualdad. Esto implica establecer mecanismos que aseguren la responsabilidad y la rendición de cuentas de los sistemas de IA y sus resultados, tanto antes como después de su implementación (Comisión Europea, 2019, pág. 24).

La auditabilidad se refiere a la capacidad de evaluar los algoritmos, los datos y los procesos de diseño con el objetivo de contribuir a la confiabilidad de esta tecnología. Es necesario garantizar la capacidad de informar sobre las acciones o decisiones que contribuyen a un resultado específico del sistema, así como responder a las consecuencias de dicho resultado. Identificar, evaluar, notificar y minimizar los posibles efectos negativos de los sistemas de IA es especialmente importante para aquellos que se vean afectados directa o indirectamente por ellos. Las decisiones deben ser fundamentadas, documentadas y revisadas constantemente para garantizar la adaptación del sistema cuando sea necesario. Cuando se presenten consecuencias negativas e injustas, resulta crucial disponer de mecanismos accesibles que garanticen una compensación adecuada. La garantía de que se puede obtener una reparación en caso de que las cosas no salgan como se esperaba es fundamental para generar confianza. Se debe prestar especial atención a las personas o grupos vulnerables (Comisión Europea, 2019, pág. 24).

2.7 Tipos de ciberataques a la Inteligencia Artificial

Los Adversarios en Machine Learning son individuos que buscan por medio de técnicas, engañar a los modelos mediante la introducción de datos engañosos. El objetivo principal es causar un mal funcionamiento en un modelo y se pueden llevar a cabo ataques adversarios mediante la presentación de datos inexactos o modificados durante el entrenamiento del modelo, o mediante la introducción de datos diseñados con malicia para engañar a un modelo ya entrenado. Algunos sistemas que están en producción podrían ser vulnerables a ataques. Por ejemplo, investigadores demostraron que, al colocar pequeñas pegatinas en el suelo, podían hacer que un automóvil autónomo se desplace hacia el carril de tráfico contrario. Otros estudios han

demostrado que, al realizar cambios imperceptibles en una imagen, se puede engañar a un sistema de análisis médico para que clasifique un lunar benigno como maligno, o que trozos de cinta puedan engañar a un sistema de visión por computadora para que clasifique erróneamente una señal de alto como una señal de límite de velocidad (Wiggers, 2021).

Existen diversos tipos de ataques adversarios dirigidos a los modelos de IA, que se pueden clasificar en tres categorías principales: influencia en el clasificador, violación de seguridad y ataques dirigidos. Estos ataques pueden ser subcategorizados como de caja blanca o caja negra. En los ataques de caja blanca, el atacante cuenta con acceso a los parámetros del modelo, mientras que, en los ataques de caja negra, el atacante no dispone de dicho acceso a los parámetros. Un ataque puede influir en el modelo, interrumpiendo su capacidad de realizar predicciones. Por otro lado, una violación de seguridad implica el suministro de datos maliciosos que se clasifican como legítimos. Por último, un ataque dirigido busca permitir una intrusión o interrupción específica, o crear un caos general (Wiggers, 2021).

Según lo expuesto por el autor Dar (2023), los conjuntos de datos utilizados en el entrenamiento de modelos de Deep Learning pueden ser comprometidos mediante un tipo de ataque adversario conocido como envenenamiento de datos. En este ataque, la información maliciosa es introducida en los datos recopilados para el entrenamiento. Los ataques de envenenamiento de datos son simples y prácticos de realizar, y solo se requieren habilidades técnicas limitadas. Estos ataques permiten a los actores maliciosos agravar sesgos o insertar una puerta trasera en el modelo para controlar su comportamiento posterior al entrenamiento.

Existen dos tipos de ataques de envenenamiento que han sido demostrados. El primero, conocido como envenenamiento de vista dividida, se aprovecha de las diferencias entre los datos observados durante el tiempo de revisión y los datos utilizados en el entrenamiento del modelo. El segundo, denominado ataque frontal, involucra la modificación de artículos en sitios web antes de que se realicen las copias de seguridad periódicas de su contenido. Estos ataques pueden influir en el modelo de IA, incluso si solo afectan a una pequeña porción de los datos. Aunque hasta el momento no se han detectado casos de envenenamiento de datos, existe un riesgo potencial debido a los incentivos económicos y la posibilidad de manipular modelos de texto en aplicaciones como los motores de búsqueda (Dar, 2023).

Wiggers (2021) afirma que los ataques de evasión son los más frecuentes, en los cuales los datos se modifican para evadir la detección o para clasificarlos como legítimos. Estos ataques no implican influir en los datos utilizados para entrenar un modelo, sin embargo, se asemejan a la manera en que los spammers y hackers ocultan el contenido de los correos no deseados y el malware. Un ejemplo de evasión es el spam basado en imágenes, donde el contenido del spam

se incrusta en una imagen adjunta para evadir el análisis de los modelos antispam. Otro ejemplo son los ataques de suplantación de identidad contra los sistemas de verificación biométrica impulsados por IA.

Por otro lado, el robo de modelos, también conocido como extracción de modelos, implica que un adversario intente reconstruir un sistema de aprendizaje automático de caja negra o extraer los datos en los que se entrenó el modelo. Esto puede generar problemas cuando los datos de entrenamiento o el propio modelo son sensibles y confidenciales. Por ejemplo, el robo de modelos podría utilizarse para extraer un modelo de comercio de acciones patentado, que el adversario podría emplear en su propio beneficio financiero (Wiggers, 2021).

2.8 Confianza en Inteligencia Artificial

La confianza en la Inteligencia Artificial es un proceso mental y físico que tiene en cuenta las características de los sistemas basados en IA y cómo interactúan con otros sistemas similares. Es importante destacar que esta definición no se centra en aspectos como la bondad o integridad, lo que permite investigar los estados mentales relevantes en la confianza. En lugar de eso, se enfoca en comprender cómo surge, evoluciona y desaparece la confianza, así como los mecanismos psicológicos subyacentes (Lukyanenko et al., 2022, pág. 16).

La investigación sobre cómo la confianza en la interacción entre humanos y sistemas de IA afecta la confianza en una IA específica, así como la transferencia de confianza de un nivel más general a uno más específico, es limitada. Es necesario llevar a cabo más investigaciones para comprender los éxitos y fracasos de la IA, así como las preocupaciones relacionadas con su invasividad y las repercusiones que puede tener. En resumen, la definición de confianza humana en la IA establece que es un requisito previo para interactuar con esta tecnología, lo que nos permite explorar las razones que generan confianza en la IA y las diferentes variables que dependen de la confianza humana en la IA (Lukyanenko et al., 2022, pág. 16).

2.9 Marco de trabajo AI TRiSM: AI Trust, Risk, and Security Management

Los sistemas de Inteligencia Artificial son susceptibles a ser atacados por ciberdelincuentes. Esto implica que los delincuentes pueden aprovecharse de los modelos de IA para llevar a cabo actividades maliciosas automatizadas, como ataques de software malicioso, violaciones de datos y estafas de suplantación de identidad (Siddiqui, 2023).

Es en este contexto donde resulta indispensable el uso de la IA TRiSM, ya que permite a las empresas utilizar modelos de IA de manera segura. El enfoque de TRiSM se basa en técnicas que establecen una base sólida para la seguridad de los modelos de IA. Mediante la

implementación de medidas como el cifrado de datos, el almacenamiento seguro de la información y la autenticación multifactorial, TRiSM garantiza la obtención de resultados precisos por parte de los modelos de IA.

En palabras de la autora Perri del grupo Gartner (2022), menciona que la Declaración de Derechos de la Inteligencia Artificial, explorado en el apartado 2.6.2 de este trabajo, es un modelo organizativo que busca proteger a la sociedad de los posibles daños que podría causar la IA. Este enfoque insta a desarrolladores y usuarios de modelos de IA a establecer garantías en sus estrategias y modelos de IA, con el objetivo de minimizar los riesgos y maximizar los beneficios de esta tecnología. Con el fin de alcanzar estos objetivos, se ha creado un marco de trabajo riguroso conocido como AI TRiSM.

Según la definición de Gartner (2022), este marco de trabajo proporciona una estructura que respalda distintos aspectos fundamentales. En primer lugar, la gobernanza es uno de los pilares principales. Su propósito radica en asegurar que la inteligencia artificial sea empleada de forma responsable y ética. Esto implica establecer políticas y directrices claras para su implementación y asegurar que se cumplan en todas las etapas del proceso. La confianza es otro aspecto crucial contemplado por el marco AI TRiSM. Se busca asegurar que los sistemas de IA sean confiables, justos y precisos. Esto implica verificar y evaluar la calidad de los datos utilizados, así como los algoritmos y modelos empleados en la construcción de dichos sistemas. Asimismo, la seguridad es una preocupación esencial en el contexto de la IA. El marco AI TRiSM se encarga de proteger los sistemas de IA de posibles ciberataques. Esto implica implementar medidas de seguridad robustas, como el cifrado de datos y la detección de intrusiones, para garantizar la integridad y confidencialidad de la información procesada por los sistemas de IA. Por último, la privacidad es un aspecto crítico para considerar. El marco AI TRiSM se enfoca en asegurar que los sistemas de IA no violen la privacidad de los individuos. Esto implica implementar salvaguardias y mecanismos que protejan la información personal y se adhieran a las regulaciones y políticas de privacidad aplicables.

Perri (2022) destaca que esta estructura incluye soluciones, técnicas y procesos para la interpretabilidad, explicabilidad y operaciones del modelo, así como para la privacidad y la resistencia a ataques adversarios para los clientes y la empresa. Su propósito es ayudar a generar confianza con los usuarios y las partes interesadas, y hacer más probable que adopten y utilicen los sistemas de IA. Ayuda a reducir la exposición al riesgo, identificando y mitigando los riesgos asociados con la IA. Esto puede ayudar a proteger a las organizaciones de ciberataques, violaciones de datos y otros riesgos potenciales. Por último, mejora del cumplimiento ayudando

a las organizaciones a llevar a cabo las regulaciones de privacidad de datos y otras leyes y regulaciones. Esto puede ayudar a proteger a las organizaciones de multas y otras sanciones.

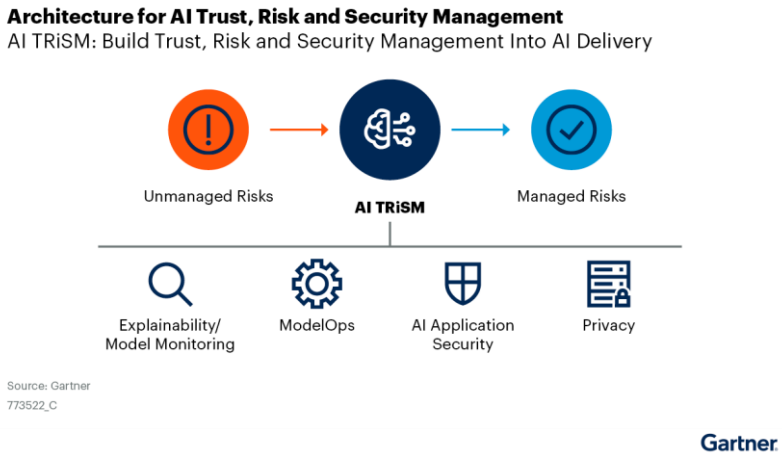
Al proporcionar una plataforma segura para la IA, las empresas pueden centrarse en utilizar estos modelos para impulsar el crecimiento, aumentar la eficiencia y crear mejores experiencias para el cliente. Por ejemplo, la IA TRiSM proporciona una manera automatizada de analizar los datos de los clientes, lo que permite a las empresas identificar rápidamente tendencias y oportunidades para mejorar sus productos y servicios (Siddiqui, 2023). De esta manera, se busca establecer un marco ético y responsable para el desarrollo y uso de la Inteligencia Artificial, asegurando que se tenga en cuenta el bienestar de la sociedad y se minimicen los riesgos asociados con esta tecnología (Perri, 2022).

