AUDITORÍA INTERNA

LA FÁBRICA DE PENSAMIENTO
INSTITUTO DE AUDITORES INTERNOS DE ESPAÑA

Auditoría Interna de la Inteligencia Artificial aplicada a procesos de negocio

ACTUALIZACIÓN 2024

The Institute of Internal Auditors

Instituto de Auditores Internos de España

El INSTITUTO DE AUDITORES INTERNOS DE ESPAÑA es una asociación profesional fundada en 1983, cuya misión es contribuir al éxito de las organizaciones impulsando la Auditoría Interna como función clave del buen gobierno. En España cuenta con más de 3.500 socios, auditores internos en las principales empresas e instituciones de todos los sectores económicos del país.

LA FÁBRICA DE PENSAMIENTO es el laboratorio de ideas del Instituto de Auditores Internos de España sobre gobierno corporativo, gestión de riesgos y Auditoría Interna, donde participan más de 150 socios y profesionales técnicos expertos.

AUDITORÍA INTERNA

OBSERVATORIO SECTORIAL

PRÁCTICAS DE BUEN GOBIERNO

BUENAS PRÁCTICAS EN GESTIÓN DE RIESGOS

El laboratorio trabaja con un enfoque práctico en la producción de documentos de buenas prácticas que contribuyan a la mejora del buen gobierno y de los sistemas de gestión de riesgos en organizaciones de habla hispana. Además de desarrollar contenido, fomenta el intercambio de conocimientos entre los socios.

ENCUENTRA TODOS LOS DOCUMENTOS DE LA FÁBRICA EN www.auditoresinternos.es

AUDITORÍA INTERNA

LA FÁBRICA DE PENSAMIENTO
INSTITUTO DE AUDITORES INTERNOS DE ESPAÑA

Auditoría Interna de la Inteligencia Artificial aplicada a procesos de negocio

Actualización - Octubre 2024

MIEMBROS DE LA COMISIÓN TÉCNICA

COORDINACIÓN:

Daniel Tortosa Illana; ICAEW, ROAC. TELEFÓNICA BRASIL.

Pablo Ausín Sánchez; PMP-PMI. INDITEX.

Luis Enrique Corredera. DELOITTE.

José Ignacio Díez Arocena; CIA, CISA, CFE, COSO CI, COSO ERM, CESCOM. INDEPENDIENTE.

Javier Escribano Alarcón; CISA, COBIT, ITIL y PMP. REPSOL.

Andrés Morales Fernández. KPMG.

Borja Rioja Mata. MAPFRE.

Juan José Villar Roldán. IBERDROLA.

ADEMÁS DE LOS ANTERIORES, PARTICIPARON EN LA ELABORACIÓN DE LA VERSIÓN PRECEDENTE (2023):

Javier Echeverría Blanco. BBVA.

Alejandro Martínez Morillo; CISA, CDPSE; Lead Auditor 27001. PwC.

Jaime Sabau Jiménez. EY.

The Institute of Internal Auditors

Instituto de Auditores Internos de España

El término Inteligencia Artificial, fue acuñado por el científico de datos John McCarthy en 1956, definiéndolo como la ciencia para hacer inteligentes a las máquinas, o simplemente, los métodos para hacer que las maquinas tomen decisiones humanas para resolver problemas. La Inteligencia Artificial incluye actividades como aprendizaje, planificación, percepción y entendimiento de lenguaje o robótica.

En este documento actualizado de LA FÁBRICA DE PENSAMIENTO abordamos aspectos relacionados con los casos de uso de Inteligencia Artificial (IA) más comunes utilizados por las empresas en sus procesos de negocio y la regulación aplicable que los legisladores están promoviendo (primera parte); describimos los principales modelos y tipologías de IA, incluyendo la nueva perspectiva de la Inteligencia Artificial generativa, que la industria viene desarrollando (segunda parte); abordamos el marco de control interno general esperado y riesgos relacionados en aquellas organizaciones desplegando tecnología basada en IA (tercera parte); y proponemos un programa de trabajo para la auditoría de aquellas estructuras de control interno diseñadas e implementadas en procesos de negocio con IA, así como los principales procedimientos de auditoría sugeridos para su revisión (cuarta parte).

INSTITUTO DE AUDITORES INTERNOS DE ESPAÑA

Índice

Introducción

¿PUEDEN LAS MÁQUINAS PENSAR?

El término Inteligencia Artificial fue acuñado en 1956 por el científico de datos John McCarthy

Es la sugerente introducción de un artículo[1] publicado en 1950 por **Alan Mathison Turing,** en el que planteaba uno de los grandes interrogantes de la comunidad científica de la época, y en particular, de los matemáticos e informáticos que iniciaban sus pasos en lo que hoy es conocido por todos como Inteligencia Artificial. La idea de máquinas pensantes que puedan imitar la inteligencia humana ha ido evolucionando hasta nuestros días, de forma que, el desarrollo científico ha permitido dar respuesta a la pregunta planteada por Alan Turing a mitad del siglo XX, dando lugar a la Inteligencia Artificial tal y como la conocemos actualmente. La Inteligencia Artificial ("IA"), a pesar de ser un concepto abstracto, se encuentra cada vez más presente en nuestra vida, desde el reconocimiento facial de nuestros teléfonos móviles, hasta los asistentes de voz que utilizamos frecuentemente.

Tratando de seguir el planteamiento de Turing en el artículo que comentábamos, el matemático inglés indicaba que para poder responder a la pregunta había que definir con precisión qué se entiende por pensar y qué se entiende por *máquina.*

El término Inteligencia Artificial, fue acuñado por el científico de datos **John McCarthy**[2] en 1956, definiéndolo como la ciencia para hacer inteligentes a las máquinas, o simplemente, los métodos para hacer que las maquinas tomen decisiones humanas para resolver problemas. La Inteligencia Artificial incluye actividades como aprendizaje, planificación, percepción y entendimiento de lenguaje o robótica. .

En el presente documento abordamos aspectos relacionados con los casos de uso de Inteligencia Artificial (IA) más comunes utilizados por las empresas en sus procesos de negocio y la regulación aplicable que los legisladores están promoviendo (primera parte); describimos los principales modelos y tipologías de IA que la industria viene desarrollando (segunda parte); abordamos el marco de control interno general esperado y riesgos relacionados en aquellas organizaciones desplegando tecnología basada en IA (tercera parte); y proponemos un programa de trabajo para la auditoría de aquellas estructuras de control interno diseñadas e implementadas en procesos de negocio con IA, así como los principales procedimientos de auditoría sugeridos para su revisión (cuarta parte).

1. Alan Mathison Turing. *Computing machinery and intelligence.* 1950.
2. http://jmc.stanford.edu/artificial-intelligence/index.html

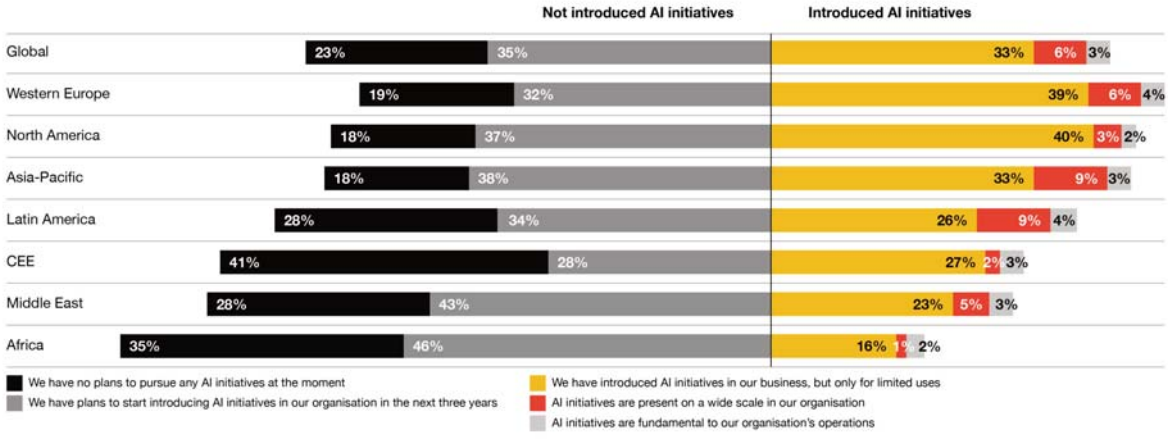
La Inteligencia Artificial en las organizaciones empresariales y sus aspectos regulatorios

1. LA PENETRACIÓN DE LA INTELIGENCIA ARTIFICIAL EN LAS ORGANIZACIONES EMPRESARIALES.

El mundo empresarial viene realizando una labor intensa en la utilización y aplicación de modelos de Inteligencia Artificial en un amplio abanico de casos de uso. Todo ello con el objeto de optimizar y dotar de mayor eficiencia a los procesos del negocio y mejorar los servicios ofrecidos a sus clientes. La integración de la Inteligencia Artificial en las organizaciones se pone de manifiesto en los resultados de la encuesta Global Anual (2019) realizada por la consultora PwC, cuyos resultados indicaban que el 85%[3] de los principales responsables de las compañías encuestadas (CEOs, *Chief Executive Officer)* consideraban que la Inteligencia Artificial cambiará significativamente la forma de hacer negocios en los próximos cinco años. Por otro lado, el 42% de los encuestados manifestaban haber introducido iniciativas de Inteligencia Artificial y el 33% tenía planes para el desarrollo de la Inteligencia Artificial en los próximos 3 años. La siguiente ilustración muestra los resultados sobre la implementación de iniciativas de Inteligencia Artificial por áreas geográficas, siendo Europa y Norte América las más activas en este sentido.

La IA se está aplicando con intensidad en la empresa para optimizar los procesos de negocio y mejorar los servicios a clientes

PLANES DE DESARROLLO DE INTELIGENCIA ARTIFICIAL



FUENTE: PwC. 22 nd Anual Global Global CEO Survey[4]

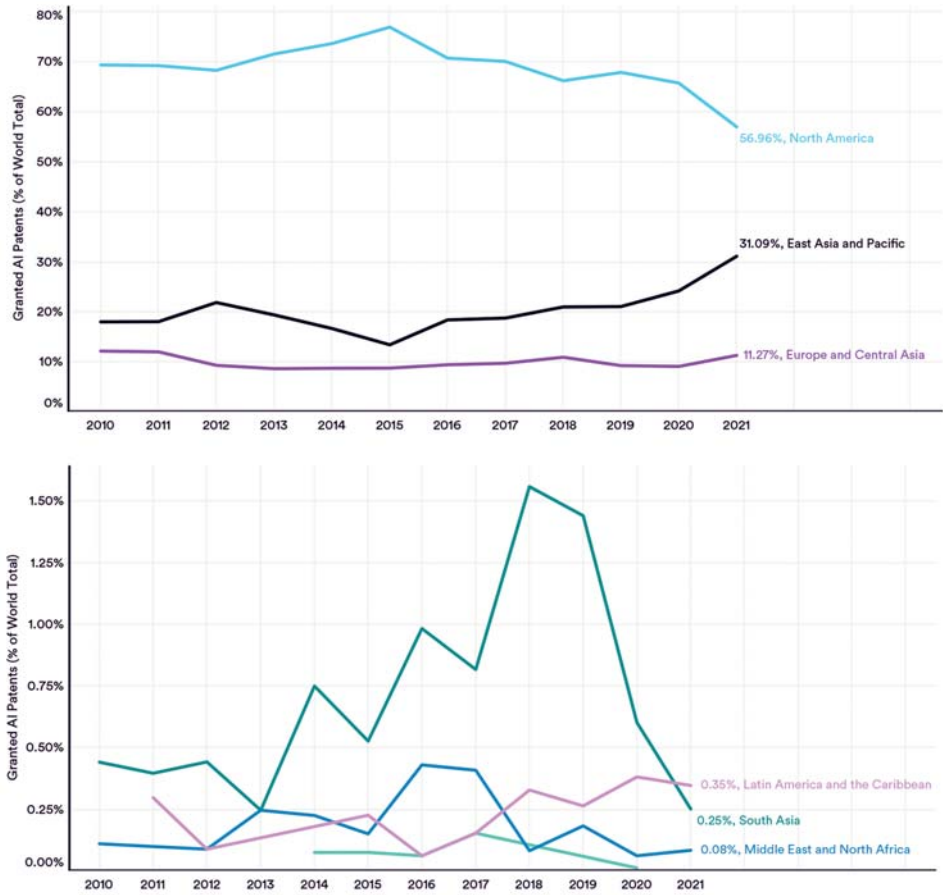
3. PwC, 22nd Annual Global Survey. 2019 (Pag. 34)
4. PwC, 22nd Annual Global Survey. 2019 (Pag. 35)

El uso de herramientas de IA en funciones empresariales ha subido del 50% al 56% en solo un año, según Mckinsey

De forma similar, los estudios de mercado realizados por la consultora McKinsey en su *Global Survey: The state of AI,* muestran índices de aplicación de modelos de Inteligencia Artificial activos y crecientes en los últimos años. En su encuesta realizada en 2020, el 50%[5] de los encuestados declararon que se habían adoptado herramientas de Inteligencia Artificial en al menos una función de la empresa. Tan solo un año más tarde, el porcentaje se incrementaba al 56%[6].