2.9.1 Objetivo de AI TRiSM

El objetivo del framework AI TRiSM es habilitar la gestión de confianza, riesgo y seguridad, con la capacidad de anticipar mejores resultados empresariales para proyectos de IA. (Kaur, 2023). En la siguiente figura, se destacan los tres marcos principales que se unen para asegurar que los proyectos de IA sean confiables, seguros y generen mejores resultados, junto a los pilares esenciales:

Figura 6

AI TRiSM: Marco para administrar la confianza, el riesgo y la seguridad de la IA



Nota. AI TRiSM se logra mediante la aplicación de prácticas y metodologías interdisciplinarias a modelos de IA y datos de respaldo, y la adopción de conjuntos de herramientas que respalden estas prácticas. Adaptado de *Market Guide for AI Trust, Risk and Security Management* de Gartner, 2023, https://www.gartner.com/doc/reprints?id=1-2CQX56DW&ct=230303&st=sb

2.9.2 Los marcos de Confianza, Riesgo y Seguridad de la IA

El primer marco es el AI Trust o también conocido como Confianza en la IA, este se asocia con la transparencia o explicabilidad, es decir, la capacidad de identificar si el modelo logró los resultados deseados con los pasos correspondientes. Esto contribuye a generar confianza y promover la transparencia (Kaur, 2023).

AI Risk, o Riesgo en la IA aplica una gobernanza precisa y estricta en la gestión de los riesgos de la IA empresarial. Registrando y gestionando las etapas de desarrollo y proceso de los modelos y verificando todas las partes del proceso de lanzamiento para comprobar la integridad y el cumplimiento (Kaur, 2023).

AI Security Management, o Gestión de seguridad en la IA tiene como propósito garantizar la seguridad en cada etapa del proceso en las operaciones del modelo de Machine Learning. La gestión de seguridad en la IA es capaz de acceder a toda la tubería del modelo, identificar anomalías, automatizar pipeline CI/CD y escanear vulnerabilidades (Kaur, 2023).

2.9.3 Los pilares de AI TRiSM

AI TRiSM tiene como objetivo optimizar la adopción de la IA para lograr una mejor confiabilidad, mejorar las funcionalidades y mantener la integridad del valor del sistema de IA en producción. A continuación, se describen los pilares de AI TRiSM.

2.9.3.1 Explicabilidad

La Explicabilidad en Inteligencia Artificial se describe como un conjunto de habilidades que ofrecen información detallada o justificaciones para aclarar el funcionamiento de un modelo ante un público determinado. Pone énfasis en respaldar la confiabilidad y previsibilidad del modelo. En primer lugar, permite describir las características y propiedades de un modelo, lo cual resulta crucial para comprender su estructura y los procesos internos que lo sustentan. Además, permite resaltar tanto los aspectos positivos como los negativos de un modelo, brindando una visión completa y equilibrada de sus fortalezas y debilidades. La explicabilidad también desempeña un papel importante en la predicción del comportamiento probable de un modelo. Al comprender cómo se generan los resultados y qué factores influyen en ellos, se puede anticipar su impacto en diferentes situaciones y tomar decisiones bien fundamentadas. Asimismo, ayuda a identificar posibles sesgos o predisposiciones en un modelo. Al analizar detalladamente los datos de entrada y los procesos de toma de decisiones del modelo, es posible detectar y abordar posibles discriminaciones o desigualdades en los resultados. Otro aspecto relevante de la explicabilidad es su capacidad para aclarar el funcionamiento de un modelo. Al

comprender cómo se procesan los datos y se generan las predicciones, se facilita la búsqueda de precisión, imparcialidad, responsabilidad, estabilidad y transparencia en la toma de decisiones basadas en algoritmos. En la actualidad, no existen herramientas ni métodos convencionales para alcanzar la explicabilidad en Inteligencia Artificial. No obstante, tanto en el ámbito académico como en el sector privado, se están llevando a cabo iniciativas prometedoras con el fin de progresar hacia este propósito (Litan et al., 2023).

2.9.3.2 Monitoreo del modelo

En línea con lo expuesto anteriormente por el mismo autor, describe el Monitoreo en IA refiriéndose a la supervisión y análisis de los datos de producción con el objetivo de lograr un rendimiento óptimo del sistema y garantizar la seguridad de las organizaciones ante posibles ataques maliciosos, desviaciones, sesgos y errores en los datos y procesos. Además, se busca establecer una conexión fluida entre el monitoreo del modelo y la explicabilidad en Inteligencia Artificial, para asegurar un funcionamiento confiable y transparente.

Las principales funciones del Monitoreo del modelo abarcan diversas áreas. En primer lugar, se busca evaluar las posibles desviaciones en los datos y asegurar la equidad en las distribuciones del modelo durante los procesos de razonamiento y producción. Esto implica analizar cuidadosamente los resultados obtenidos y verificar que no existan sesgos o disparidades indeseables en los resultados generados por el modelo. Asimismo, el monitoreo del modelo tiene como objetivo detectar cualquier filtración de datos que pueda comprometer la calidad y la confiabilidad de los resultados. La identificación anticipada de fugas permite tomar medidas correctivas y mantener la integridad de los datos utilizados en el modelo (Litan et al., 2023).

Otra función importante del Monitoreo del modelo es la identificación del envenenamiento de datos, es decir, la presencia de datos maliciosos o incorrectos que puedan influir negativamente en el rendimiento del modelo. Mediante un monitoreo constante, se busca detectar y mitigar este tipo de amenazas, garantizando la integridad de los datos y la precisión de los resultados. Además, el monitoreo del modelo se encarga de verificar el cumplimiento en el uso de los datos, asegurando que se respeten los acuerdos y las políticas establecidas en cuanto a su utilización. Esto implica revisar de manera periódica y sistemática el manejo de los datos, garantizando su correcta manipulación y protección. Por último, una función clave del monitoreo del modelo es medir los cambios en la distribución de los datos. Esto implica capturar tanto las alteraciones en la probabilidad previa como las fluctuaciones en las variables predictivas que pueden impactar en el rendimiento y la precisión del modelo. Al detectar estos cambios, se pueden tomar acciones correctivas y asegurar la actualización y ajuste adecuado del modelo. De

esta manera, el monitoreo del modelo se convierte en una herramienta fundamental para garantizar un rendimiento óptimo y una toma de decisiones fundamentada en los sistemas de Inteligencia Artificial (Litan et al., 2023).

2.9.3.3 ModelOps

El pilar de ModelOps descripto por se centra en la gestión integral y el ciclo de vida de todos los modelos analíticos, modelos basados en Machine Learning, grafos de conocimiento, reglas, optimización, lingüística y agentes. En la actualidad, la mayoría de los proveedores de ModelOps ofrecen funcionalidades que van desde la implementación hasta la producción de los modelos (Litan et al., 2023).

Según los resultados de la Encuesta sobre IA en Organizaciones de Gartner 2021, las principales barreras para la implementación de la IA se deben a la dificultad para medir su valor y la falta de comprensión de sus beneficios y usos. ModelOps ayuda a superar estas barreras al proporcionar soporte para diferentes tipos de metricas de rendimiento de los modelos. Algunas organizaciones consolidadas en el ámbito de la IA están adaptando su estrategia de ModelOps hacia la llamada IA compuesta, la cual implica la combinación de diferentes técnicas de IA para mejorar la eficiencia del aprendizaje y ampliar el nivel de representaciones de conocimiento (Litan et al., 2023).

ModelOps permite la operacionalización y gestión eficiente de los modelos de Inteligencia Artificial. En primer lugar, la gestión de dependencias y recursos desempeñan un papel crucial en el proceso de operacionalización del modelo. Esto implica asegurar que todas las dependencias necesarias para el funcionamiento del modelo estén correctamente gestionadas, así como la catalogación de modelos con un inventario detallado y la evaluación de riesgos asociados. La integración en sistemas busca establecer una integración fluida del modelo en los sistemas existentes, permitiendo su interoperabilidad y sinergia con otras soluciones tecnológicas. La catalogación de modelos implica un registro organizado y estructurado de los modelos disponibles, así como su evaluación continua para asegurar su calidad y rendimiento (Litan et al., 2023).

El monitoreo del rendimiento del modelo es otra capacidad crítica de ModelOps. Mediante un seguimiento constante, se evalúa el rendimiento del modelo en tiempo real, se realizan reentrenamientos cuando sea necesario y se realiza un seguimiento de los indicadores clave de rendimiento para garantizar la calidad y la eficacia del modelo en su desempeño operativo. La automatización es un componente clave de ModelOps, permitiendo una gestión eficiente, explicación y gobernanza de los modelos. Esto incluye la automatización de tareas

como pruebas, revisiones, administración y explicación del modelo, lo que garantiza una implementación y gestión más eficiente y transparente (Litan et al., 2023).

Antes de implementar en producción, se llevan a cabo pruebas exhaustivas para evaluar el rendimiento del modelo en diferentes escenarios y situaciones. Estas pruebas previas a la producción son fundamentales para identificar posibles problemas y asegurar que el modelo cumpla con los estándares y requisitos establecidos. La ejecución del modelo en diferentes entornos técnicos y la gestión de su ciclo de vida implica la adaptabilidad del modelo para funcionar en diferentes entornos, la gestión adecuada de las actualizaciones y modificaciones a lo largo de su ciclo de vida. La reproducibilidad del modelo busca capturar la línea de tiempo y los metadatos asociados al modelo, lo que permite una comprensión clara y precisa de su evolución y facilita la reproducción de resultados en diferentes momentos y contextos. El monitoreo de la infraestructura técnica es constante para garantizar una implementación eficiente y escalable del modelo, asegurando así su disponibilidad y rendimiento óptimo (Litan et al., 2023).

También, se proporcionan librerías de modelos y plantillas predefinidas como parte de ModelOps, lo que permite una implementación rápida y transparente de los modelos. Estas librerías facilitan el acceso a modelos preentrenados y plantillas estandarizadas, acelerando el proceso de implementación y reduciendo la complejidad asociada. Además, se establecen alertas personalizadas y notificaciones para un seguimiento y resolución eficientes de incidencias relacionadas con los modelos. Esto permite una respuesta rápida y eficaz ante posibles problemas, garantizando la disponibilidad y el rendimiento óptimo del modelo en todo momento. Por último, ModelOps incluye funciones de seguridad integradas, como autenticación, control de acceso y registro de auditoría. Estas funciones aseguran la protección y la integridad de los modelos, así como el cumplimiento de las normativas y políticas de seguridad establecidas (Litan et al., 2023).

2.9.3.4 Seguridad de Aplicaciones de IA

La Seguridad de Aplicaciones de IA, se refiere a las prácticas, políticas y herramientas diseñadas para proteger las aplicaciones de IA contra ataques especializados. Esto incluye la detección y prevención de amenazas, así como la evaluación de la resistencia de los modelos frente a ataques adversarios (Litan et al., 2023).