Por otro lado, y analizando las patentes concedidas en Inteligencia Artificial en el *Artificial Intelligence Index Report 2022* emitido por la Universidad de Standford, se obtiene también una visión de la penetración esperada geográficamente de iniciativas en Inteligencia Artificial que, eventual y potencialmente, terminarán en nuevos casos de uso cambiando la forma de hacer negocios e impactando nuestro día a día. En este sentido, Norte América es la región que concentra un mayor número

PATENTES DE INTELIGENCIA ARTIFICIAL POR REGIONES



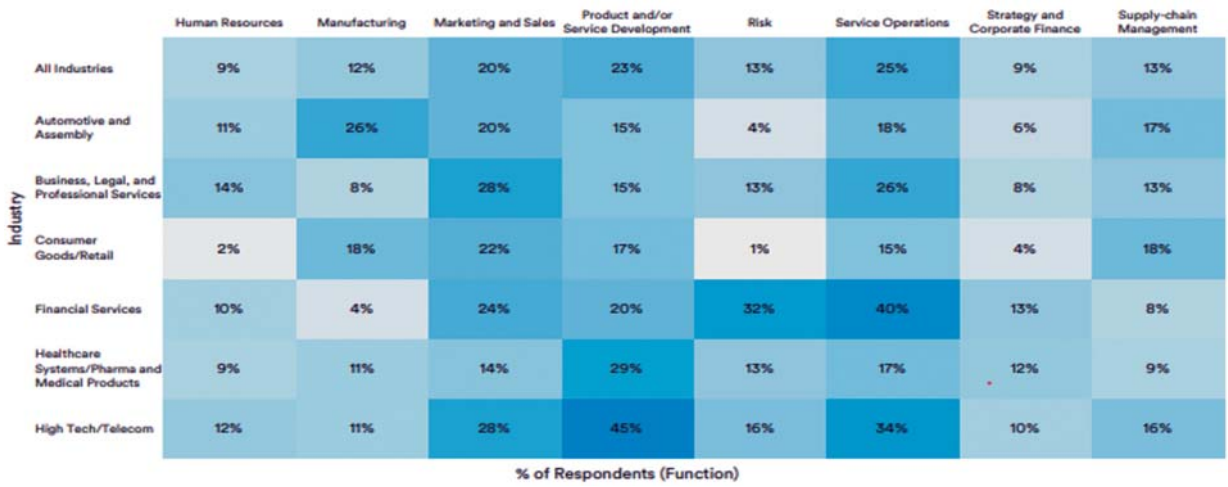
FUENTE: Standford University, Artificial Intelligence Index Report 2022

5. McKinsey Digital, Global Survey: The state of AI in 2020. (Pag. 2)
6. McKinsey Digital, Global Survey: The state of AI in 2020. (Pag. 2)

de patentes en Inteligencia Artificial en el año 2021, con un 57% del total mundial concedidas. Seguida por la región del Este de Asia con un 31%, fundamentalmente por la influencia de China, y con Europa y la región de Centro de Asia totalizando 11%. El resto de las regiones analizadas en el *Index Report* mencionado (Latinoamérica, África y Oriente Medio) muestran porcentajes de patentes concedidas muy residuales, 1% entre las tres, y alejadas de las regiones anteriores mencionadas.

En cuanto a la **implementación por industrias y áreas departamentales de las empresas,** los sectores con alto componente tecnológico *(high tech)* y empresas de telecomunicaciones son las que muestran mayores índices de adopción de modelos de Inteligencia Artificial, en las áreas de desarrollo de producto y servicios, seguido del sector financiero y bancario, tal y como muestra la ilustración siguiente, que resumen gráficamente las respuestas de los encuestados del *Global Survey: The state of AI in 2021* de la consultora McKinsey, de las industrias y áreas departamentales con mayores índices de adopción de iniciativas basadas en Inteligencia Artificial.

AI ADOPTION BY INDUSTRY AND FUNCTION, 2021



FUENTE: Global Survey: The state of AI in 2021

2. CONOCIMIENTO Y HABILIDADES DE LOS AUDITORES INTERNOS EN MATERIA DE INTELIGENCIA ARTIFICIAL.

En el contexto empresarial actual, la **profesión de Auditoría Interna** se encuentra en un proceso de evolución en el que los riesgos y **los procesos tienen un componente cada vez más tecnológico** y, en particular, la implementación de modelos de Inteligencia Artificial en los procesos de negocio supone un desafío añadido que obliga a la función de Auditoría Interna a dar una respuesta adecuada, aprovechando su posicionamiento en la compañía.

Es recomendable que el equipo de Auditoría Interna disponga de una combinación de conocimientos técnicos y humanísticos

De la capacidad de adaptación y actualización, dependerá que Auditoría Interna siga siendo un actor relevante en un área que, de forma constante y continua, aporta valor en las organizaciones. En este sentido, uno de los aspectos relevantes a considerar es la necesidad de contar con representantes de Auditoría Interna con el **talento y el conocimiento necesario** para abordar desde el principio, a través de auditorías de diseño, proyectos de auditoría en los que nos enfrentemos a procesos de negocio con estructuras de control basadas en modelos de Inteligencia Artificial. Dicho talento es una mezcla de entendimiento de los principios básicos de auditoría, así como de la posesión de los conocimientos específicos capaces de abordar los requerimientos técnicos que los modelos de Inteligencia Artificial necesitan.

Actualmente puede resultar difícil acceder en el mercado laboral a este tipo de perfiles holísticos. Es por ello, que se hace relevante que la disposición de este conocimiento, dentro de la función de Auditoría Interna, provenga también del *reskilling* de los profesionales de esta disciplina. A este respecto, es crucial que la Inteligencia Artificial se encuentre presente en los planes de formación anual de los auditores internos, sobre todo de aquellos que abordarán proyectos de revisión en procesos en los que se encuentre presente la Inteligencia Artificial, u otros aspectos tecnológicos actuales.

En cuanto a los conocimientos específicos necesarios, es recomendable que, en la medida de lo posible, el equipo de Auditoría Interna disponga de una combinación de conocimientos técnicos y otros de tipo más *humanístico*. Conjugar aspectos como las matemáticas, las tecnologías de la información, la computación, la programación y la neurociencia con conocimientos en los campos de la lógica y la filosofía, la filología e, incluso, la psicología, garantizan disponer de las capacidades necesarias para poder evaluar correctamente la Inteligencia Artificial que se encuentra presente en los diferentes procesos de negocio, no solamente desde un punto de vista de su desarrollo técnico, sino también desde la perspectiva de aquellos aspectos más ligados a la ética presente en su diseño y aplicación práctica.

3. LA "INTELIGENCIA ARTIFICIAL ACT": EL CAMINO DE EUROPA HACIA UNA REGULACIÓN COMUNITARIA SOBRE LA INTELIGENCIA ARTIFICIAL.

«En cuanto a la Inteligencia Artificial ["IA"], la confianza es una obligación, no un adorno. Mediante estas reglas de referencia, la UE lidera la formulación de nuevas normas mundiales para que garanticen que se pueda confiar en la IA. Al establecer las normas, podremos facilitar el advenimiento de una tecnología ética en todo el mundo y velar por que la UE siga siendo competitiva. Nuestras normas, que son a prueba de futuro y propicias a la innovación, intervendrán cuando sea estrictamente necesario, esto es, cuando estén en juego la seguridad y los derechos fundamentales de los ciudadanos de la UE».

Margrethe Vestager
Vicepresidenta Ejecutiva responsable de la cartera de una Europa Adaptada a la Era Digital.

El desarrollo y creciente uso de la Inteligencia Artificial en los distintos ámbitos privados y empresariales han llevado a la Comisión Europea a emprender un camino hacia su regulación. En abril del 2021, la Comisión Europea aprobó una propuesta de **Reglamento por el que se establecen normas armonizadas en la Unión Europea en materia de Inteligencia Artificial.**

Posteriormente, en mayo de 2024[7], el Consejo de la Unión Europea aprobó definitivamente la Ley de Inteligencia Artificial. Tras las aprobaciones pertinentes en la esfera europea, los estados miembros tendrán un periodo de transición de 24 meses para su aplicación efectiva en cada territorio nacional respectivo.

La propuesta de la Comisión Europea nace con el espíritu de establecer un marco reglamentario sobre Inteligencia Artificial que sigue un enfoque basado en el riesgo y establece un marco legal uniforme y horizontal para la IA que tiene como objetivos[8]:

- Garantizar que los sistemas de IA introducidos y usados en el mercado de la UE sean seguros y respeten la legislación vigente en materia de derechos fundamentales y valores de la Unión.
- Garantizar la seguridad jurídica para facilitar la inversión e innovación en IA.
- Mejorar la gobernanza y la aplicación efectiva de la legislación vigente en materia de derechos fundamentales y los requisitos de seguridad aplicables a los sistemas de IA.
- Facilitar el desarrollo de un mercado único para hacer un uso legal, seguro y fiable de las aplicaciones de IA y evitar la fragmentación del mercado.

Los **sistemas de IA de Alto Riesgo** son aquellos que presentan un alto riesgo de menoscabar la salud y la seguridad o los derechos fundamentales de las personas, teniendo en cuenta tanto la gravedad del posible perjuicio como la probabilidad de que se produzca. Dentro de los tipos de sistemas de Inteligencia Artificial merecen especial atención aquellos considerados de **Alto Riesgo** porque son el tipo permitido sobre el que más obligaciones establece el futuro reglamento.

La futura regulación europea sobre el uso de la Inteligencia Artificial clasifica los sistemas en función de la finalidad de uso y para cada tipo establece una serie de requisitos y limitaciones. Por otro lado, la propuesta del Reglamento considera situaciones en las que los sistemas de IA pueden utilizarse para varios fines diferentes (IA de propósito general), y donde la tecnología de IA de propósito general se integra posteriormente en otro sistema de alto riesgo.

Dicha clasificación se reproduce en la ilustración de la página siguiente.

El nuevo Reglamento Europeo sobre Inteligencia Artificial establece un marco legal uniforme y horizontal bajo un enfoque basado en el riesgo

7. Reglamento de Inteligencia Artificial: *El Consejo da luz verde definitiva a las primeras normas del mundo en materia de inteligencia artificial* - Consilium (europa.eu). https://www.consilium.europa.eu/es/press/press-releases/2024/05/21/artificial-intelligence-ai-act-council-gives-final-green-light-to-the-first-worldwide-rules-on-ai/

8. Comisión Europea. Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de inteligencia artificial (Ley de Inteligencia Artificia). Pag.3 https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0008.02/DOC_1&format=PDF

TIPOS DE SISTEMAS DE INTELIGENCIA ARTIFICIAL RECOGIDOS EN LA "AI ACT"

1	2	3	4
Sistemas de Inteligencia Artificial de Riesgo inaceptable (Art. 5)	Sistemas de Inteligencia Artificial de Alto Riesgo (HRAIS, Art. 6)	Sistemas de Inteligencia Artificial con obligaciones de transparencia específicas (Art. 52)	Sistemas de Inteligencia Artificial de Riesgo Inexistente o Mínimo
Prohibido · Manipulación del comportamiento, las opiniones y las decisiones humanas. · Clasificación de personas en función de su comportamiento social. · Identificación biométrica masiva a distancia y en tiempo real, salvo ciertas excepciones.	Permitido si se cumplen los requisitos de la IA de la evaluación ex-ante de conformidad · Principales aspectos del reglamento (anexo III). · Regímenes comunes con los que ya están sujetos a una norma armonizada de la UE. · Lista adicional que debe ser revisada cada año por la EAIB (art. 84).	Permitido pero sujeto a obligaciones de información/ transparencia · Interacción con humanos. · Uso para detectar emociones o determinar categorías basadas en datos biométricos. · Generación de contenidos manipulados.	Permitido sin restricciones
Ejemplo: Social scoring	Ejemplo: Contratación	Ejemplo: Personificación (bots)	Ejemplo: Mantenimiento predictivo

FUENTE:Elaboración Propia. T*ipos de sistemas de Inteligencia Artificial recogidos en la "AI Act".*

Conforme a la propuesta del Reglamento del Parlamento serán considerados sistemas de **Riesgo Inaceptable, y por lo tanto prohibidos,** los siguientes sistemas de IA:

a) Sistemas de IA destinados a la **manipulación del comportamiento** que se sirvan de técnicas subliminales que trasciendan la conciencia de una persona o que aproveche alguna de las vulnerabilidades de un grupo específico de personas (debido a su edad o discapacidad física o mental) para **alterar de manera sustancial su comportamiento** de un modo que provoque, o sea probable que provoque, perjuicios físicos o psicológicos a esa persona o a otra.

b) Sistemas de IA diseñados para **clasificar personas físicas en función de su comportamiento social** utilizados por organismos públicos (o en su representación) atendiendo a su conducta social, características personales o de su personalidad (conocidas o predichas).

c) Sistemas de IA de **identificación biométrica masiva remota** «en tiempo real» en espacios de acceso público con fines de aplicación de la ley, salvo en las siguientes excepciones:
   1) la búsqueda selectiva de posibles víctimas concretas de un delito, incluidos menores desaparecidos;
   2) la prevención de una amenaza específica, importante e inminente para la vida o la seguridad física de las personas físicas o de un atentado terrorista;
   3) la detección, la localización, la identificación o el enjuiciamiento de personas físicas.

Cualquier uso concreto de un sistema de identificación biométrica remota estará supeditado a la concesión de una autori-

Cualquier uso concreto de un sistema de identificación biométrica remota estará supeditado a autorización judicial previa

zación previa por parte de una autoridad judicial o una autoridad administrativa independiente del Estado miembro.

Además de los sistemas de IA de Riesgo Inaceptable ya mencionados, es importante **considerar la categoría de IA Generativa.** Estos sistemas pueden ser clasificados como de Riesgo Alto dependiendo de su aplicación y alcance. Por ejemplo, la IA Generativa General puede incluirse en el Grupo 3 (Sistemas de Inteligencia Artificial con obligaciones de transparencia específicas) del diagrama de riesgos anterior, dado que su uso indiscriminado podría tener implicaciones significativas en la privacidad y la autenticidad de la información. Por otro lado, la IA Generativa Fundacional, diseñada con propósitos específicos y controlados, podría requerir una evaluación más detallada para determinar su nivel de riesgo.

Un amplio rango de sistemas de IA de Alto Riesgo estaría autorizado para acceder al mercado de la UE, siempre que cumplan con requisitos y obligaciones establecidos. Estos requisitos han sido refinados por los legisladores para asegurar su aplicabilidad técnica y reducir la carga de cumplimiento para los interesados. Sin embargo, es importante destacar que todos los sistemas de Alto Riesgo requerirán una evaluación *ex-ante.*

Los sistemas de IA Generativa pueden ser clasificados como de Riesgo Alto, dependiendo de su aplicación y alcance

SISTEMAS DE INTELIGENCIA ARTIFICIAL CONSIDERADOS DE ALTO RIESGO[9]	
1. Identificación biométrica y categorización de personas físicas.	Sistemas de IA destinados a utilizarse en la identificación biométrica remota en tiempo real, o en diferido, de personas físicas, conforme lo descrito en la letra d) anterior.
2. Gestión y explotación de infraestructuras críticas.	Modelos de IA vinculados a infraestructuras transportes u otros similares que puedan poner en peligro la vida y la salud de los ciudadanos.
3. Educación y formación profesional	Sistemas de IA que puedan determinar el acceso a la educación y la carrera profesional de una persona.
4. Empleo, gestión de trabajadores y acceso al autoempleo	Sistemas de IA destinados a la contratación –por ejemplo, para anunciar las vacantes, seleccionar o filtrar las solicitudes, evaluar a los candidatos en el curso de las entrevistas o pruebas– así como para tomar decisiones sobre la promoción y la terminación de las relaciones contractuales relacionadas con el trabajo, para la asignación de tareas y para el seguimiento y la evaluación del rendimiento y el comportamiento en el trabajo.
5. Acceso y disfrute de los servicios privados esenciales y de los servicios y prestaciones públicas	Sistemas de IA destinados a ser utilizados por las autoridades o en nombre de ellas para evaluar el derecho a prestaciones y servicios de asistencia pública, así como para conceder, revocar o reclamar dichas prestaciones y servicios.
6. Asuntos relacionados con la aplicación de la ley	Modelos de IA destinados a ser utilizados para realizar evaluaciones de riesgo individuales, u otras predicciones destinadas a ser utilizadas como prueba, o para determinar la fiabilidad de la información proporcionada por una persona con vistas a prevenir, investigar, detectar o perseguir un delito o adoptar medidas que afecten a la libertad personal de un individuo.
7. Gestión de la migración, el asilo y el control de fronteras	Sistemas de IA destinados a ser utilizados para predecir la ocurrencia de delitos o eventos de malestar social con el fin de asignar los recursos dedicados al patrullaje y la vigilancia de los territorios.
8. Administración de justicia y procesos democráticos	Sistemas de IA destinados a ayudar a una autoridad judicial en la investigación e interpretación de hechos y de la ley, así como en la aplicación a un conjunto concreto de hechos.

9. Conforme el listado de Sistemas de IA del Anexo III del Reglamento (ANEXO III - *Sistemas de IA de Alto Riesgo a que se refiere el artículo 6, apartado 2).*

En los casos de uso citados anteriormente, se permite su uso, pero siempre previa sumisión a una **evaluación de conformidad** (antes de que entre en el mercado o se utilice) de cara a acreditar que el sistema cumple con los requisitos exigibles a una IA confiable. En este sentido, el Reglamento establece un procedimiento de emergencia que permite desplegar una herramienta de IA de Alto Riesgo sin la evaluación previa de conformidad, introduciendo mecanismos específicos para garantizar que los derechos fundamentales estarían suficientemente protegidos contra cualquier posible mal uso de dicho sistema.

Por otro lado, la Comisión, en colaboración con los Estados miembros, creará y mantendrá una base de datos para el **registro de sistemas de IA de Alto Riesgo.** El objetivo no es otro que, en caso de infracción, permitir que las autoridades nacionales sean capaces de acceder a la información necesaria para investigar si el uso de la IA fue, o no, conforme con la legislación nacional que resulte de aplicación.

Los **sistemas de IA de alto riesgo** estarán sujetos a obligaciones estrictas antes de que puedan comercializarse; dichas obligaciones son detalladas a continuación:

1. Sistemas adecuados de evaluación y mitigación de riesgos.
2. Alta calidad de los conjuntos de datos.
3. Registro de la actividad para la trazabilidad de los datos.
4. Documentación detallada.
5. Información clara y adecuada al usuario
6. Medidas apropiadas de supervisión humana para minimizar riesgos.
7. Alto nivel de solidez, seguridad y precisión.

El Reglamento sobre Inteligencia Artificial, establece un **régimen sancionador en su artículo 99,** basado en la fijación de umbrales que las autoridades nacionales deberán tener en cuenta en sus procedimientos sancionadores[10]:

a) Incumplimiento relativo a prácticas prohibidas: hasta **35 millones de euros o el 7% del volumen de negocio anual** total a escala mundial del ejercicio financiero anterior;

b) Incumplimiento de cualquier otro requisito u obligación del Reglamento: hasta **15 millones de euros o el 3% del volumen de negocio anual** total a escala mundial del ejercicio financiero anterior;

c) Suministro de información incorrecta, incompleta o engañosa a los organismos notificados y/o autoridades nacionales: hasta **7,5 millones de euros o el 1% del volumen de negocio anual** total a escala mundial del ejercicio financiero anterior.

A efectos comparativos, el Régimen General de Protección de Datos, establece un régimen sancionador menor, dado que las sanciones más graves se sancionan con hasta 20 millones de euros o el 4% del volumen de facturación anual, y de hasta 10 millones de euros o el 2% de la facturación anual, para las menos graves.

A efectos comparativos, el Régimen General de Protección de Datos, establece un régimen sancionador menor

10. Reglamento del Parlamento y del Consejo Europeo sobre normas armonizadas en materia de Inteligencia Artificial [https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=OJ:L_202401689]

Por otro lado, el Reglamento considera un régimen sancionador (por definir) más proporcionado para pequeñas y medianas empresas, así como empresas emergentes, en caso de infracciones a las disposiciones de la Ley de IA.

4. LA ANTICIPACIÓN DE LAS ORGANIZACIONES EMPRESARIALES A LOS MARCOS REGULATORIOS ESPERADOS SOBRE LA INTELIGENCIA ARTIFICIAL.

Las voces sobre la necesidad de desarrollar regulación específica en materia de Inteligencia Artificial no únicamente provienen del sector público, sino también de entes privados. Sundar Pichai, CEO de Google[11], se ha manifestado a favor de la regulación sobre la Inteligencia Artificial, promoviendo una regulación internacional convergente entre la Unión Europea y los Estados Unidos. Los *Principios de Google*[12] *sobre la Inteligencia Artificial* publicados en 2018 son, en esencia, los mismos valores que también vienen impulsando e inspirando la Ley de Inteligencia Artificial que aspira aprobar la Comisión Europea.

El tejido empresarial debe jugar un papel clave en esta fase de desarrollo de regulación, no sólo para contribuir con su perspectiva económica sobre los impactos de la Inteligencia Artificial en la sociedad, sino para adaptar, a su vez, sus procesos y mecanismos internos a este nuevo marco regulatorio en materia de Inteligencia Artificial.

Modelos de Inteligencia Artificial

Uno de los aspectos importantes a la hora de abordar la auditoría de un proceso de negocio que cuente, en alguna de sus fases, con Inteligencia Artificial implementada, es tener un conocimiento técnico mínimo del tipo de modelos de Inteligencia Artificial, con la finalidad de realizar un análisis de riesgo apropiado, de forma que nos permita diseñar un plan de trabajo acorde con los riesgos específicos del proceso e intrínsecos de los sistemas de Inteligencia Artificial utilizados.