La seguridad de aplicaciones de IA abarca diferentes etapas, desde el desarrollo y validación hasta la implementación y pruebas finales en tiempo real. La resistencia a ataques adversarios es un elemento central de la seguridad de aplicaciones de IA. Esto implica enseñar a

los modelos a ignorar o responder de manera diferente ante entradas adversarias durante su desarrollo, pruebas y producción. Se pueden utilizar técnicas de entrenamiento o herramientas específicas de IA para fortalecer los modelos y permitirles tolerar ciertos niveles de ruido y datos adversarios sin que su rendimiento se vea demasiado afectado. Los ataques adversarios modifican las salidas de los algoritmos de Machine Learning para beneficiar al atacante. Esto se logra mediante la introducción de entradas maliciosas o perturbaciones en los datos una vez que el modelo está en producción. Algunas herramientas de resistencia a ataques adversarios pueden resaltar la vulnerabilidad de los modelos frente a estas entradas maliciosas y proporcionar entrenamiento para mitigarlas. Es importante contar con métodos y herramientas para generar muestras adversarias y comprender las capacidades clave en la seguridad de aplicaciones de IA. Esto ayudará a proteger las aplicaciones de IA de amenazas especializadas y garantizar su confiabilidad (Litan et al., 2023).

En el ámbito de la defensa adversarial, se emplean diversas técnicas y métodos para proteger los sistemas de Inteligencia Artificial contra posibles ataques o manipulaciones. Estas estrategias se enfocan en detectar y contrarrestar amenazas externas, asegurar la integridad de los datos y modelos, y generar alertas con explicaciones claras. Una de las herramientas clave en la defensa adversarial es la detección de envenenamiento externo, que busca identificar intentos de manipulación de datos o modelos por parte de actores malintencionados. Además, se implementan medidas de protección de datos y modelos para prevenir accesos no autorizados y garantizar la confidencialidad de la información.

La generación de muestras adversarias es otra estrategia empleada en el ámbito de la defensa adversarial. Consiste en crear ejemplos de datos especialmente diseñados para engañar al modelo y revelar posibles vulnerabilidades. Esta práctica permite mejorar la robustez del modelo al exponerlo a situaciones desafiantes durante el entrenamiento. Asimismo, se recurre al uso de redes generativas adversarias para entrenar el modelo utilizando muestras o datos adversarios generados. Esto ayuda al modelo a tolerar el ruido y las perturbaciones durante la producción, fortaleciendo su capacidad de respuesta ante posibles ataques.

Además de estas técnicas, se trabaja en mejorar el almacenamiento del modelo y establecer protocolos para su retiro inmediato en caso de identificar posibles actores maliciosos. Esto se logra mediante la detección temprana de anomalías y la generación de alertas que permiten a las organizaciones protegerse de riesgos financieros y salvaguardar su reputación (Litan et al., 2023).

2.9.3.5 Privacidad

El pilar de la Privacidad es crucial en los proyectos de Inteligencia Artificial debido a las restricciones y riesgos de cumplimiento normativo asociados al acceso a datos personales. Las leyes de protección de datos están en constante crecimiento y buscan minimizar las violaciones de privacidad. Se recomienda utilizar datos sintéticos en lugar de información identificable durante las etapas iniciales del entrenamiento del modelo para abordar este problema. Además, el control del riesgo de reidentificación puede lograrse mediante el uso de datos sintéticos generados por redes generativas adversarias u otras técnicas de IA generativa. El uso de datos sintéticos también puede ayudar a equilibrar sesgos y artefactos en el modelo y puede combinarse con enfoques distribuidos de Machine Learning (Litan et al., 2023).

Los controles de privacidad en la IA no incluyen las capacidades convencionales de protección de datos, ya que se espera que las organizaciones ya hayan implementado estos controles de seguridad estándar. Se deben seleccionar las medidas de mitigación adecuadas según el apetito de riesgo de la organización, la robustez percibida de la tecnología y las características del caso de uso. Existen diversas metodologías y enfoques para acceder y compartir información sin utilizar datos identificables, cumpliendo así con los requisitos de privacidad y protección de datos. No existe una solución única para abordar todos los problemas de privacidad (Litan et al., 2023).

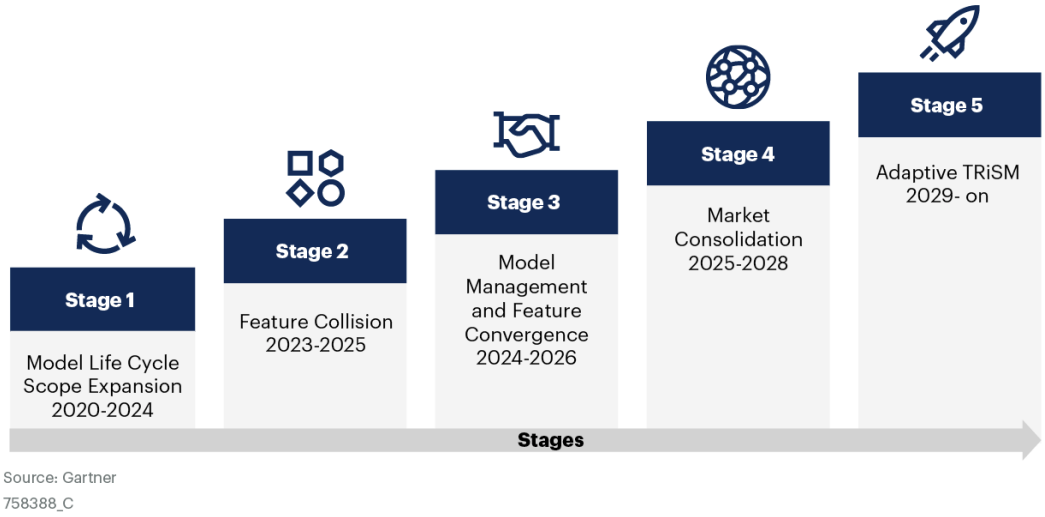
2.9.4 Progreso de AI TRiSM

La Guía de Mercado de AI TRiSM de Gartner desarrollada por Litan et at. (2023), predice que para el año 2026, las organizaciones que operacionalicen la transparencia, confianza y seguridad de la IA verán una mejora del 50% en la adopción, metas empresariales y aceptación del usuario en sus modelos de IA. Para el año 2027, al menos dos proveedores que ofrezcan funcionalidades de gestión de riesgos de IA van a ser adquiridos por proveedores de gestión de riesgos empresariales que brinden una funcionalidad más amplia. Tambien al menos una empresa global verá prohibida la implementación de su IA por un regulador debido al incumplimiento de la legislación de protección de datos o gobernanza de IA. Se espera que el mercado evolucione y crezca de manera constante, impulsado en parte por regulaciones crecientes y requisitos para operacionalizar y mejorar el rendimiento de los modelos de IA en proliferación.

La Figura 7 muestra las cinco fases de progreso del mercado de AI TRiSM según Gartner (2023), las cuales se superponen en gran medida, mientras el mercado se organiza y establece.

Figura 7

Fases de progreso del mercado de AI TRiSM



Nota. La figura presenta las cinco fases del ciclo de vida del modelo y su evolución en el mercado de AI TRiSM, desde la expansión del alcance en la Fase 1 hasta la consolidación del mercado en la Etapa 4 y la aparición de sistemas de AI TRiSM mejorados por IA en la Etapa 5. Adaptado de *Market Guide for AI Trust, Risk and Security Management* de Gartner, 2023, https://www.gartner.com/doc/reprints?id=1-2CQX56DW&ct=230303&st=sb

En la Fase 1, denominada Ampliación del Alcance del Ciclo de Vida del Modelo y que abarca el período de 2020 a 2024, los autores Litan et al. (2023) mencionan que se destaca la colaboración entre científicos de datos, desarrolladores de IA, líderes de seguridad y actores comerciales. Durante esta etapa, se esta trabajando de manera conjunta para garantizar la comprensión de los beneficios de las herramientas de TRiSM y su integración desde el inicio del diseño hasta la producción. Además, se hace uso de las herramientas de AI TRiSM para validar y probar modelos de terceros, asegurando su correcto funcionamiento según lo anunciado y requerido.

La Fase 2, conocida como Colisión de Funciones y que tiene lugar entre 2023 y 2025, se caracteriza por la superposición de distintos elementos clave. ModelOps, explicabilidad, monitoreo del modelo, pruebas continuas, privacidad y seguridad de aplicaciones de IA se entrelazan entre sí. Durante este período, estos módulos y capacidades se entrecruzan, generando un solapamiento en su implementación y desarrollo (Litan et al., 2023).

En la Fase 3, denominada Gestión del Modelo y Convergencia de Funciones abarcando el período de 2024 a 2026, los autores previamente mencionados, predicen que se enfocará en la

gestión integral del modelo. A medida que más científicos de datos, desarrolladores de IA y usuarios comerciales adaptan las herramientas de AI TRiSM, los proveedores de ModelOps se alinearán para brindar soporte a lo largo de todas las etapas del desarrollo del modelo. Esta adaptación implica la incorporación de capacidades como el inventariado, la explicabilidad, el monitoreo, las pruebas y otras funciones útiles de AI TRiSM.

La Etapa 4, denominada Consolidación del Mercado y que tendrá lugar entre 2025 y 2028, se caracteriza por una continua expansión de las funciones en los pilares del mercado. Durante este período, el mercado se consolidará en torno a ModelOps y las funciones de privacidad. ModelOps respalda la IA compuesta, la explicabilidad y monitoreo, las pruebas y las herramientas de seguridad de aplicaciones de IA. Además, se integran nuevas funciones de gestión de riesgos de IA en la gestión de riesgos empresariales, estableciendo una relación estrecha entre ambos aspectos (Litan et al., 2023).

Finalmente, la Etapa 5, denominada TRiSM Mejorado por IA y que se iniciará a partir de 2029, se caracteriza por la emergencia de sistemas empresariales de TRiSM completamente gestionados de extremo a extremo. Estos sistemas utilizan la propia IA para corregirse bajo supervisión humana. En esta etapa, las capacidades de TRiSM se incorporan a las disciplinas más amplias de ingeniería y gobernanza de IA, estableciendo un vínculo integral entre ambos campos (Litan et al., 2023).

Capítulo 3 – Desarrollo

3.1 Introducción

En el capítulo anterior, se desarrolló el marco teórico abordando diversos aspectos relacionados con la Inteligencia Artificial. Se exploraron los antecedentes históricos de la IA, pasando por diferentes etapas y avances significativos en el campo. Además, se analizaron los beneficios y riesgos asociados a la IA, así como los diferentes tipos de IA y las áreas de aplicación. En cuanto a los fundamentos de una IA fiable, se examinaron los principios éticos y los requisitos necesarios para garantizar la confiabilidad de los sistemas de IA. Entre estos requisitos se encuentran la necesidad de contar con la intervención y supervisión humanas, la robustez técnica y la seguridad, la gestión adecuada de la privacidad y los datos, la transparencia, la no discriminación y la equidad, entre otros aspectos.

Se exploraron también los tipos de ciberataques a la IA y se definió el concepto de confianza en la IA. Además, se presentó el marco de trabajo AI TRiSM, cuyo objetivo es abordar la confianza, el riesgo y la seguridad en la IA. Se detallaron los pilares fundamentales de AI TRiSM, como la explicabilidad, el monitoreo del modelo, ModelOps, la seguridad de aplicaciones de IA y la privacidad. Con base en este fundamento teórico, en este capítulo se abordarán las recomendaciones y estrategias para mejorar la seguridad en los modelos de IA. Además, se examinarán casos de uso para verificar la efectividad de este enfoque, siguiendo las directrices de AI TRiSM.