En la presente sección, desarrollamos aquellos modelos de Inteligencia Artificial más comunes y utilizados por las organizaciones con

11. Why Google thinks we need to regulate Artificial Intelligence. [Artículo de opinión de Sundar Pichai publicado en el *Financial Times* (20 junio 2020)]
12. Google, AI Principles 2020 update. 2020

el objetivo de proporcionar una base técnica mínima para el desarrollo de los trabajos de revisión de procesos con estructuras de Inteligencia Artificial.

1. ANÁLISIS PREDICTIVO TRADICIONAL.

El **análisis predictivo tradicional** está caracterizado por el uso de métodos y técnicas estadísticas avanzadas que permitan asignar probabilidades de ocurrencia de eventos futuros basándose en datos históricos.

Normalmente, los datos históricos[13] se utilizan para crear un modelo matemático que capture las tendencias importantes. Este modelo predictivo se usa entonces con los datos actuales para predecir lo que pasará a continuación, o bien para sugerir acciones que llevar a cabo con el fin de obtener resultados óptimos.

Este área del análisis predictivo tradicional engloba técnicas clásicas de minería de datos como la regresión lineal/logística, *clustering*, análisis factorial o series temporales[14].

El análisis predictivo utiliza métodos y técnicas estadísticas avanzadas para asignar probabilidades de ocurrencia de eventos futuros basándose en datos históricos

2. INTELIGENCIA ARTIFICIAL Y MACHINE LEARNING.

El **término Inteligencia Artificial (IA)** se refiere comúnmente a la actividad inteligente llevada a cabo por máquinas diseñadas para reproducir las capacidades del cerebro humano por medio de combinaciones de algoritmos, permitiendo percibir el entorno que les rodea y responder de forma similar a un humano. Esto implica capacidad de ejecutar funciones de razonamiento, observación, aprendizaje y resolución de problemas. Es decir, la IA es una máquina que parece humana y que puede imitar el comportamiento de las personas.

Por otro lado, el *Machine Learning* es un tipo de Inteligencia Artificial caracterizado por el aprendizaje de máquinas de conocimientos y comportamientos que serían difíciles de llevar a cabo por el ser humano, llegando a ir incluso, en algunos aspectos, mucho más allá de la inteligencia humana.

Desde la década de los años 60, se ha relacionado el término de *Machine Learning* o Aprendizaje Automático como una rama de la Inteligencia Artificial focalizada en el reconocimiento de patrones y aprendizaje por parte de las computadoras.

Con el paso de los años, esta disciplina fue desarrollándose en otras materias relacionadas con el razonamiento probabilístico, la investigación estadística y, en especial, la profundización en el reconocimiento de patrones relacionados con procesos de ingeniería, matemáticas y computación.

13. La utilización de datos históricos presenta ciertas limitaciones, puesto que, dependiendo de las variables utilizadas en el modelo, el pasado no necesariamente tiende a repetirse en el futuro. No obstante, suelen ser los únicos datos disponibles o los más fácilmente obtenibles en determinadas circunstancias.

14. Ver definición en Glosario de Términos del Anexo II.

Hoy en día, el *Machine Learning* es una disciplina científica del ámbito de la Inteligencia Artificial que tiene como principal objetivo crear sistemas que aprendan automáticamente para que, a posteriori, y en base a ese aprendizaje obtenido por ese sistema o máquina, el mismo sea capaz de resolver un problema dado, con precisión en base al aprendizaje obtenido.

El aprendizaje de las maquinas en este contexto significa adquirir la capacidad de identificar patrones complejos en millones de datos, de ahí la relación tan estrecha del *Machine Learning* con la disciplina del *Big Data* en el presente. Esta combinación de disciplinas empieza a tomar forma y denominarse en algunos círculos como *"Machine Big Data"* cuyo potencial sin límites se basa en los siguientes conceptos:

- Uso de la lógica y la estadística que nos permitan razonar y abordar un problema, realizando un modelo predictivo que nos faculte para dar respuesta al mismo con el fin de proporcionar una solución a un grupo de personas y sus necesidades.
- Utilización del *Big Data* que nos permita realizar un manejo eficiente de los datos, independientemente de su estructura (datos estructurados y/o no estructurados)[15] así como de su tipología (registros históricos, datos recientes o incluso datos recogidos en tiempo real).
- Resolución del problema dado, obteniendo la mejor respuesta al problema mediante el uso de los diferentes algoritmos de *Machine Learning* y eligiendo aquel que se adapta mejor al problema.

Esta realidad hace que el factor clave en todo este proceso sean los datos, ya que el objetivo final que persigue esta actividad consiste en automatizar, mediante algoritmos complejos, aquellos patrones o tendencias que esconden los mismos, y que son difíciles de identificar mediante un análisis tradicional efectuado por cualquier ser humano.

El ámbito de utilización del *Machine Learning* es muy diverso dado que pueden aplicarse a universos de datos de multitud de tipologías. Por tanto, este tipo de técnicas son utilizadas hoy en día en campos como la Medicina, para detectar enfermedades prematuramente; la lucha contra el terrorismo, con objeto de predecir un ataque terrorista; o en el ámbito empresarial, para obtener ventajas competitivas frente a otras empresas.

A continuación, mostramos unos ejemplos de caso de uso para ilustrar los avances de la IA en distintos campos:

1. **Medicina.** Realizar pre-diagnósticos médicos o detectar enfermedades de forma precoz. Un sistema basado en algoritmos de *Machine Learning* puede aprender del historial médico de pacientes ligados a diagnósticos previos correctos, con el objetivo de aprender a identificar futuros pacientes enfermos, y de esta forma, ayudar al personal médico en la toma de decisiones ante la aparición de síntomas de enfermedades. Por ejemplo, en el año 2016 se desarrolló un algoritmo de *Machine Learning* basado en aprendizaje supervisado (concepto explicado más adelante) capaz de detectar si una persona sufre depresión analizando sus publicaciones en la red social Instagram.

Machine Learning es una disciplina de la IA: crea sistemas que aprenden automáticamente para, después, resolver un problema dado

15. Ver definición en Glosario de Términos del Anexo II.

2. **Lucha contra el terrorismo.** Actualmente se han desarrollado softwares basados en la aplicación de *Machine Learning* capaces de identificar patrones ante un posible atentado terrorista, mediante la extracción y combinación de información existente en diferentes medios de internet como redes sociales. También se han desarrollado sistemas capaces de monitorizar y detectar transacciones ligadas a grupos terroristas.

3. **TI.** También se desarrollan estas técnicas en el ámbito de TI con vistas a mejorar las infraestructuras tecnológicas ya establecidas. En este ámbito, se pueden destacar la identificación de correo electrónico no deseado, técnicas de reconocimiento de voz, detectar intrusiones en una red de comunicaciones de datos, predecir fallos en equipos tecnológicos o, incluso, modificar el funcionamiento o aspecto de una aplicación móvil para que se adapte de forma personalizada a las costumbres y necesidades de cada usuario

4. **Ámbito empresarial.** Las últimas tendencias del uso de *Machine Learning* en el ámbito empresarial están muy ligadas a los departamentos de RR.HH. y Marketing Digital, donde se están comenzando a crear modelos predictivos que permitan determinar, por ejemplo, qué empleados serán los más productivos en los próximos años o identificar clientes potenciales mediante la identificación de patrones basados en sus comportamientos en las redes sociales. Cabe destacar también, dentro de las organizaciones financieras, el uso de este tipo de herramientas para la detección de fraude, con el fin de mitigar el mismo y reducir su impacto económico en este tipo de organizaciones. Por último, y más directamente relacionado con el ámbito empresarial, su uso en la denominada Industria 4.0, optimizando las operaciones mediante la anticipación a picos y valles de demanda; mejorando el mantenimiento predictivo de las instalaciones y su fiabilidad; o consiguiendo ahorros de costes a lo largo de la cadena de suministros.

Las últimas tendencias del uso de *Machine Learning* en el ámbito empresarial están muy ligadas a RRHH y Marketing Digital

5. **La aplicación de la IA Generativa en el ámbito empresarial** ha sido impulsada en los últimos años por importantes empresas tecnológicas como Microsoft, Amazon, IBM y Google, desarrollando soluciones que han mejorado la eficiencia, productividad y creatividad en diversas áreas de trabajo. Algunos de los ejemplos son los siguientes:

   A) **Microsoft Copilot** impulsado por el modelo de lenguaje GPT-4, es un chatbot que ofrece funciones avanzadas de generación de código, integrándose en aplicaciones de Microsoft como Visual Studio Code y Word para mejorar la experiencia de desarrollo.

   B) **Amazon Q** integrado en la infraestructura de AWS, está diseñado para abordar necesidades empresariales específicas, proporcionando soluciones y asistencia en tiempo real dentro de la consola de administración de AWS, con acceso a diversos modelos de IA para respuestas más precisas.

   C) **IBM WatsonX** es una plataforma integral de datos e IA, ofrece herramientas para el desarrollo de soluciones personalizadas, un almacén de datos eficiente y un kit de herramientas para la gobernanza de la IA.

3. TIPOS DE ALGORITMOS DE MACHINE LEARNING: APRENDIZAJE SUPERVISADO, APRENDIZAJE NO SUPERVISADO Y APRENDIZAJE POR REFUERZO. REDES NEURONALES. INTELIGENCIA ARTIFICIAL GENERATIVA.

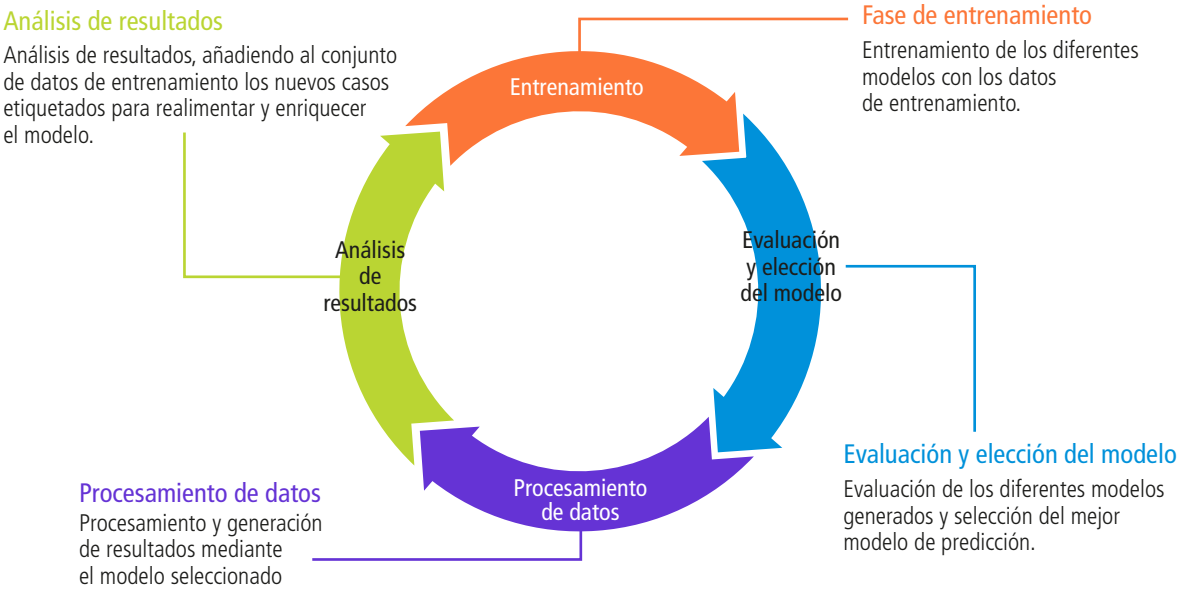
Dentro de la disciplina del *Machine Learning* podemos destacar sus dos principales tipos de aprendizaje, el aprendizaje supervisado y el aprendizaje no supervisado.

Aprendizaje supervisado

En el aprendizaje supervisado se realizan predicciones basándose en patrones, comportamientos o características que ya se han visto en datos históricos y etiquetados. Es decir, existe un conocimiento previo de los datos.