3.2 Recomendaciones principales

Antes de adentrarse por completo, los autores Litan et al. del grupo Gartner (2023), proveen algunas recomendaciones para la correcta aplicación del marco de trabajo AI TRiSM. En primer lugar, se sugiere asignar roles y responsabilidades a personas específicas dentro de la organización para que sean responsables de gestionar la confianza, el riesgo y la seguridad de la IA. Esta sugerencia podría resultar insuficiente si no se define claramente el alcance de sus funciones y si no se establecen canales efectivos de colaboración con todas las unidades organizativas involucradas. Es imprescindible establecer objetivos e intenciones comunes definiendo los controles de la IA, trabajando en conjunto con diferentes áreas para abarcar los temas de privacidad, cumplimiento, seguridad y legal. Esta colaboración permitirá una evaluación más completa y precisa de los riesgos. Sin embargo, esto plantea interrogantes sobre cómo se definiría claramente el alcance de sus funciones y cómo se establecerían canales efectivos de colaboración con todas las unidades organizativas involucradas.

Los autores Litan et al. (2023) también mencionan que es de vital importancia, documentar e implementar las intenciones del modelo. Se deben utilizar herramientas adecuadas para documentar y poner en práctica las intenciones colectivas establecidas previamente para el modelo de IA. Estas herramientas pueden tomar la forma de una guía o documento que describa los lineamientos y principios éticos que se deben seguir al desarrollar e implementar el modelo. Es deseable clasificar y utilizar soluciones PET (Privacidad, Ética y Transparencia), que implican evaluar y utilizar soluciones ya disponibles que se centren en los aspectos de privacidad, ética y transparencia, alineándose con los requisitos del proyecto de IA. Estas soluciones pueden ayudar a prevenir problemas de cumplimiento normativo y respaldar resultados exitosos y éticos. No obstante, es esencial que estas intenciones se traduzcan en acciones concretas y no se queden solo en un documento teórico.

Como bien mencionan los autores previamente citados, la evaluación de riesgos debe ser integrada en los procesos de desarrollo de IA como una etapa clave en el ciclo de vida del desarrollo de IA. Esto garantizará que los riesgos se aborden desde las primeras etapas del proceso y se tengan en cuenta en cada fase del desarrollo. Aunque esto suena razonable en la teoría, la implementación de evaluaciones periódicas de los riesgos de la IA puede ser complicada. Además, los cambios rápidos en el entorno tecnológico y normativo plantean desafíos adicionales para mantener actualizadas las medidas implementadas, lo que pone en duda la viabilidad de esta recomendación en la realidad.

Litan et al. (2023) indican que iniciar un programa de seguridad de aplicaciones de IA es imperativo en este ámbito, con el propósito de realizar un análisis exhaustivo de la seguridad de las aplicaciones de IA, incluyendo tanto las desarrolladas internamente como las que contengan IA de terceros. Es fundamental implementar medidas de mitigación de amenazas y asegurarse de que existan compromisos de seguridad previos a la implementación de la IA en los entornos correspondientes. Para aplicar estas acciones de manera efectiva, se sugiere promover la conciencia y la formación, brindando capacitación a los responsables de la implementación de IA, así como a los equipos involucrados, sobre la importancia de la evaluación inicial de riesgos y las mejores prácticas asociadas.

Es importante tener en cuenta que la seguridad de la IA es un campo en constante evolución, y las amenazas y vulnerabilidades asociadas también evolucionan rápidamente. Por lo tanto, es necesario adoptar un enfoque holístico y proactivo para mantenerse al día con las últimas medidas de mitigación de amenazas y asegurar la continuidad de la seguridad a lo largo del tiempo.

3.3 Recomendaciones complementarias

Con el objetivo de facilitar la aplicación de las recomendaciones de Litan et at. (2023), sugerimos tomar un enfoque ágil de gestión de proyectos que ha demostrado ser altamente efectivo en el mercado actual. Según Schwaber y Sutherland (2020), Scrum es un marco de trabajo ágil que permite a los equipos desarrollar, entregar y mantener productos de manera iterativa e incremental. Se basa en principios de transparencia, inspección y adaptación, fomentando la colaboración y la autoorganización del equipo para maximizar el valor entregado al cliente. Scrum proporciona una estructura definida de roles, eventos y artefactos que ayudan a los equipos a gestionar de manera efectiva la complejidad y a responder rápidamente a los cambios en los requisitos y en el entorno de desarrollo. A continuación, en la Tabla 1, proporcionamos nuestras sugerencias complementarias a las recomendaciones de los autores anteriormente citados, para llevar a cabo la implementación de AI TRiSM.

Tabla 1

Sugerencias a las recomendaciones de Litan et at. con un enfoque ágil

Acciones	**Descripción**
Asignar roles y responsabilidades	Establecer un equipo multidisciplinario que sea responsable de gestionar la confianza, el riesgo y la seguridad de la IA. El rol de Product Owner será el responsable de priorizar las tareas. El Scrum Master actuará como el facilitador de la colaboración entre las unidades organizativas involucradas. Los miembros técnicos que formarán el equipo de desarrollo ejecutarán las tareas a realizar. Comunicar claramente las responsabilidades a cada rol.
Establecer canales de comunicación	Crear mecanismos de comunicación fluida y colaboración entre todo el equipo involucrado. Establecer reuniones periódicas en base a los principios de Scrum.
Establecer objetivos	Facilitar la colaboración entre diferentes áreas para establecer metas claras en términos de privacidad, cumplimiento normativo, seguridad y legalidad. Promover la identificación y mitigación del sesgo y la equidad.

Documentar y poner en práctica	Utilizar herramientas adecuadas para documentar y compartir las intenciones colectivas establecidas previamente. Implementar soluciones PET que se alineen con los requisitos del proyecto de IA.
Integrar la evaluación de riesgos	Utilizar las ceremonias de Scrum para realizar revisiones periódicas para evaluar y actualizar los riesgos asociados a la IA. Realizar ajustes necesarios en función de los cambios en el entorno tecnológico y normativo.
Iterar	Mantener el backlog actualizado para realizar un análisis exhaustivo de vulnerabilidades y amenazas. Priorizar medidas de mitigación de amenazas según su impacto y probabilidad de ocurrencia. Promover conciencia y formación en seguridad de IA mediante capacitaciones regulares para el equipo y responsables de la implementación.

Nota. La tabla proporcionada resume los pasos complementarios sugeridos para aplicar las recomendaciones de Litan et al. (2023) utilizando un marco de trabajo ágil, como Scrum.

Es importante destacar que estos pasos pueden ser adaptados y ajustados según las necesidades, teniendo en cuenta que el mundo de la IA está en constante evolución, lo que dificulta la comprensión completa de sus implicaciones en términos de seguridad. La metodología ágil proporciona flexibilidad y permite realizar ajustes continuos para asegurar la generación de valor. En la siguiente sección, se detallarán las fases de aplicación para aplicar AI TRiSM, que denominaremos el backlog dentro del marco de trabajo Scrum.

3.4 Implementando Explicabilidad y Monitoreo

En esta sección, nos basamos en las premisas planteadas por Litan et al. (2023) con el objetivo de abordar la evaluación de la explicabilidad en los modelos de IA, con el objetivo de generar confianza y transparencia en su uso. La combinación de explicabilidad y monitoreo de modelos se considera crucial, pero es primordial asegurarse de que los proveedores que soportan explicabilidad también brinden opciones de monitoreo. Estas funcionalidades respaldan objetivos diferentes, mientras que la explicabilidad se centra en la confiabilidad y previsibilidad del modelo, el monitoreo del modelo se enfoca en su rendimiento. Por lo tanto, es necesario

evaluar la viabilidad de compartir el código de explicabilidad con el monitoreo del modelo, considerando cuidadosamente los factores de influencia y las características relevantes.

La Comisión Europea (2019) destaca la importancia de divulgar los elementos relevantes de un sistema de IA, como los datos, el sistema mismo y los modelos de negocio, para garantizar la transparencia. Además, al evaluar la confianza en la IA, es imprescindible considerar el impacto en el bienestar de las personas. Los sistemas de IA deben diseñarse y utilizarse de manera que no causen daño, protegiendo la seguridad, la salud mental y el bienestar general de los usuarios (pág. 22). Sin embargo, es necesario cuestionar hasta qué punto la divulgación de estos elementos realmente aumenta la transparencia y cómo se pueden abordar los desafíos asociados con la documentación exhaustiva de los procesos y conjuntos de datos.

Para lograr una comprensión más profunda de los modelos de IA, se requiere la aplicación de técnicas de explicabilidad e interpretabilidad. Por otro lado, es importante reconocer que no existe una solución única que se adapte a todos los escenarios y tipos de modelos de IA. La elección de la técnica de explicabilidad adecuada debe ser objeto de un análisis crítico que tenga en cuenta el contexto específico y los requisitos particulares del problema en cuestión. Además, cada técnica de explicabilidad presenta ventajas y limitaciones propias, lo que subraya la necesidad de una evaluación rigurosa para determinar cuál es la más idónea en cada caso.

La Tabla 2 proporciona algunos métodos de explicabilidad en Machine Learning, junto con sus descripciones correspondientes. Estos métodos ofrecen diferentes técnicas para obtener explicabilidad sobre las predicciones de los modelos. No obstante, es fundamental analizar críticamente su aplicabilidad y limitaciones, así como su capacidad para proporcionar explicaciones confiables y comprensibles que impulsen la confianza en los sistemas de IA.

Tabla 2

Métodos de Explicabilidad

Método	**Descripción**
SHAP	Método utilizado para descomponer predicciones individuales de un modelo complejo. Calcula la contribución de cada característica para la predicción con el objetivo de identificar el impacto de cada entrada.

LIME	Técnica que proporciona explicabilidad local en Machine Learning. Crea modelos locales interpretables para cada predicción como caja negra del modelo original.
TreeSHAP	Enfoque rápido para analizar modelos de árboles de decisión en la biblioteca SHAP. Brinda explicaciones para comprender las características o variables más relevantes o influyentes en la toma de decisiones de esos modelos.
Deep Explainer	Técnica para modelos de redes neuronales. Utiliza una versión basada en SHAP para explicar las predicciones de los modelos de redes neuronales.
Expected Gradients	Técnicas utilizadas con modelos de redes neuronales que permiten atribuir las predicciones del modelo a las características de entrada y visualizar la relación entre las características de entrada y las predicciones del modelo.
Linear SHAP	Enfoque de explicabilidad diseñado para modelos de regresión lineal.
KernelSHAP	Enfoque de Shapley lento basado en alteraciones que teóricamente funciona para todo tipo de modelos, pero suele ser poco utilizado debido a su lentitud y complejidad.
Surrogate Model	Enfoque de explicabilidad que consiste en construir un modelo transparente basado en las predicciones de un modelo real. Se utiliza cuando el resultado de interés no se puede medir directamente.
Individual Conditional Expectation	Técnica que visualiza cómo cambia la predicción de una instancia cuando se modifica una característica. Se muestra una línea por instancia para representar estos cambios.

Nota. La aplicación de técnicas de explicabilidad e interpretabilidad contribuye a una mayor comprensión de los modelos y permite tomar decisiones basadas en el impacto de los datos de entrada. Adaptado de *What Are the Prevailing Explainability Methods?* de Towards Data Science, 2021, https://towardsdatascience.com/what-are-the-prevailing-explainability-methods-3bc1a44f94df

Siguiendo las indicaciones de Litan et at. (2023), es de suma importancia llevar a cabo una supervisión continua de los datos de producción del modelo de IA con el fin de identificar desviaciones, sesgos, errores, ataques y problemas en el procesamiento de entrada. Asimismo, se sugiere evaluar la precisión del modelo después de cada predicción para mantener una vigilancia constante sobre su rendimiento a lo largo del tiempo. Es recomendable utilizar herramientas que resalten y den aviso acerca de posibles inconvenientes y anomalías en los datos antes de tomar decisiones basadas en el modelo. A continuación, en la Tabla 3, se ofrece una lista de herramientas de monitoreo de modelos, las cuales posibilitarán el análisis de los datos de producción y la detección de cualquier degradación en las características clave del modelo.

Tabla 3

Herramientas de Explicabilidad y Monitoreo de IA

Herramienta	**Descripción**
Arize AI	Arize AI es una plataforma de observabilidad. Permite detectar y solucionar problemas relacionados con modelos en producción, monitorear el rendimiento de Machine Learning, detectar patrones emergentes y problemas de integridad en sistemas de almacenamiento.
Qwak	Qwak es una herramienta gestionada de MLOps que brinda soporte en todo el ciclo de vida de desarrollo de modelos de Machine Learning. Permite transformar y almacenar datos, entrenar, implementar y monitorear modelos. También ofrece automatización del monitoreo de datos de los modelos.
Why Labs	Why Labs es una plataforma de observabilidad que se enfoca en preservar la privacidad mientras monitorea la calidad de los datos, cambios en el concepto y la precisión del modelo. Ayuda a capturar datos faltantes, cambios en el esquema, y genera alertas automáticas.
Evidently AI	Evidently AI es una plataforma de Machine Learning de código abierto que se centra en la depuración de problemas con modelos. Ofrece herramientas para evaluar, probar y monitorear modelos, detectando cambios en los datos y en el comportamiento del modelo a lo largo del tiempo.

Neptune AI	Neptune AI es un almacén de metadatos que realiza un seguimiento de experimentos de Machine Learning, almacena modelos y centraliza los datos relacionados en un solo lugar.
Qualdo	Qualdo es una plataforma que monitorea la calidad de datos y modelos de Machine Learning. Trabaja con múltiples nubes y con bases de datos de los principales proveedores. Identifica anomalías en los datos, genera informes y alertas. Incluye una herramienta de monitoreo de modelos que detecta degradación y fallos del modelo.
Fiddler AI	Fiddler AI es una plataforma de gestión del rendimiento de modelos que ofrece visibilidad en tareas de entrenamiento e inferencia. Además del monitoreo de métricas, se enfoca en explicar las predicciones realizadas. Proporciona explicabilidad y monitoreo de métricas de modelos en producción, especialmente en visión por computadora y procesamiento de lenguaje natural.

Nota. La tabla presenta una selección de herramientas de monitoreo de IA que brindan soporte en diferentes aspectos del ciclo de vida de desarrollo de modelos de Machine Learning. Adaptado de *Top ML Model Monitoring Tools* de Alon Lev, Qwak, 2023, https://www.qwak.com/post/top-ml-model-monitoring-tools

En el próximo análisis, examinaremos el caso de uso de America First Credit Union creado por Arize (2022), en el cual el equipo de ciencia de datos y Machine Learning ha implementado más de 30 modelos de IA en producción, afectando áreas como la otorgación de préstamos, el riesgo crediticio, la prevención de fraudes con tarjetas de crédito, entre otros. Gracias a la aplicación de Arize AI, el equipo de IA de America First Credit Union ahora puede detectar de manera inmediata cualquier degradación del rendimiento o cambio en el comportamiento de los modelos, reciben alertas en tiempo real cuando el rendimiento de un modelo cae por debajo del nivel esperado, lo que les permite tomar medidas rápidas para resolver cualquier problema.

Un beneficio clave es la capacidad de realizar análisis de causa raíz más rápidos. En lugar de realizar consultas SQL que consumen mucho tiempo y escribir scripts personalizados para analizar los datos, con Arize pueden acceder a información relevante con tan solo unos clics.

Esto les permite ahorrar tiempo y resolver los problemas de manera más eficiente, lo que tiene un impacto significativo en el rendimiento general de los modelos. Otro aspecto destacado es la detección y solución proactiva de problemas de cambio en los modelos. America First Credit Union utiliza Arize para monitorear los cambios en los modelos de préstamos y valor de vida útil, utilizando una métrica de desviación para identificar posibles problemas. Si se detecta un cambio significativo, el equipo puede revertir los cambios y realizar ajustes en los datos de entrada para asegurarse de que los modelos sigan siendo estables y precisos (pág. 4).

Olmein (2023) describe como se implementó el monitoreo de Machine Learning utilizando la plataforma WhyLabs, también con el objetivo de detectar fraudes financieros en tiempo real. Los pasos clave incluyeron una validación continua de la calidad de los datos utilizados para entrenar el modelo de ML, asegurando la precisión y consistencia de los datos para evitar resultados inexactos. Además, se realizó un monitoreo constante del rendimiento del modelo a lo largo del tiempo, utilizando técnicas como gráficos de control estadístico y comparación con modelos de referencia, con el fin de detectar cualquier cambio que pudiera afectar su precisión y confiabilidad. Asimismo, se llevó a cabo la identificación de desviación de datos, monitoreando cualquier cambio en la distribución de los datos utilizados por el modelo. Esto permitió adaptarse a las tácticas cambiantes de los defraudadores y mantener la efectividad del modelo en la detección de fraudes financieros. Por último, se utilizó la detección de anomalías para identificar desviaciones significativas en el comportamiento del modelo o en los datos de entrada, lo cual facilitó una detección temprana de actividades fraudulentas en tiempo real y una respuesta rápida para prevenir daños adicionales. Como beneficios obtenidos, este enfoque de monitoreo de ML permitió una mayor precisión del modelo, una reducción de los falsos positivos, una detección más rápida de fraudes y una mejora en la eficiencia operativa en la detección y prevención de fraudes financieros. En conjunto, estos pasos y beneficios contribuyeron a proteger a las instituciones financieras y empresas contra pérdidas económicas y daños reputacionales causados por actividades fraudulentas.

En conclusión, la implementación de Arize AI en America First Credit Union ha demostrado ser beneficiosa al permitir la detección inmediata de degradación del rendimiento o cambios en el comportamiento de los modelos de IA. Esto proporciona al equipo de IA la capacidad de tomar medidas rápidas para resolver cualquier problema y mantener la estabilidad y precisión de los modelos. Además, Arize AI ofrece la ventaja de realizar análisis de causa raíz de manera más rápida y eficiente, lo que ahorra tiempo y mejora el rendimiento general de los modelos. La capacidad de monitorear los cambios en los modelos de préstamos y valor de vida

útil también es destacable, ya que permite una solución proactiva de problemas y ajustes necesarios para mantener la estabilidad de los modelos.

Por otro lado, la implementación de WhyLabs en la detección de fraudes ha demostrado ser efectiva al utilizar técnicas como la validación continua de la calidad de los datos, el monitoreo constante del rendimiento del modelo, la identificación de deriva de datos y la detección de anomalías. Estas medidas permiten adaptarse a las tácticas cambiantes de los defraudadores y mantener la efectividad del modelo en la detección de fraudes financieros. Además, el enfoque de WhyLabs ha logrado una mayor precisión del modelo, una reducción de los falsos positivos y una detección más rápida de fraudes, lo que contribuye a la protección de las instituciones financieras contra pérdidas económicas y daños reputacionales.

Ambos enfoques presentan beneficios significativos en la detección y prevención de fraudes financieros, así como en la mejora de la eficiencia operativa. Por otro lado, es importante considerar algunas debilidades y desafíos asociados con cada enfoque. En el caso de Arize AI, aunque ofrece una solución lista para usar, puede limitar la adaptabilidad y personalización a las necesidades específicas de una organización. Por otro lado, la implementación de WhyLabs puede requerir más recursos en términos de tiempo, personal y costos debido a su enfoque personalizado y la necesidad de una validación continua de la calidad de los datos. En última instancia, la elección entre Arize AI y WhyLabs dependerá de las necesidades y prioridades particulares de una organización. Es crucial evaluar cuidadosamente los recursos disponibles, los requisitos del monitoreo de modelos y las características específicas de cada enfoque antes de tomar una decisión informada. Además, es importante considerar la escalabilidad y la capacidad de integración con los sistemas existentes en la organización. También es fundamental considerar los aspectos éticos y de privacidad asociados con el monitoreo de modelos de IA. La protección de datos y la transparencia son temas clave para tener en cuenta al seleccionar una herramienta de monitoreo. Es importante evaluar cómo cada herramienta aborda la privacidad de los datos y si proporciona mecanismos para garantizar la explicabilidad y equidad en los modelos de IA.

Por último, siguiendo las recomendaciones de Litan et al. (2023), con el objetivo de asegurar una supervisión efectiva del modelo, es esencial verificar si se están cumpliendo distribuciones equitativas y realizar comprobaciones de desviación de datos. Es importante detectar posibles fugas y envenenamiento de datos en el modelo, cumpliendo con los estándares establecidos para el consumo de datos. Asimismo, se debe medir cualquier cambio en la distribución de datos y en las covariables utilizando índices y estadísticas apropiadas, con el objetivo de tener un control completo sobre el rendimiento del modelo y tomar decisiones

informadas. A continuación, en la Tabla 4, se sugieren recomendaciones para implementar Explicabilidad y Monitoreo en los modelos de AI, siguiendo el modelo ágil propuesto en las Recomendaciones Complementarias.

Tabla 4

Recomendaciones para implementar Explicabilidad y Monitoreo

Acciones	**Descripción**
Identificar los objetivos	Generar confianza y transparencia en el uso de los modelos. Garantizar la divulgación de elementos relevantes del sistema de IA. Proteger el bienestar de los usuarios.
Establecer iteraciones	Dividir el trabajo en iteraciones cortas. Establecer entregables claros para cada iteración.
Priorizar acciones	Priorizar las acciones a implementar en cada iteración según su importancia y valor. Realizar un análisis crítico de las acciones considerando factores de influencia y características relevantes.
Implementar técnicas	Seleccionar las técnicas de explicabilidad adecuadas según el contexto y los requisitos del problema. Evaluar su aplicabilidad, limitaciones y capacidad para proporcionar explicaciones confiables y comprensibles.
Evaluar herramientas	Evaluar las herramientas de monitoreo de modelos mencionadas. Analizar capacidades y adaptabilidad a las necesidades específicas del contexto.
Realizar pruebas y ajustes	Realizar pruebas de las técnicas de explicabilidad y herramientas de monitoreo seleccionadas en un entorno controlado. Ajustar y personalizar las soluciones según sea necesario.
Implementar en producción	Implementar las técnicas de explicabilidad y herramientas de monitoreo en el entorno de producción de los modelos de IA. Establecer mecanismos para la supervisión continua de los datos de producción y la identificación de desviaciones, sesgos, errores, ataques y problemas.

Realizar seguimiento	Realizar un seguimiento continuo del rendimiento de los modelos. Recopilar retroalimentación de los usuarios y partes interesadas. Utilizar la retroalimentación para mejorar y ajustar las técnicas de explicabilidad y las herramientas de monitoreo.
Mejora continua	Mantener una mejora continua a través de la retroalimentación y la adaptación a medida que se identifiquen áreas de mejora.

Nota. Estas acciones son una guía general para la implementación de una planificación ágil en el contexto de evaluación de la Explicabilidad y Monitoreo en los modelos de IA y pueden adaptarse según las necesidades y particularidades de cada proyecto.

3.5 Implementando Privacidad

Según Litan et al. (2023), el primer paso para proteger la privacidad es considerar el uso de datos sintéticos generados por GANs o generative AI como una estrategia preventiva contra problemas de privacidad. No obstante, se debe tener en cuenta que no todos los tipos de datos sintéticos ofrecen los mismos resguardos contra la reidentificación. Por lo tanto, es crucial evaluar críticamente la efectividad y la idoneidad de los datos sintéticos utilizados en cada caso específico, teniendo en cuenta la calidad y la representatividad en relación con los datos reales.