Estos algoritmos tienen la capacidad de predecir el valor de un conjunto de datos, tras ser entrenados con otro conjunto de datos suficientemente amplio, donde la variable objetivo ya ha sido etiquetada. A partir de estos datos, donde su etiqueta ya es conocida, se obtienen las relaciones entre la misma y el resto de las variables del modelo. Es decir, se establecen una serie de reglas o patrones, en función de un universo de datos conocido, que serán aplicables para realizar una predicción sobre nuevos datos.

MEJORA CONTINUA DEL MODELO BASADO EN APRENDIZAJE SUPERVISADO



FUENTE: Elaboración propia.

Algunos de los algoritmos que se suelen utilizar en el aprendizaje supervisado son los Árboles de Decisión, *Gradient Boosting, Random Forest,* SVM *(Support Vector Machines)* o *Naive Bayes*[16].

16. Ver definición en Glosario de Términos del Anexo II.

Entre algunos de los casos más comunes del aprendizaje supervisado, destacan, por ejemplo; su aplicación para la determinación del *scoring* de clientes en entornos financieros; el mantenimiento predictivo en el ámbito industrial; la detección de enfermedades; la ciberseguridad, etc.

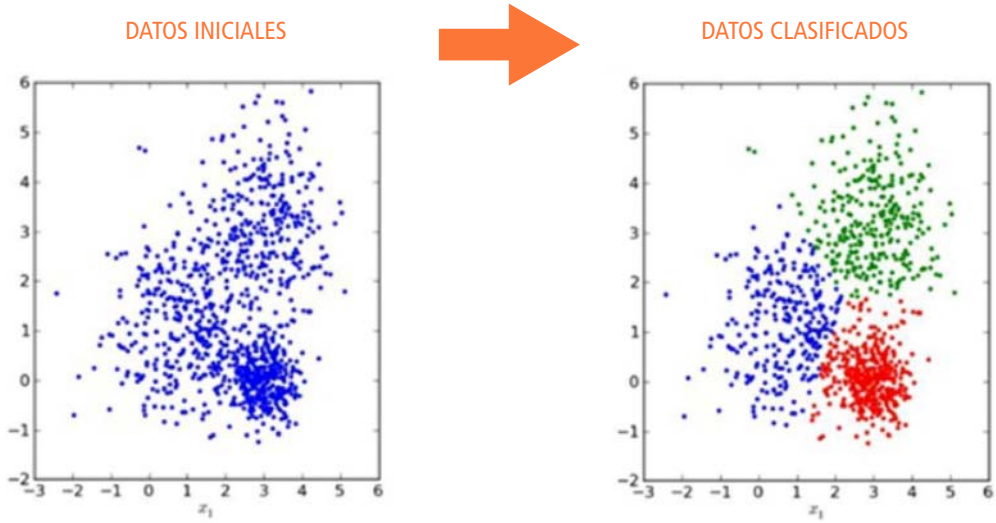
Aprendizaje no supervisado

Con los algoritmos de aprendizaje no supervisado se produce conocimiento, únicamente, por los datos que se proporcionan como entrada. Es decir, no existe un conocimiento previo de los datos.

Este tipo de algoritmos tienen la capacidad de autoorganizarse en grupos, en función de sus características comunes. En estos casos, se exploran los datos buscando relaciones en los mismos para estructurarlos u organizarlos en función de sus características.

Algunos de los algoritmos más comunes que engloba el aprendizaje no supervisado son el *K-Means, Isolation Forest* o las redes neuronales[17].

CLASIFICACIÓN DE DATOS EN ENTORNOS NO SUPERVISADOS



FUENTE: Elaboración propia.

La capacidad que nos otorgan las máquinas en el presente está permitiendo avanzar a gran velocidad en el uso de este tipo de técnicas, rompiendo con las restricciones o limitaciones que existían en su uso hasta hace solo unos pocos años.

Alguna aplicación de este tipo de métodos podría ser el reconocimiento de imagen y voz, las predicciones médicas, la predicción de información anómala, etc.

17. Ver definición en Glosario de Términos del Anexo II.

Estos avances en técnicas como redes neuronales (desarrolladas a finales de los años 50) únicamente han sido posible utilizarlas con los últimos avances tecnológicos, permitiendo avanzar en la subdisciplina del *Machine Learning* conocida como ***Deep Learning.*** La base principal del *Deep Learning* consiste en la utilización de técnicas avanzadas de procesamiento que se aproximen a la percepción humana, mediante unidades de proceso o neuronas artificiales especializadas en detectar características existentes en los datos u objetos recibidos.

Aprendizaje por refuerzo

Además de los dos principales tipos de algoritmos que engloban la disciplina del *Machine Learning* (supervisado y no supervisado) encontramos otro tipo de aprendizaje denominado aprendizaje por refuerzo o *Reinforcement Learning.*

Su característica principal destaca por la no necesidad de grandes cantidades de datos para poder realizar su entrenamiento. El aprendizaje por refuerzo se basa en un sistema de prueba *vs.* error basado en recompensas, que permite reforzar el comportamiento deseado.

Su funcionamiento se fundamenta en una exploración del entorno sobre el que se toman decisiones en las que el algoritmo ("agente inteligente") recibe "recompensas" o "penalizaciones". De esta forma, los algoritmos de *Reinforcement Learning* aprenden y redefinen su estrategia de actuación, iterando el número de veces necesario hasta encontrar la estrategia que conduzca al resultado óptimo.

Aunque resulta intuitivo de comprender, hay tres grandes **retos en el aprendizaje por refuerzo:**

1) Si el "agente inteligente" no produce un comportamiento suficientemente diverso, tiene riesgo de producir el espejismo de que no hay mejor forma para conseguir una recompensa mayor y quedarse estancado en una solución no óptima.

2) En segundo lugar, el "agente inteligente" debe encontrar un equilibrio entre comportarse por las políticas que está aprendiendo y la exploración de nuevas estrategias que podrían dar mejores resultados. Estas exploraciones pueden reducir la recompensa, o incluso llegar a sustituir comportamientos previamente aprendidos. Los humanos nos enfrentamos con frecuencia a estas decisiones: ¿cenamos en uno de nuestros restaurantes habituales o nos arriesgamos en el nuevo restaurante de moda?

3) El tercer reto es que la recompensa puede llegar mucho tiempo después de las acciones. Por ejemplo, en ajedrez un movimiento en medio de la partida puede condicionar completamente el desenlace.

A pesar de estas dificultades, el aprendizaje por refuerzo se está usando con éxito en distintos ámbitos donde los aprendizajes supervisados no tenían éxito por la dificultad de disponer de datos suficientes y donde los algoritmos de aprendizaje no supervisado simplemente no son adecuados para el problema.

El aprendizaje por refuerzo se usa, por ejemplo, para el desarrollo de agentes de *trading* algorítmico, donde existen una gran cantidad

La principal característica del *Reinforcement Learning* es que no necesita grandes cantidades de datos para realizar su entrenamiento

de datos históricos disponibles que sirven para hacer el *backtest* y generar las recompensas de manera algorítmica. También en otros problemas como entrenamiento de *chatbots*, automatización de navegación web autónoma para fines de *testing* o *scraping*.

Otros ámbitos donde la recompensa se puede obtener directamente del entorno son los videojuegos, donde sus propias reglas sirven para dar esa recompensa o penalización a las acciones realizadas. Asimismo, pueden darse situaciones asimilables en el mundo real, donde el buen o mal resultado se pueda medir de forma sistemática, como son los casos de la conducción autónoma, de la recomendación en comercio electrónico, y en servicios de *streaming*, o para optimizar la publicidad que se muestra a los visitantes de servicios online, entre otros usos.

Las redes neuronales

Son una aproximación a la Inteligencia Artificial mediante la cual, una entidad pretende imitar un sistema nervioso, basado en neuronas, por medio de algoritmos "bio-inspirados". La unidad fundamental de una red neuronal es la neurona, que a su vez se organizan típicamente en capas. Cada neurona recibe unas entradas de la capa anterior que combina mediante una operación matemática sencilla y las envía a la siguiente capa. En la primera capa, la de entrada, cada neurona recibe un dato de la entrada a procesar, por ejemplo, un píxel de una foto; y en la capa más exterior produce la salida, que típicamente es un valor numérico y que para cada uso representará algo diferente (una predicción de un valor, o la siguiente palabra para generar un texto, o el valor de un pixel en una imagen, etc.). Aquellas capas que se encuentran entre la entrada y la salida se denominan capas ocultas.

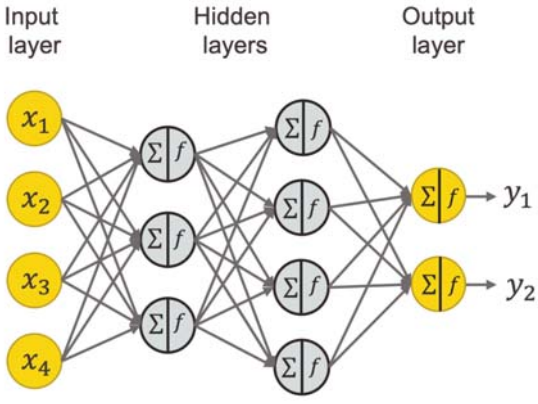
Los parámetros de cada neurona se ajustan en un proceso de entrenamiento denominado *back propagation* (propagación hacia atrás), que es computacionalmente costoso, pero que produce buenos resultados en múltiples problemas complejos.

Las redes neuronales son algoritmos cuya "explicabilidad" es muy baja: es difícil o incluso imposible explicar de forma directa por qué una red neuronal produce un resultado.

Cuando el número de capas ocultas es muy grande se le denomina redes neuronales profundas y a su entrenamiento se le llama normalmente en inglés ***deep learning*** **(aprendizaje profundo).** Cuantas más capas tiene una red neuronal, más compleja resulta, más difícil es explicar su comportamiento y computacionalmente es más demandante.

La capacidad e idoneidad de resolver problemas de las redes neuronales está muy ligada a diferentes factores: a la forma en que estas capas están dispuestas entre sí; cómo están conectadas; qué funciones matemáticas im-

EJEMPLO DE RED NEURONAL SENCILLA CON 2 CAPAS OCULTAS



FUENTE: https://www.knime.com/blog/a-friendly-introduction-to-deep-neural-networks

plementan; qué dimensiones tienen, y cómo se organizan entre sí.

Por lo que hay muchos tipos de arquitecturas diferentes para distintos tipos de redes, y la comunidad de práctica de Inteligencia Artificial comparte, en gran medida, sus conocimientos, sus hallazgos y sus modelos[18] entrenados, de forma que otras personas puedan apoyarse en ellos para realizar sus trabajos y mejoras.

Una forma de trabajar bastante optimizada para enfrentarse a un problema consiste en usar como punto de partida un modelo previamente entrenado para una tarea (por ejemplo, generar texto) y entrenarlo de manera específica para matizar su comportamiento (por ejemplo, generar textos jurídicos, o generar textos médicos). Mientras que el sistema base aportaría el conocimiento básico del lenguaje (como son las construcciones gramaticales, concordancia, etc.), el entrenamiento especializado aportaría conocimiento específico. A esto se le conoce como ***transfer learning*** **(aprendizaje transferido).**

Si bien el *transfer learning* ha demostrado ser muy útil para múltiples casos de uso relacionados con el procesamiento de textos, generación de lenguaje, reconocimiento y generación de imágenes, etc.; el uso de este tipo de técnicas comporta el riesgo de heredar los sesgos existentes en el modelo base. El sesgo algorítmico ocurre cuando un sistema informático refleja los valores de los humanos que están implicados en la codificación y recolección de datos usados para entrenar el algoritmo, tal y como se explica más adelante en este documento.