En segundo lugar, el enfoque descentralizado de Machine Learning se propone como una solución para proteger la privacidad cuando los datos están dispersos en múltiples dispositivos (Litan et at., 2023). Sin embargo, es necesario abordar de manera crítica la utilización de datos sintéticos en este enfoque y cómo se manejan los posibles sesgos que puedan surgir. Es esencial garantizar que los datos sintéticos utilizados no introduzcan sesgos y sean representativos de los datos reales dispersos.

En tercer lugar, los autores previamente citados, mencionan el cifrado homomórfico y la computación segura multiparte (sMPC) como tecnologías relevantes para preservar la privacidad durante el uso de los datos en la IA. Por otro lado, es importante evaluar críticamente la aplicabilidad y las consideraciones prácticas de implementación de estas tecnologías en el contexto específico de cada proyecto de IA. Factores como la escalabilidad, el rendimiento y la compatibilidad con los requisitos legales y éticos deben ser considerados cuidadosamente.

Por último, la elección de las medidas de mitigación adecuadas depende del riesgo de la organización, la robustez de la tecnología y las características del caso de uso (Litan et at., 2023). Consideramos necesario realizar evaluaciones exhaustivas que aborden tanto los aspectos

técnicos como los legales y éticos. Se deben considerar aspectos como el impacto de las medidas en la precisión y el rendimiento del modelo, así como la capacidad de cumplir con las regulaciones de privacidad aplicables. A continuación, en la Tabla 5, se describe una lista de herramientas que ofrecen variantes de datos, datos sintéticos, aprendizaje federado, privacidad diferencial, computación segura multiparte y cifrado homomórfico.

Tabla 5

Herramientas de Privacidad de IA

Producto	**Ofrece**
Diveplane GEMINAI	Datos sintéticos.
Fluid Secure AI Platform	Aprendizaje descentralizado, privacidad diferencial.
ZeroReveal Machine Learning	Aprendizaje descentralizado y cifrado homomórfico.
Synthetic Data Platform	Datos sintéticos y privacidad diferencial.
XOR Platform	Computación segura multiparte y cifrado homomórfico.
Synthetic Data Platform	Datos sintéticos.
Private AI	Datos sintéticos.

Nota. La tabla proporciona una lista de empresas y sus respectivos productos relacionados con la protección de la privacidad. Adaptado de *Market Guide for AI Trust, Risk and Security Management* de Litan et al., Gartner, 2023, https://www.gartner.com/doc/reprints?id=1-2CQX56DW&ct=230303&st=sb

Siguiendo las sugerencias de Litan et al. (2023) en referencia a los datos sintéticos, se analizó el caso de uso de Fraudprotect (Eder Labs, Inc., sf), una startup enfocada en construir modelos y soluciones antifraude para empresas financieras y de pagos. La problemática se basaba en la obtención y acceso a datos confidenciales de transacciones financieras para desarrollar un sistema robusto de detección de fraude. A pesar de eso, la naturaleza sensible y

regulada de estos datos dificultaba su utilización y colaboración con las partes interesadas. Para abordar esta problemática, se implementó la plataforma Fluid, la cual proporciona una solución centrada en la privacidad y la colaboración segura. La plataforma Fluid permitió a Fraudprotect y al proveedor de servicios de pago en línea trabajar juntos de manera eficiente, asegurando la protección de los datos sensibles y cumpliendo con las regulaciones de privacidad.

Entre las evidencias presentes en el caso, se destaca que los datos compartidos estaban protegidos y tenían acceso limitado, lo que garantizaba la confidencialidad de la información. Además, la encriptación de los datos y el acceso controlado aseguraban que solo se accediera a la información de manera autorizada y segura.

Asimismo, la plataforma Fluid ofreció una evidencia tecnológicamente verificable de que el acceso a los datos se realizaba de acuerdo con los propósitos y controles establecidos, brindando confianza y asegurando el cumplimiento normativo. La generación de pruebas de privacidad listas para auditorías desde el inicio demostró el compromiso de las partes involucradas en proteger la confidencialidad de los datos sensibles (Eder Labs, Inc., sf).

El siguiente caso de uso analizado de Mostly (2022), se trata de Synthetic Data Platform, haciendo mención de que los sistemas tradicionales basados en reglas generan un alto número de falsos positivos y requieren un proceso laborioso de seguimiento. Es común que los fraudes raros de alto valor pasen desapercibidos y que las señales que alertan sobre actividades fraudulentas sean engañosas. Para abordar estos desafíos, Mostly propone utilizar datos de entrenamiento sintéticos para aumentar el rendimiento del Machine Learning en la detección de fraudes. Se argumenta que los conjuntos de datos de entrenamiento sintéticos son mejores que los datos reales porque, al equilibrar el conjunto de datos, el modelo puede detectar casos de fraude de manera más eficiente, lo que se refleja en un número alto y constante de AUC (Área Bajo la Curva). El AUC se utiliza para evaluar el rendimiento de un clasificador binario de Machine Learning, como un algoritmo de detección de fraudes, midiendo la tasa de verdaderos positivos frente a la tasa de falsos positivos. El objetivo es tener un AUC alto y constante para que cualquier umbral seleccionado tenga una alta tasa de verdaderos positivos y una baja tasa de falsos positivos. Se menciona la importancia de generar regularmente lotes grandes de datos de entrenamiento sintéticos a partir de conjuntos de datos transaccionales sin procesar para recalibrar el algoritmo y capturar nuevos patrones y señales.

Como evidencia de la efectividad de la plataforma Synthetic Data Platform, se mencionan diversos estudios de casos en los que se ha demostrado una mejora consistente en la curva AUC en comparación con el uso de datos sin procesar y desequilibrados. Se destaca que una mejora del 10% en el AUC podría resultar en una disminución del 10% en los falsos positivos. Se plantea

un escenario en el que un modelo tiene una tasa de falsos positivos del 1% y ha identificado 100,000 casos de fraude positivos, de los cuales aproximadamente 1000 podrían no ser realmente fraudes. Si se reduce la tasa de falsos positivos al 0.9%, solo se identificarían incorrectamente alrededor de 900 casos (Mostly, 2022).

Estas medidas parecen cumplir con las recomendaciones de Litan et al., (2023) en términos de proteger la privacidad y utilizar enfoques preventivos para evitar problemas de privacidad. No obstante, los casos no proporcionan información detallada sobre cómo se generan y equilibran los datos sintéticos, lo que dificulta la evaluación de su representatividad y su capacidad para capturar los patrones y señales necesarios en la detección de fraudes. Es importante considerar que los resultados de los estudios de casos pueden variar en diferentes escenarios y contextos, por lo que se requiere una evaluación cuidadosa de la efectividad y las limitaciones de esta solución.

En la Tabla 6, se sugieren recomendaciones para implementar Privacidad en los modelos de AI, siguiendo el modelo de AI TRiSM.

Tabla 6

Recomendaciones para implementar Privacidad

Acción	Descripción
Analizar el uso de datos sintéticos	Realizar una evaluación de la efectividad e idoneidad de los datos sintéticos en cada caso, considerando su calidad y representatividad.
Evaluar el enfoque descentralizado de Machine Learning	Considerar la opción de utilizar el enfoque descentralizado de Machine Learning como una medida de protección de la privacidad en situaciones donde los datos están distribuidos en diferentes dispositivos. Evaluar de manera crítica la utilización de datos sintéticos en este enfoque y asegurar que sean representativos y no introduzcan sesgos.
Evaluar la aplicabilidad de tecnologías de privacidad	Evaluar cuidadosamente si el cifrado homomórfico y la computación segura multiparte son adecuados para cada proyecto de IA, teniendo en cuenta aspectos como la capacidad de escalar, el rendimiento y la conformidad con requisitos legales y éticos.

Llevar a cabo evaluaciones minuciosas	Realizar evaluaciones que aborden aspectos técnicos, legales y éticos al seleccionar medidas de mitigación de privacidad. Evaluar el impacto en la precisión y rendimiento del modelo, así como la capacidad de cumplir con las regulaciones de privacidad aplicables.
Investigar y considerar herramientas de privacidad	Investigar y evaluar herramientas de privacidad de IA que ofrezcan variantes de datos, datos sintéticos, aprendizaje federado, privacidad diferencial, computación segura multiparte y cifrado homomórfico. Seleccionar las herramientas adecuadas según las necesidades y requisitos del proyecto de IA.

Nota. La tabla proporciona una descripción detallada de diversas acciones y recomendaciones para garantizar la Privacidad en proyectos de IA.

3.6 Implementando Seguridad en las Aplicaciones

En el contexto de la Seguridad de las Aplicaciones de IA, es fundamental implementar nuevas técnicas y marcos de trabajo para detectar y detener los ataques adversarios, como señalan los autores Litan et al. (2023). Estos ataques maliciosos pueden provocar diversos daños a las organizaciones, como pérdidas financieras, daño a la reputación, pérdida de propiedad intelectual o filtración de información de las personas. Por lo tanto, los líderes de seguridad deben incorporar controles y prácticas especializadas para proteger las aplicaciones de IA, además de las medidas de protección de datos y seguridad de aplicaciones que ya se implementan para otros tipos de aplicaciones.

La seguridad de las aplicaciones de IA debe aplicar prácticas especializadas, políticas, herramientas de detección y prevención de amenazas diseñadas para proteger contra amenazas nuevas y especializadas dirigidas a estas aplicaciones. Se deben implementar controles durante el desarrollo, validación, implementación, pruebas finales y ejecución de las aplicaciones, como mencionan Litan et al. (2023). Esto garantiza que los modelos de IA estén protegidos en todas las etapas de su ciclo de vida. Un elemento central en la seguridad de las aplicaciones de IA es la resistencia a los ataques adversarios. Esto implica enseñar a los modelos a ignorar o responder de manera diferente ante entradas adversarias durante su desarrollo, pruebas y producción. Es posible aplicar técnicas de entrenamiento o herramientas de IA específicas para fortalecer la resistencia de los modelos y minimizar el impacto negativo en su rendimiento.

Los ataques adversarios pueden modificar los resultados de los algoritmos Machine Learning en beneficio del atacante, como menciona Litan et al. (2023). Estos ataques pueden realizarse mediante la introducción de entradas maliciosas o perturbaciones de datos una vez que el modelo está en producción. Algunas herramientas de resistencia a los ataques adversarios pueden identificar las vulnerabilidades de los modelos y proporcionar entradas de entrenamiento para mitigar estos ataques. En la Tabla 7, evaluaremos las soluciones que existen en el mercado, que se pueden aplicar para mitigar los ataques adversarios.

Tabla 7

Herramientas de Seguridad en las Aplicaciones

Producto	**Descripción**
AIShield	Proporciona protección integral, incluyendo análisis de vulnerabilidades, detección en tiempo real y resistencia ante ataques adversarios.
VESPR Validate	Incluye un mecanismo integrado de validación para datos y modelos.
HiddenLayer MLSec Platform	Incluye detección y respuesta de ML, aseguramiento de la integridad y el escaneo de vulnerabilidades de los activos de ML, y la generación de informes de auditoría y priorización de vulnerabilidades.
Stained Glass Transform	Contribuye a salvaguardar tanto los datos utilizados para el entrenamiento como los datos de inferencia, al mismo tiempo que ofrece funcionalidades para contrarrestar ataques adversariales.