Inteligencia Artificial Generativa

En los últimos años han proliferado los llamados **Modelos de Lenguaje Grande** ***(Large Language Models,*** "LLM", por sus siglas en inglés). Además de poder simular conversaciones con un elevado nivel de complejidad y conocimiento, las capacidades que tienen los LLM para procesar textos han superado ampliamente a las de otros enfoques en la creación y generación de resúmenes de textos, identificación de sentimientos o palabras clave, entre otras funcionalidades.

Los Modelos de Lenguaje Grande utilizados por la IA Generativa tienen su origen en una arquitectura similar a las redes neuronales profundas

Los LLMs tienen su origen en una arquitectura denominada *Transformers*[19], que son similares a las redes neuronales profundas, pero que tienen la capacidad de extraer información estadística sobre cómo unas palabras se relacionan con otras a partir de textos y, con el suficiente entrenamiento y datos, los textos generados por estos sistemas parecen estar dotados de inteligencia. Los *Transformers* tienen unas características técnicas que les permiten ejecutarse en paralelo en potentes computadores y GPUs con mayor eficiencia que otras redes neuronales. Además, la forma de entrenar estos *Transformers* con nuevos textos se lleva a cabo simplemente eliminando palabras en textos existentes y entrenándolos para predecir las palabras que han sido suprimidas, o bien prediciendo cuáles serán las siguientes palabras para el texto proporcionado. Esto permite generar conjuntos de entrenamiento de forma automatizada, lo que es una gran ventaja sobre otros sistemas que

18. Un ejemplo de repositorio público de modelos es https://huggingface.co/models
19. Cornell University - Attention is al you need: [1706.03762] Attention Is All You Need (arxiv.org)

requieren el etiquetado por parte de expertos. Dado que los *Transformers* se pueden paralelizar de forma muy eficiente para entrenarse y ejecutarse en hardware extremadamente potente, y que los datos de entrenamiento se pueden crear de manera automática a partir de textos existentes, ha sido posible entrenar enormes modelos de lenguaje, con decenas o cientos de billones de parámetros, usando trillones de palabras extraídas de toda la información pública disponible en Internet (como el conjunto de datos Common Crawl)[20].

La capacidad de la IA Generativa para procesar y generar lenguaje humano de manera eficaz, facilita su adopción en entornos empresariales para transformar procesos

El punto de inflexión que ha permitido democratizar el uso de los LLM ha sido combinar estas características con unas interfaces de usuario tipo "chat", que permiten hablar con ellos para hacerles preguntas, como si se tratara de personas al otro lado del terminal. Esto se ha conseguido combinando sus propias características con técnicas de aprendizaje por refuerzo, para las que se ha usado conjuntos de preguntas y respuestas elaboradas por humanos de manera específica, y las valoraciones de las respuestas que pueden hacer los usuarios después de cada interacción.

Gracias a la capacidad de los LLMS para procesar y generar lenguaje humano de manera eficaz, se están adoptando en entornos empresariales para transformar procesos. Algunos de los casos de uso más relevantes serían:

- **Automatización del servicio al cliente.** Los LLM se utilizan para potenciar los chatbots y asistentes virtuales, proporcionando respuestas rápidas y precisas a las consultas de los clientes. Esto mejora la experiencia del cliente al ofrecer un servicio disponible las 24 horas del día, los 7 días de la semana, al tiempo que reduce la carga de trabajo sobre el personal humano.
- **Análisis de texto y datos.** Los modelos pueden analizar grandes volúmenes de texto, como correos electrónicos, documentos legales, o *feedback* de clientes, para extraer *insights,* tendencias y patrones significativos. Esto apoya la toma de decisiones basada en datos y ayuda a identificar áreas de mejora en productos y servicios.
- **Generación de contenido.** Los LLM pueden generar contenido escrito de alta calidad, como informes, correos electrónicos, artículos de marketing y más, ahorrando tiempo y recursos. Esto permite a las empresas mantener una presencia activa y profesional en múltiples plataformas con menor esfuerzo.
- **Entrenamiento y desarrollo.** Los LLM se emplean para crear simulaciones y escenarios de entrenamiento personalizados que ayudan en el desarrollo de habilidades del personal. Pueden facilitar escenarios interactivos para la capacitación en servicio al cliente, ventas, y más, adaptándose a las respuestas de los usuarios para ofrecer una experiencia de aprendizaje más efectiva.
- **Optimización de procesos internos.** Los modelos pueden ayudar a automatizar y optimizar una variedad de tareas administrativas y repetitivas, como la entrada de datos, la programación de reuniones, y la gestión de correos electrónicos. Esto libera al personal para concentrarse en tareas de mayor valor y mejora la eficiencia operativa.

20. Cornell University - Language Models are Few-Shot Learners: [2005.14165] Language Models are Few-Shot Learners (arxiv.org)

Pero la rápida adopción para aprovechar las ventajas que proporciona este tipo de Inteligencia Artificial también lleva asociada la asunción de nuevos riesgos. Además de los problemas inherentemente técnicos, existen otros debates relevantes, como aquellos relacionados con la propiedad intelectual de los contenidos con los que han sido entrenados, dado que los LLM generan contenidos basados en las relaciones estadísticas de esos contenidos en alguna proporción. También es importante considerar el ángulo de la responsabilidad. Si se emplean este tipo de sistemas para realizar recomendaciones en ámbitos como la salud, la fiscalidad, el entorno legal, etc. ¿quién es responsable de las recomendaciones que emitan y de sus resultados de aplicación?

Los modelos de *machine learning* pueden heredar los problemas y dificultades que acechan al software tradicional, añadiendo otros propios

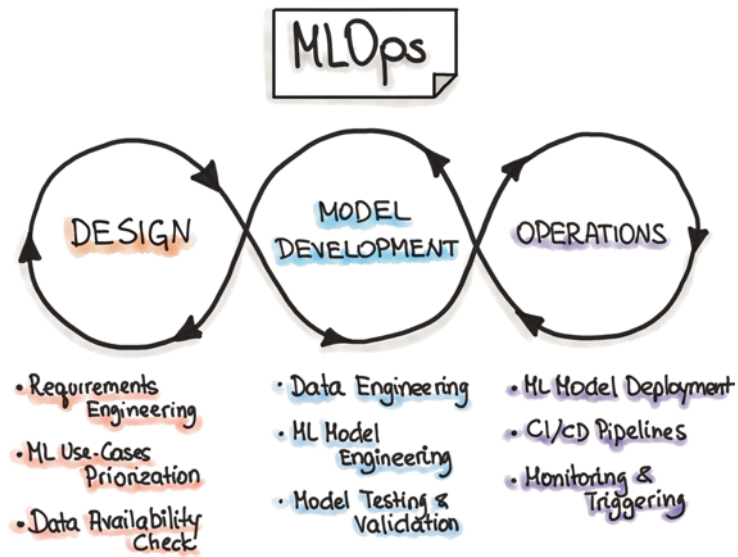
4.MLOps COMO RESPUESTA A LAS NECESIDADES DE ADAPTACIÓN.

De forma análoga a como ha ocurrido con otras disciplinas como el software, el *machine learning* se ha visto acelerado en los últimos años. La necesidad de hacer frente a nuevos retos como son las necesidades cambiantes y la recogida automatizada de flujos de información, se yuxtaponen a la forma tradicional de trabajo en la que tanto modelos como datos, eran seleccionados y tratados de forma cuasi artesanal.

Cabe destacar que el desarrollo de modelos de *machine learning* es susceptible de heredar los problemas y dificultades que acechan también a la ingeniería de software tradicional, añadiendo además los propios problemas que trae consigo la Inteligencia Artificial y el aprendizaje automatizado.

Como respuesta a esta necesidad de dinamismo surgen las prácticas de **MLOps**[21], acrónimo de ***Machine Learning Operations.*** Podría decirse que MLOps es al *machine learning* lo que DevOps es al desarrollo de software en general: siendo un conjunto de prácticas que pretenden agilizar el ciclo de vida de las soluciones de Inteligencia Artificial, mediante la automatización de los procesos de desarrollo, entrenamiento y despliegue de modelos, integrando aspectos de data, desarrollo, infraestructura y seguridad, minimizando los tiempos de despliegue y mantenimiento, así como posibles errores.

EL PROCESO ITERATIVO INCREMENTAL DE ML-OPS



FUENTE: https://Ml-ops.org

21. MLOps, al igual que DEVOps, se trata de una evolución de la metodología AGILE para el desarrollo de software; metodología que ha sido incorporada recientemente y de manera exitosa a los trabajos de auditoría interna.

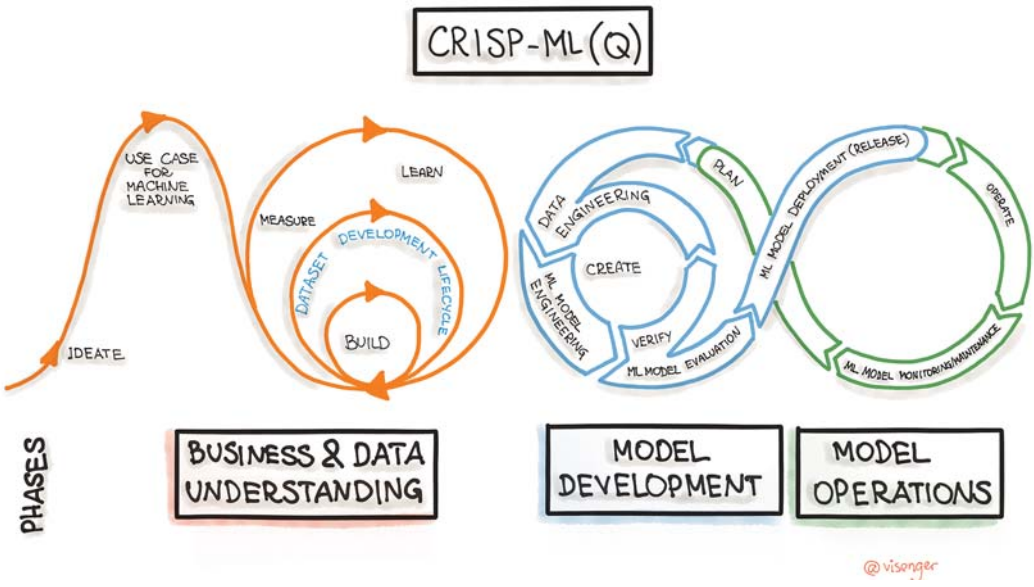
Algunas de las actividades y prácticas que recoge MLOps serían las listadas a continuación:

- Desarrollo de pruebas específicas para validar el modelo de aprendizaje automático.
- Integración continua *(continuous integration)* para probar y validar tanto los modelos como sus datos.
- Entrega continua *(continuous delivery)* con la finalidad de automatizar la puesta en marcha de los nuevos modelos.
- Entrenamiento continuo *(continuous training)* con el objetivo de garantizar un aprendizaje continuo en el tiempo.
- Monitorización de los modelos en diferentes métricas de precisión, especificidad, impacto de disparidad, etc. para detectar derivas no esperadas de los modelos de IA, poder analizarlas de manera temprana y, en su caso, aplicar las correcciones necesarias.

En lo relativo al desarrollo del modelo, al igual que ocurre con el software, no hay una metodología que sobresalga de las demás en todos los aspectos y para todas las situaciones, debiendo recurrirse en la mayoría de las ocasiones, a aproximaciones *ad-hoc* para los distintos problemas.

No obstante, y con el objetivo de definir un punto de partida genérico, se ha propuesto la **metodología CRISP-ML(Q)**[22] *(Cross-Industry Standard Process for development of Machine Learning applications with Quality Assurance),* que alinea aspectos de negocio con el desarrollo y la operación de los modelos con la finalidad de asegurar la mayor calidad posible de los modelos desarrollados y que se ajusten a las expectativas de los promotores.