Nota. La tabla proporciona una lista de soluciones relacionadas con la protección de la seguridad en las aplicaciones. Adaptado de *Market Guide for AI Trust, Risk and Security Management* de Litan et al., Gartner, 2023, https://www.gartner.com/doc/reprints?id=1-2CQX56DW&ct=230303&st=sb

A continuación, analizaremos un caso de uso de CalypsoAI (2022), utilizando la solución VESPR Validate, aplicada en la Administración de Seguridad del Transporte. Esta entidad del Departamento de Seguridad Nacional (DHS) de los Estados Unidos se encarga de salvaguardar

la seguridad de los sistemas de transporte a nivel nacional, con especial atención en los aeropuertos donde se realizan los procedimientos de control de seguridad. La institución tiene una aplicación llamada Screening at Speed (SaS) que utiliza Inteligencia Artificial y Machine Learning para mejorar la eficiencia y seguridad de la experiencia de los pasajeros. Estos modelos ayudan a los oficiales a detectar elementos prohibidos y personas u objetos de interés, al tiempo que reducen los retrasos y las inspecciones físicas que ralentizan las filas de pasajeros en las condiciones actuales.

El desafío que enfrentan consiste en garantizar que los modelos de Machine Learning utilizados en el programa funcionen de manera efectiva para mantener la seguridad de los viajes aéreos en los Estados Unidos. Un solo fallo o ataque a un modelo podría poner en riesgo la vida de las personas, tanto en el aire como en tierra. Este departamento requería pruebas y validación de estos modelos, entre otros, para avanzar con el programa SaS. Sin un proceso establecido de pruebas y validación, podría llevar meses o más identificar las debilidades de los modelos, lo que retrasaría el proyecto en su conjunto. Es aquí donde VESPR Validate entró en acción en el flujo de trabajo del programa SaS, con el objetivo de validar y acelerar de manera eficiente los modelos para su producción. Dado el gran número de modelos desarrollados para SaS, así como la criticidad y sensibilidad de la misión, contar con una solución de prueba y validación repetible es esencial para generar confianza generalizada en los modelos. Con confianza, los modelos pueden ser implementados (CalypsoAI, 2022).

Por otro lado, el caso de uso de Bosch AIShield (2022) de la industria automotriz, se destaca la necesidad de realizar tratamientos de gases de escape en motores diésel para reducir las emisiones de NOx. El uso de la Reducción Catalítica Selectiva (SCR) puede reducir eficientemente el NOx hasta un 98%. Sin embargo, este proceso requiere el uso costoso de Diesel Exhaust Fluid (DEF), aproximadamente $10-12 para un viaje de 700 millas. Para evitar estos costos, las empresas de transporte instalan dispositivos ilegales conocidos como Tampering Dongles para eliminar el uso de DEF. Para detectar estas manipulaciones, se implementa una red neuronal en la Unidad de Control del Motor (ECU) / Unidad de Control Digital (DCU). Por otro lado, existe el riesgo de que hackers extraigan este algoritmo y lo utilicen para evadir o manipular los sistemas, lo que tendría consecuencias negativas para el medio ambiente.

Ante estas amenazas, Bosch AIShield (2022) identificó y demostró las vulnerabilidades de seguridad del algoritmo de detección de manipulaciones a través de análisis de riesgos y amenazas del modelo y sistema de IA. Se desarrolló un modelo de defensa específico para Machine Learning y se integró en el entorno de la ECU. Además, se mejoró la función existente de detección de manipulaciones basada en reglas y conectividad. Como resultado, se logró

mejorar la seguridad en general, cumplir con estándares más altos de cumplimiento y sostenibilidad, y prevenir la pérdida de confianza y negocio por parte de los clientes de la empresa automotriz.

En conclusión, CalypsoAI (2022) proporciona la solución VESPR Validate para la Administración de Seguridad del Transporte de los Estados Unidos. La implementación de esta solución tiene como objetivo mejorar la eficiencia y la seguridad en los controles de seguridad en los aeropuertos. No obstante, existen algunas limitaciones y desafíos que deben abordarse. Una limitación importante es la dependencia de los modelos de Machine Learning para detectar elementos prohibidos y personas u objetos de interés. Estos modelos pueden tener falsos positivos o falsos negativos, lo que puede resultar en interrupciones innecesarias para los pasajeros o pasar por alto amenazas potenciales. Además, la precisión de los modelos puede verse comprometida debido a cambios en los patrones de amenazas o la aparición de nuevos métodos para eludir la detección.

Otro desafío es garantizar la seguridad de los modelos de Machine Learning utilizados en el programa SaS. Un solo fallo o ataque podría poner en riesgo la vida de las personas. Por lo tanto, es crucial implementar técnicas y marcos de trabajo para detectar y detener los ataques adversarios, como recomienda Litan et al. (2023). Esto implica proteger los modelos de IA contra ataques maliciosos que podrían manipular los resultados de los algoritmos o comprometer la seguridad del sistema. La validación y prueba de los modelos de manera efectiva es fundamental para generar confianza en su desempeño y garantizar la seguridad de los viajes aéreos.

En el segundo caso de uso, Bosch AIShield (2022) aborda la necesidad de reducir las emisiones de NOx en motores diésel mediante el uso de la Reducción Catalítica Selectiva. Sin embargo, se plantea el problema de los dispositivos ilegales conocidos como Tampering Dongles que eliminan el uso de Diesel Exhaust Fluid (DEF) y comprometen la eficacia del sistema. AIShield propone una solución que mejora la seguridad y detecta manipulaciones a través de un modelo de defensa específico para Machine Learning. Una limitación en este caso de uso es la posibilidad de que los hackers extraigan el algoritmo de detección de manipulaciones y lo utilicen para evadir o manipular los sistemas. Esto resalta la importancia de implementar medidas de seguridad adicionales, como el análisis de riesgos y amenazas del modelo y sistema de IA. Además, es necesario mantenerse actualizado sobre las técnicas y tácticas utilizadas por los hackers para adaptar continuamente las medidas de seguridad.

En ambos casos, se evidencia el potencial de la IA y el Machine Learning para mejorar la eficiencia y la seguridad en diferentes sectores, como la seguridad del transporte y la reducción de emisiones en la industria automotriz. Estas soluciones pueden agilizar los procesos, detectar

amenazas o manipulaciones, y mejorar la confiabilidad de los sistemas. Por otro lado, es fundamental abordar las limitaciones y desafíos presentes en estos casos de uso. La seguridad y la confiabilidad de los modelos de IA son aspectos críticos que deben ser considerados en todas las etapas de desarrollo, implementación y operación. Las soluciones propuestas, como VESPR Validate y el modelo de defensa de Bosch AIShield, son pasos en la dirección correcta, pero es importante seguir mejorando y fortaleciendo las medidas de seguridad.

Las recomendaciones de Litan et al. (2023) en cuanto a la seguridad de las aplicaciones de IA son relevantes para ambos casos. Es necesario implementar técnicas y marcos de trabajo para detectar y detener los ataques adversarios. En el caso de CalypsoAI, se sugiere fortalecer la resistencia de los modelos de Machine Learning ante ataques adversarios. Esto implica enseñar a los modelos a ignorar o responder de manera diferente ante entradas adversarias. En el caso de Bosch AIShield, se debe poner un énfasis especial en la seguridad de los algoritmos de detección de manipulaciones. Además de las medidas de seguridad ya implementadas, se puede considerar la aplicación de herramientas de resistencia a ataques adversarios que identifiquen vulnerabilidades y proporcionen entradas de entrenamiento para mitigar estos ataques. Estas recomendaciones buscan garantizar la seguridad y confiabilidad de las aplicaciones de IA en diferentes sectores. En la Tabla 8, se presenta un resumen de las sugerencias para implementar seguridad en las aplicaciones de IA.

Tabla 8

Recomendaciones para implementar Seguridad en las Aplicaciones

Acción	**Descripción**
Implementar técnicas y marcos de trabajo	Detectar y prevenir ataques maliciosos en aplicaciones de IA. Aplicar medidas especializadas para proteger las aplicaciones de IA en todas las etapas de desarrollo.
Reforzar la capacidad de los modelos para resistir ataques	Enseñarles a ignorar o responder de manera distinta a entradas adversarias. Reducir el impacto negativo en el rendimiento de los modelos.
Aplicar recursos de defensa	Implementar herramientas de resistencia a ataques adversarios que identifiquen vulnerabilidades en los modelos y proporcionen entradas de entrenamiento para mitigar estos ataques.

Efectuar pruebas y verificaciones	Llevar a cabo pruebas y validaciones efectivas de los modelos de Machine Learning para generar confianza en su desempeño y garantizar la seguridad en aplicaciones críticas.
Mantenerse actualizado sobre técnicas	Estar al día sobre las técnicas y tácticas utilizadas por los hackers para adaptar continuamente las medidas de seguridad. Esto es importante para casos como la detección de manipulaciones, donde los hackers pueden extraer algoritmos y utilizarlos para evadir o manipular los sistemas.
Mejorar la precisión de los modelos	Implementar enfoques adicionales para mejorar la precisión de los modelos de Machine Learning y reducir falsos positivos y falsos negativos. Adaptar los modelos a cambios en los patrones de amenazas o la aparición de nuevos métodos para eludir la detección.

Nota. La tabla proporciona una lista de recomendaciones relacionadas con la protección de la seguridad en las aplicaciones. Adaptado de *Market Guide for AI Trust, Risk and Security Management* de Litan et al., Gartner, 2023, https://www.gartner.com/doc/reprints?id=1-2CQX56DW&ct=230303&st=sb

Capítulo 4 – Conclusiones

4.1 Conclusión

En el presente Trabajo Final de Carrera, se han cuestionado las prácticas convencionales de seguridad en la Inteligencia Artificial al priorizar la prevención proactiva de ciberataques y la eliminación de vulnerabilidades desde las primeras etapas del ciclo de vida de un sistema de IA. Estas herramientas permiten detectar tempranamente degradaciones en el rendimiento, cambios inesperados en el comportamiento y desviaciones en los datos de producción de los modelos, lo cual no solo acelera la respuesta a problemas potenciales, sino que también cambia significativamente la forma en que se enfrentan y mitigan las amenazas en el campo de la IA.

Además, se destaca la relevancia de la privacidad y la protección de datos personales en las aplicaciones de IA. La investigación resultó en la implementación de medidas sólidas de seguridad para resguardar la información confidencial de los usuarios y prevenir posibles fugas o accesos no autorizados. Al mismo tiempo, se fomenta la conciencia de los usuarios sobre los permisos y configuraciones de privacidad en las aplicaciones de IA, capacitándolos para tomar decisiones informadas sobre la protección de sus datos personales.

Se examinaron minuciosamente las recomendaciones y estrategias específicas propuestas por Gartner con el fin de asegurar la confianza, gestionar el riesgo y fortalecer la integridad en el ámbito de la IA. Este análisis ha tenido un impacto relevante al enriquecer la comprensión y perspectiva sobre estos aspectos.

Por ejemplo, en el caso de Arize AI, ha permitido a America First Credit Union detectar degradaciones en el rendimiento de los modelos de IA de manera inmediata y tomar medidas rápidas para resolver problemas. En el caso de WhyLabs, se ha logrado una detección más rápida de fraudes financieros a través del monitoreo continuo en el rendimiento de los modelos.

Estas soluciones agilizan los procesos, mejoran la eficiencia operativa y permiten una respuesta más rápida ante situaciones problemáticas.

Sin embargo, este trabajo no está exento de limitaciones significativas. La falta de recursos adecuados para la implementación y validación práctica de estas recomendaciones ha sido un obstáculo importante. La ejecución de pruebas en entornos controlados y la implementación de soluciones prácticas requieren inversiones financieras y tecnológicas considerables, lo que afecta la capacidad para proporcionar evidencia empírica sólida sobre la eficacia de las medidas.