CICLO DE VIDA DEL DESARROLLO DE MACHINE LEARNING SEGÚN LA METODOLOGÍA CRISP-ML(Q)



FUENTE: https:/ml-ops.org

22. Studer, S.; Bui, T.B.; Drescher, C.; Hanuschkin, A.; Winkler, L.; Peters, S.; Mu̲ller, K.-R.; Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology. Preprints 2021, 1, 0. https://doi.org/. Accesible en: https://arxiv.org/pdf/2003.05155.pdf

Las evidencias generadas como parte del proceso que define la metodología deberían ser de interés para el auditor, de cara a la revisión de las actividades realizadas para garantizar el buen funcionamiento y el alineamiento con el negocio.

5. ARQUITECTURA DE DATOS Y TI

Los estándares de la industria aconsejan que toda solución basada en Inteligencia Artificial venga acompañada de una arquitectura que permita su automatización y facilite su uso y explotación. Desde un punto de vista de control interno, es muy relevante conocer la arquitectura de datos y TI de los sistemas de IA objeto de auditoría.

A estos efectos, este apartado tiene como objetivo desarrollar una arquitectura modelo de sistemas de IA, cuyos detalles técnicos deben ser analizados y entendidos por Auditoría Interna, a la hora de afrontar trabajos de revisión de estructuras de IA.

Si definimos una arquitectura desde el punto de vista funcional, podemos encontrarnos las siguientes capas o *layers:*

A) **Data Governance o Modelo de Gobierno del Dato.**

B) **Origen de Datos.**

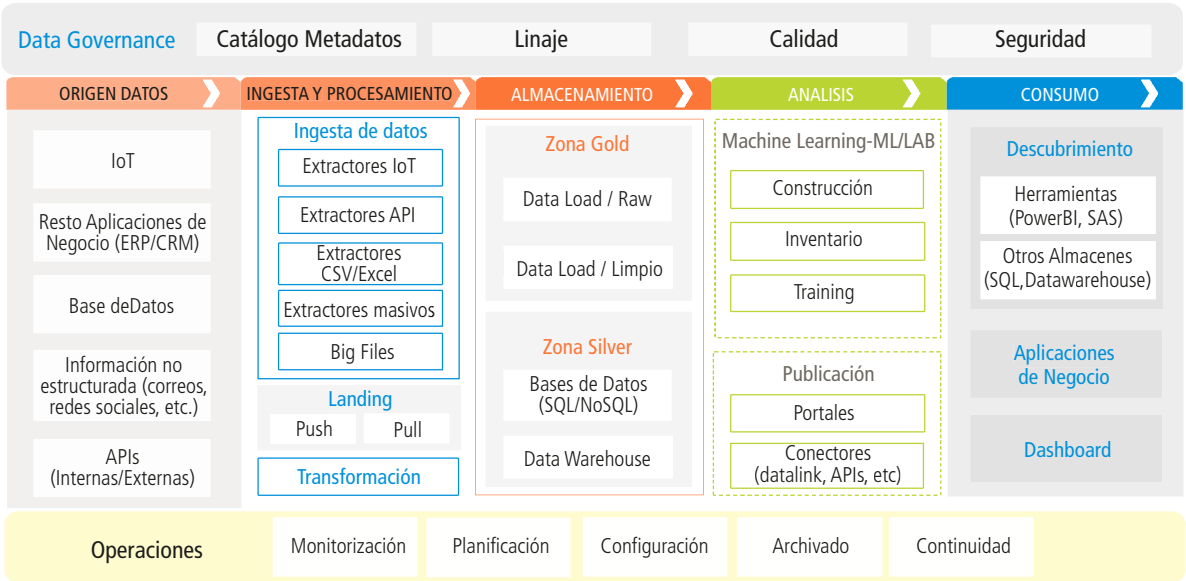
C) **Plataforma de Gestión de Datos.**

- C.1) Ingesta de datos y procesamiento.
- C.2) Almacenamiento.
- C.3) Análisis (Machine Learning - ML/LAB).

D) **Consumo de datos.**

E) **Operaciones.**

ARQUITECTURA DE DATOS Y TI



FUENTE: Elaboración Propia

Desde un punto de vista de control interno, es crítico conocer la arquitectura de datos y TI de los sistemas de IA objeto de auditoría

A. Data Governance o Modelo de Gobierno del Dato

Este dominio suele estar compuesto por diferentes componentes con el fin de cubrir las principales disciplinas del gobierno del dato:

- **Catálogo Metadatos.** Datos disponibles de forma unificada y sus metadatos.
- **Linaje.** Trazabilidad de los datos desde su origen hasta sus diferentes transformaciones.
- **Seguridad:**
  - **Permisos:** control de acceso a la información.
  - **Cumplimiento normativo aplicable:** aplicación y chequeo del cumplimiento con las necesidades requeridas por diferentes normativas en términos de uso y protección de los datos, como por ejemplo RGPD y PCI-DSS *(Payment Card Industry Security Standards Council),* entre otras normativas aplicables.
- **Calidad del dato.** Conjunto de reglas y métricas destinadas a evaluar la calidad de la información disponible.

B. Orígenes de datos

Las fuentes de datos que alimentan los sistemas de IA pueden ser múltiples, considerando datos financieros y no financieros, de distintos sistemas contables (ERP - *Enterprise Resource Planning)* y/u otras aplicaciones de negocio de las organizaciones tanto internas como externas. En este sentido, se hace especialmente relevante un conocimiento preciso de las bases de datos que alimentan los modelos de IA, para un entendimiento adecuado del origen de los datos, su estructura (información estructurada o no estructurada) y cómo estos alimentan la capa de Plataforma de Gestión de Datos explicada a continuación.

C. Plataforma de Gestión de Datos

Las Plataformas de Gestión de Datos habituales de los sistemas de IA pueden presentar los siguientes cuatro dominios que describimos a continuación:

c.1) **Ingesta de datos y procesamiento.**

Capa lógica formada por el conjunto de componentes responsables del acceso a los sistemas de IA de los distintos orígenes de datos. Para que la arquitectura sea lo más automatizable posible, es deseable que esta capa sea lo más completa posible, conteniendo el mayor número de extractores y herramientas de carga necesarios, bien sea con orígenes de datos internos a la organización o externos, vía APIs[23] o conexiones directas. Dispone de motores de procesamiento de alto rendimiento que transforman los datos depositados en la *Landing Zone,* zona de entrada de datos en modalidad de persistencia generados por los extractores *(pull)* o depositados por procesos externos *(push).*

c.2) **Almacenamiento.**

Se trata de la zona *core* de toda plataforma, estando compuesta por diferentes tecnologías y capacidades de almacenamiento, formando varias capas lógicas de datos y motores de procesamiento que transicionan los datos entre ellas.

23. Ver definición en Glosario de Términos del Anexo II.

Desde el punto de vista de almacenamiento destacan las siguientes zonas:

- ***Gold Zone:*** Repositorio donde se almacenan y procesan los datos uno a uno, sin necesidad de normalizar ni aplicar reglas de negocio.
- ***Silver Zone:*** Zona de acceso y proceso de datos, donde se aplican transformaciones y se aplican reglas de negocio, normalizaciones, etc. De cara a la explotación final, la información se puede estructurar según convenga.

c.3) **Laboratorio de Machine Learning (ML-LAB).**

En soluciones orientadas a analítica avanzada, cobra especial interés la zona de *Machine Learning,* proporcionando capacidades, tales como entorno de desarrollo, librerías, repositorios de código abierto, etc., que facilitan el diseño, entrenamiento, *testing,* gobierno y entrega de modelos, formando entre ellos el *Laboratorio de Machine Learning* (ML-Lab). En determinadas arquitecturas, y en función de su criticidad, podría ser necesario disponer de un *sandbox,* entorno aislado que permite ejecutar pruebas de forma segura sin comprometer el resto de la arquitectura.

D. Consumo de datos

Esta última capa representaría el *output* de los datos como resultado de las ejecuciones realizadas sobre los datos habiendo transcurrido las capas y dominios anteriores. Representaría, por tanto, el resultado de los modelos de Inteligencia Artificial, para su análisis mediante herramientas de visualización o *big data,* almacenamiento o utilización por otras aplicaciones de negocio.

6. ARQUITECTURA GLOBAL EN MODELOS DE IA GENERATIVA.

Al tratarse de una tecnología disruptiva, todas las organizaciones se encuentran actualmente construyendo y enriqueciendo su arquitectura partiendo de una inicial que sirva de referencia. Estas nuevas arquitecturas, vienen integrándose como parte del ecosistema de los entornos de TI de cada organización, donde los modelos de Inteligencia Artificial se implementan y crecen de forma estructurada.

En este sentido, para poder dar una solución confiable y que cumpla con los estándares esperados de seguridad y gobierno, se están empezando a diseñar arquitecturas que contemplen los **cuatro flujos principales en cualquier solución de IA Generativa:**

- **Extracción e ingesta** de información estructurada y no estructurada.
- Generación de *chunks* o "troceado" de datos ***(chunking).***
- Vectorización o ***embedding*** de los datos de entrada.
- **Interpretación del *prompt*** del usuario y respuesta generativa.

En toda arquitectura que preste servicio a casos de uso de IA Generativa, hay tres activi-

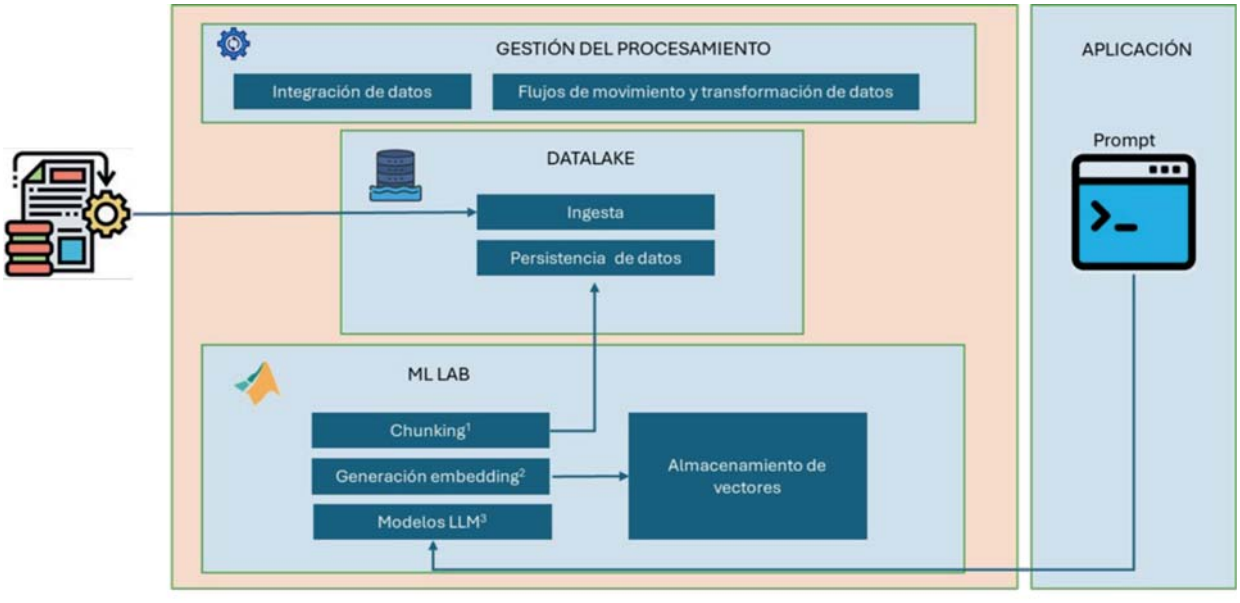
Las nuevas arquitecturas de la IA Generativa se integran en el ecosistema de los entornos de TI de cada organización

dades que son clave a la hora de disponer de un entorno eficiente y que cubra las expectativas y valor esperado:

- ***Chunking* o fragmentado de datos:** es una técnica que permite combinar o agrupar múltiples unidades de información en un número limitado de fragmentos, de forma que sea más fácil procesar y recordar la información. Esta técnica intenta asimilar los métodos de almacenamiento al funcionamiento de la memoria de las personas, ya que la mayoría pueden recordar listas de cinco palabras durante 30 segundos, por lo que, al tener una lista de más palabras, éstas se dividen en fragmentos más pequeños con el fin de mejorar el rendimiento.
- ***Embedding:*** se trata de una técnica de procesamiento de lenguaje natural que convierte el lenguaje humano en vectores matemáticos, lo que permite que las computadoras procesen el lenguaje de manera más efectiva al tratar las palabras como datos.
- ***Prompt:*** un *prompt* es una instrucción o texto inicial que se le proporciona a una herramienta de IA generativa para guiar su generación de respuestas o resultados, según los formatos en los que se especialice la herramienta.