Se reconoce, además, la complejidad de implementar estas medidas en sistemas de IA existentes. Las organizaciones pueden enfrentar desafíos logísticos y técnicos al adoptar estas prácticas en su infraestructura actual, lo que limita la aplicabilidad inmediata de la solución.

Cada aplicación de IA puede presentar desafíos únicos que no se han considerado en profundidad en este estudio. Esta limitación destaca la necesidad de adaptar y personalizar estas recomendaciones para abordar circunstancias específicas de uso.

La evolución constante en el campo de la IA es otro desafío que se ha identificado. Los avances tecnológicos y los cambios regulatorios ocurren a un ritmo rápido, lo que implica que las soluciones sugeridas en este estudio podrían volverse obsoletas en un futuro cercano. Esta naturaleza dinámica de la IA plantea desafíos continuos en la gestión de la seguridad que deberán abordarse a medida que surgen.

A pesar de los obstáculos y limitaciones enfrentados, este trabajo representa un paso fundamental hacia un enfoque más proactivo y comprensivo de la seguridad en la IA. Se ha desarrollado un marco de referencia integral que aborda una amplia gama de amenazas y desafíos en la seguridad de la IA, proporcionando una base teórica y práctica para guiar a profesionales de la seguridad y desarrolladores de sistemas de IA en la comprensión y mitigación de riesgos en entornos de IA.

La investigación también ha brindado un análisis crítico de casos de uso, lo cual es valioso para comprender el tema y sentar las bases para trabajos futuros. La falta de recursos para probar la hipótesis también abre oportunidades para futuras investigaciones que puedan abordar esta cuestión en un entorno práctico, en tanto y en cuanto, se tengan los recursos necesarios para llevar a cabo la implementación y pruebas correspondientes.

Esta investigación ha adoptado un enfoque que considera tanto los aspectos técnicos como los éticos, legales y de cumplimiento normativo de la seguridad en la IA. Se reconoce la complejidad de este campo y la necesidad de abordar múltiples dimensiones para proteger adecuadamente los sistemas de IA. Estas recomendaciones y herramientas se convertirán en estándares que revolucionarán la forma en que se aborda la seguridad en la IA, permitiendo un entorno más seguro y confiable para su aplicación en diversas áreas. Se está comprometido a continuar avanzando en esta dirección y a contribuir a un futuro más seguro en el mundo de la Inteligencia Artificial.

4.2 Líneas Futuras de Investigación

Las futuras líneas de investigación deben enfocarse en la aplicación práctica de las recomendaciones propuestas en este trabajo para corroborar la hipótesis planteada. Se sugiere llevar a cabo estudios empíricos que implementen el marco AI TRiSM en diferentes contextos de aplicaciones de Inteligencia Artificial. Estos estudios podrían evaluar el impacto de la implementación de AI TRiSM en términos de mejoras en la seguridad y confianza de los modelos

de IA, así como en la mitigación de riesgos y la promoción de la equidad. Sería importante examinar cómo las recomendaciones principales e implementar el marco de trabajo, pueden influir en la calidad y confiabilidad de los sistemas de IA.

Además, se podrían realizar estudios comparativos entre diferentes técnicas de explicabilidad para determinar cuál es la más adecuada en diferentes escenarios y contextos de aplicación. Estas investigaciones podrían evaluar las ventajas y limitaciones de cada técnica, así como su impacto en la comprensión y confianza de los usuarios en los modelos de IA.

Acrónimos

AGI	Artificial General Intelligence
AI	Artificial Intelligence
ANI	Artificial Narrow Intelligence
ASI	Artificial Superintelligence
DBA	Danish Business Authority
DP	Deep Learning
DPI	Data Protection Impact
FPGA	Field-Programmable Gate Array
GPU	Graphics Processing Unit
IA	Inteligencia Artificial
ICE	Individual Conditional Expectation
LIME	Local Interpretable Model-Agnostic Explanations
LS	Transport Layer Security
ML	Machine Learning
MLOps	Machine Learning Operations
PET	Privacidad, Ética y Transparencia
RMF	Risk Management Framework
SHAP	Shapley Additive exPlanations
sMPC	Secure Multiparty Computation
SSL	Secure Sockets Layer
TEV	Prueba, Evaluación y Validación
TPU	Tensor Processing Unit
UC Irvine	University of California, Irvine

Referencias Bibliográficas

Arize. (Junio de 2022). *America First Credit Union Pioneers A New Era of AI-Led Growth.* Arize: https://arize.com/wp-content/uploads/2022/06/America-First-Credit-Union-Arize-Case-Study.pdf

Ashoori, M., & Weisz, J. D. (5 de Diciembre de 2019). *In AI We Trust? Factors That Influence Trustworthiness of AI-infused Decision-Making Processes.* arXiv.org: https://arxiv.org/pdf/1912.02675.pdf

Biswal, A. (13 de Febrero de 2023). *7 Types of Artificial Intelligence That You Should Know in 2023*. Simplilearn: https://www.simplilearn.com/tutorials/artificial-intelligence-tutorial/types-of-artificial-intelligence

Bosch AIShield. (3 de Junio de 2022). *BOSCH AISHIELD CASE STUDIES.* https://www.boschaishield.com/media/downloads/case_study.pdf

CalypsoAI. (25 de Agosto de 2022). *How CalypsoAI and the U.S. Department of Homeland Security are Developing the Next Generation of Air Passenger Screening*. https://calypsoai.com/how-calypsoai-and-the-u-s-department-of-homeland-security-are-developing-the-next-generation-of-air-passenger-screening/

Comisión Europea. (8 de Abril de 2019). *Directrices Éticas para una IA fiable.* https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60423

Dar, P. (24 de Marzo de 2023). *Protecting AI Models from "Data Poisoning"*. IEEE Spectrum: https://spectrum.ieee.org/ai-cybersecurity-data-poisoning

Dhinakaran, A. (Diciembre de 2021). *What Are the Prevailing Explainability Methods?* Towards Data Science: https://towardsdatascience.com/what-are-the-prevailing-explainability-methods-3bc1a44f94df

Durgia, C. (30 de Diciembre de 2021). *Using SHAP for Explainability — Understand these Limitations First*. Towards Data Science: https://towardsdatascience.com/using-shap-for-explainability-understand-these-limitations-first-1bed91c9d21

Eder Labs, Inc. (sf). *Securely sharing sensitive intelligence among multiple parties for AI-driven fraud prevention.* https://www.eder.io/_files/ugd/2aa7e2_9513834d12074c4ca8f0f17cb7d19042.pdf

Gamco. (2021). *Types of artificial intelligence according to their capabilities and functionality*. https://gamco.es/en/types-of-artificial-intelligence-capacity-functionality/

Gartner. (Diciembre de 2021). *What Is Artificial Intelligence (AI)*. https://www.gartner.com/en/topics/artificial-intelligence

Google. (2022). *¿Qué es la inteligencia artificial o IA?* https://cloud.google.com/learn/what-is-artificial-intelligence

Gretel. (19 de Mayo de 2022). *How to create differentially private synthetic data.* https://gretel.ai/videos/how-to-create-differentially-private-synthetic-data

Javatpoint. (2021). *Types of Artificial Intelligence*. https://www.javatpoint.com/types-of-artificial-intelligence

Joshi, N. (19 de Junio de 2019). *7 Types Of Artificial Intelligence*. Forbes: https://www.forbes.com/sites/cognitiveworld/2019/06/19/7-types-of-artificial-intelligence

Kaur, J. (5 de Enero de 2023). *AI TRiSM Challenges and its Framework for Businesses in 2023*. Xenonstack: https://www.xenonstack.com/blog/ai-trism

Lateef, Z. (3 de Marzoo de 2023). *Types Of Artificial Intelligence You Should Know*. Edureka: https://www.edureka.co/blog/types-of-artificial-intelligence

Lev, A. (12 de Enero de 2023). *Top ML Model Monitoring Tools*. Qwak: https://www.qwak.com/post/top-ml-model-monitoring-tools

Litan, A. (18 de Enero de 2023). *How can you trust ChatGPT or any other model for that matter? AI TRiSM*. Gartner: https://blogs.gartner.com/avivah-litan/2023/01/18/how-can-you-trust-chatgpt-or-any-other-model-for-that-matter-ai-trism/

Litan, A., D'Hoinne, J., Willemsen, B., & Agarwal, S. (16 de Enero de 2023). *Market Guide for AI Trust, Risk and Security Management.* Gartner: https://www.gartner.com/doc/reprints?id=1-2CQX56DW&ct=230303&st=sb

Lukyanenko, R., Maass, W., & Storey , V. C. (28 de Noviembre de 2022). Trust in artificial intelligence: From a Foundational Trust Framework to emerging research opportunities. https://doi.org/10.1007/s12525-022-00605-4

Mostly. (7 de Julio de 2022). *Power Up Fraud Detection Models with Synthetic Data.* https://mostly.ai/case-study/synthetic-training-data-for-machine-learning-fraud-detection

Olmeim, K. (7 de Marzo de 2023). *Detecting Financial Fraud in Real-Time: A Guide to ML Monitoring.* WhyLabs: https://whylabs.ai/blog/posts/detecting-financial-fraud-in-real-time-a-guide-to-ml-monitoring

Perri, L. (19 de Octubre de 2022). *Qué se necesita para que la inteligencia artificial sea segura y efectiva.* Gartner: https://www.gartner.es/es/articulos/que-se-necesita-para-que-la-inteligencia-artificial-sea-segura-y-efectiva

Russell, S., & Norvig, P. (2020). *Artificial Intelligence: A Modern Approach (Fourth Edition).* Pearson.

Schwaber, K., & Sutherland, J. (Noviembre de 2020). *La Guía de Scrum.* La Guía Definitiva de Scrum: Las Reglas del Juego: https://scrumguides.org/docs/scrumguide/v2020/2020-Scrum-Guide-Spanish-Latin-South-American.pdf

Siddiqui, L. (28 de Marzo de 2023). *Splunk.* AI TRiSM Explained: AI Trust, Risk & Security Management: https://www.splunk.com/en_us/blog/learn/ai-trism-ai-trust-risk-security-management.html

Sokhach, D. (21 de Marzo de 2023). *4 Ways to Preserve Privacy in Artificial Intelligence*. Landbot: https://landbot.io/blog/preserve-privacy-artificial-intelligence

Stefanini Group. (10 de Diciembre de 2022). *Gartner® Top Tech Trends For 2023: AI TRiSM*. https://stefanini.com/en/insights/news/gartner-top-tech-trends-for-2023-ai-trism

Techliance. (2023). *Types of Artificial Intelligence: Categories of AI*. https://blog.techliance.com/types-of-artificial-intelligence

U.S. Department of Commerce. (Enero de 2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). Estados Unidos. https://doi.org/10.6028/NIST.AI.100-1

White House Office of Science and Technology Policy. (Octubre de 2022). *Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People.* https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf

Wiggers, K. (29 de Mayo de 2021). *Adversarial attacks in machine learning: What they are and how to stop them*. Venture Beat: https://venturebeat.com/security/adversarial-attacks-in-machine-learning-what-they-are-and-how-to-stop-them/

Wiles, J. (15 de Septiembre de 2022). *Novedades del Hype Cycle de Gartner para la inteligencia artificial de 2022*. Gartner: https://www.gartner.es/es/articulos/novedades-del-hype-cycle-de-gartner-para-la-inteligencia-artificial-2022