  El *prompt* funciona como una "entrada de información" con la cual el usuario le especifica el contexto y la tarea que se espera que la herramienta complete. Al proporcionar un *prompt,* se condiciona al modelo de IA para que genere una "salida" (resultado) coherente y relevante, en función de lo que el usuario necesita.

ARQUITECTURA GLOBAL IA GENERATIVA



¹Chunking es el proceso de dividir un contenido en partes más pequeñas y manejables
²Procesamiento de lenguaje natural a través conversión de lenguaje humano en vectores matemáticos
³Modelo lingüístico de gran tamaño que facilita la compresión y generación de lenguaje humano

FUENTE: Elaboración Propia

El consumo energético de la IA Generativa

Es importante tener en cuenta que, a día de hoy, detrás de la IA Generativa y de los LLM, en general, se produce un consumo energético importante, llegando a alcanzar cifras realmente sorprendentes. No cabe duda de que será uno de los grandes retos a la hora de optimizar esta tecnología y reducir su gran peso para la huella de carbono, para lo que serán relevantes otras tecnologías como las energías renovables.

Para hacernos una pequeña idea, un modelo pequeño, diez veces más pequeño que, por ejemplo, chatGPT, tiene un consumo de energía de 650.000 kWh solo para entrenarse, mientras que en 2022 el consumo per cápita en España fue de 5.214kWh. En esta misma línea, el entrenamiento de *Llama*[24] necesitó 2.638 MWh, lo que equivaldría al consumo de un frigorífico de 150 watios durante 2.008 años.

El consumo energético de la IA Generativa será uno de los grandes retos de esta tecnología, de cara a reducir su peso en la huella de carbono

Marco de control interno y riesgos de los procesos de negocio con Inteligencia Artificial

Las organizaciones que inician su andadura en el despliegue de sistemas de IA deben establecer un marco y unas estructuras de control interno adecuados para la identificación y mitigación de los riesgos asociados.

Una adecuada evaluación del desempeño de los sistemas de IA debe contar con aspectos para la monitorización continua de los riesgos de este tipo de tecnologías, así como con estructuras de control interno que garanticen la supervisión continua de los propios sistemas de IA. A modo ilustrativo, se sugieren a continuación una serie de actividades clave que deberían formar parte del marco de control interno de los sistemas de IA desplegados en las organizaciones.

24. *Llama* son una familia de modelos de inteligencia artificial generativa desarrollados por la empresa Meta, que los ha publicado para que puedan ser descargados y usados libremente incluso con fines comerciales. Más información: https://llama.meta.com

MONITORIZACIÓN CONTINUA DE RIESGOS	SUPERVISIÓN DE LA ARQUITECTURA DE DATOS Y TI	REVISIÓN DE LOS MODELOS DE IA	SUPERVISIÓN DE LA IMPLEMENTACIÓN	MONITORIZACIÓN TRAS LA PUESTA EN PRODUCCIÓN
Identificación de riesgos generales por el despliegue de sistemas de IA, y *risk-assessment* específico en función de la complejidad de los algoritmos y objetivos perseguidos con cada modelo de IA	• Actividades definidas para garantizar el acceso, tratamiento, privacidad, protección y destrucción de los datos, especialmente en aquellos en modelos predictivos sobre comportamiento humano. • Proporcionar seguridad razonable sobre los sistemas de TI y prevención de ciber-ataques.	• Asegurar un entendimiento adecuado de la operatividad de los algoritmos y *output* esperado. • Definición de métricas e indicadores de desempeño para la monitorización continua. • Revisión del comportamiento de los modelos en fases tempranas del despliegue de los motores.	• Desarrollar las pruebas de implementación suficientes para garantizar que el despliegue de los motores atiende los objetivos esperados. • Aprobaciones pertinentes de los Comités de Proyecto establecidos previo al *go-live*.	• Revisión periódica de las métricas e indicadores de desempeño. • Mecanismos de control para la identificación de comportamientos anómalos de desempeño, incluyendo las acciones correctivas necesarias a los algoritmos.

Se propone la utilización del Marco Integrado de Control Interno emitido en 2013 por el *Committee of Sponsoring Organizations of the Treadway Commission* (COSO), que es la referencia más habitual de mejores prácticas en esta materia, y permite desarrollar un enfoque eficaz y eficiente para diseñar e implementar el marco de control para evaluar y gestionar los riesgos asociados con los sistemas de Inteligencia Artificial, en base a sus 17 principios y 5 componentes de control interno. La función de Auditoría Interna debe proporcionar aseguramiento independiente sobre los riesgos, la gobernanza y los controles de la IA. Debe evaluar las políticas y procedimientos relacionados con la IA implementados, verificando que los objetivos de control sean adecuados y funcionen según lo diseñado. A continuación, desarrollamos los 5 componentes COSO, desde la óptica del marco de control interno en los modelos de Inteligencia Artificial.

1. ENTORNO DE CONTROL (GOBIERNO Y CULTURA)

COSO indica como objetivos principales del marco de control, establecer buenas prácticas de gobierno y fortalecer la rendición de cuentas, responsabilidad y supervisión. Adicionalmente, es relevante entender el perfil de riesgo de la IA en la Compañía y conocer cómo se encuentra gestionando los riesgos asociados.

Conforme al *Marco Integrado de Control Interno* de COSO, se debe establecer un conjunto de normas, procesos y estructuras que constituyan las bases sobre las que se desarrolle el control interno en todos los niveles de la organización. El Consejo y la Alta Dirección son quienes establecen el *"Tone at the top"* sobre la importancia del control interno, incluidos los estándares de conducta que se consideran aceptables.

Por otro lado, fomentar en la Compañía su compromiso con la integridad y los valores éticos, sociales y legales, ayuda a garantizar que las actividades de IA y las decisiones y acciones relacionadas con los sistemas de IA, sean coherentes con los valores y las responsabilidades éticas, sociales y legales de la organización.

Un adecuado modelo de Gobierno sobre los aspectos de IA debe tener en consideración los siguientes aspectos:

- Los **volúmenes ingentes de datos** utilizados por los modelos de IA; es por ello, que se hace relevante contar con un sistema de gobernanza de los datos a lo largo de su vida (creación, transformación, uso y destrucción de los datos como inputs de los modelos).

- La **cultura empresarial debe estar ligada a los componentes éticos** de los sistemas de IA; es por ello que el modelo de gobierno debe garantizar que los valores éticos de las organizaciones y sus políticas internas se extienden a su vez, a los objetivos perseguidos por las estructuras de IA. En este sentido se considera necesaria la existencia de un Código Ético, donde se refleje el respeto de los derechos humanos, la protección de los datos personales, el fomento de la igualdad de oportunidades, la transparencia y la libertad de elección, así como las necesidades y expectativas de las personas.

- El modelo de gobierno de los sistemas de IA debe, también, considerar el **cumplimiento regulatorio** aplicable en cada jurisdicción donde operan las organizaciones.

- La existencia de una **cultura consciente del riesgo,** del control, de las políticas, procesos y estructuras que guían a las personas en todos los niveles en el desempeño de sus responsabilidades de una manera que sea consistente con el compromiso de la entidad con la **integridad y los valores éticos.**

- El establecimiento de unos **niveles de competencia adecuados** para la gestión y supervisión de los sistemas de Inteligencia Artificial. Es importante destacar que el Consejo de Administración y el Comité de Dirección deben de conocer los conceptos clave de los modelos de Inteligencia Artificial.

- **La IA Generativa añade un componente adicional de amplificación de riesgos,** que el modelo de gobierno definido debe tener en cuenta tanto en su diseño, como en su mantenimiento y mejora continua de las estructuras de control interno.

Desde el punto de vista de los **principales órganos de gobierno (Consejo de Administración / Comisión de Auditoría)** el papel más relevante dentro del marco del control interno de sistemas de IA se resume a continuación:

- Asesorar sobre si la estrategia considera adecuadamente las **amenazas y oportunidades de los sistemas de IA.** Sería aconsejable contar con un **Comité de IA** específico, o asignar responsabilidades a uno ya existente, para la supervisión y gestión de los riesgos de los modelos de Inteligencia Artificial, donde se asuman responsabilidades correspondientes en este ámbito.

- Impulsar un programa de **capacitación en conocimientos y habilidades** específicas de los sistemas de IA, así como adoptar políticas y prácticas para atraer, desarrollar y retener a profesionales competentes.

- **Identificar los riesgos** de los sistemas de Inteligencia Artificial, e incorporarlos dentro de los modelos de riesgos de la organi-

El modelo de gobierno debe garantizar que los valores éticos y las políticas internas se extienden a la IA

zación, definiendo el apetito de riesgo para este tipo de riesgos.

- Supervisar y evaluar, con la periodicidad y profundidad que se estime oportuna, la **efectividad del control interno del modelo de Gobierno** sobre todos los aspectos de los modelos de IA, tanto desde un punto de vista del diseño, como de la operatividad de las estructuras de control interno implementadas.

La **dirección ejecutiva,** como extensión de los objetivos generales de los órganos de gobierno anteriores, establece la estrategia de implementación de iniciativas de IA, incluyendo los objetivos específicos y *road-map* de despliegue de los sistemas de IA prioritarios para la compañía. Adicionalmente, tanto el Consejo como la dirección ejecutiva, deben establecer con claridad la estrategia sobre el diseño y la implementación de sistemas de IA alineada con la estrategia general de la organización, considerándose esta estrategia como elemento crítico para el éxito de las iniciativas en modelos de IA que las empresas deseen emprender.

Los objetivos que llevan a las empresas a invertir en sistemas de IA deben estar también alineados con la estrategia y los objetivos empresariales

De la misma forma, existen un amplio abanico de objetivos que motivan a las empresas a la utilización de sistemas de IA, desde la optimización y reducción de costes, hasta la diferenciación de los productos y servicios que ofrecen a sus clientes, generando, de esta forma, fuentes de ingresos con mayor diversificación. En cualquiera de sus facetas, los objetivos específicos que llevan a las empresas a invertir en sistemas de IA deben encontrarse, a su vez, alineados con los objetivos empresariales a corto y largo plazo.

ENTORNO DE CONTROL (GOBIERNO Y CULTURA): MEJORES PRÁCTICAS Y RECOMENDACIONES
Existencia de una estrategia definiendo los principales órganos de gobierno para la implementación de sistemas de IA, iniciándose en aquellos modelos de mayor retorno e incrementando la inversión a medida que el *know-how* y *expertise* interno generado asegure el atendimiento de los objetivos establecidos.
La estrategia en la implementación de modelos de IA debe contar con las premisas generales de medición de desempeño de los modelos y, por extensión, de atendimiento de los objetivos perseguidos con el despliegue de los modelos de IA.
Plan estratégico de supervisión y evaluación del modelo de Gobierno para la mejora continuada del entorno de control y su reporte los órganos de administración correspondientes.

2. EVALUACIÓN DE RIESGOS

Conforme a lo que indica el *Marco de Control Interno* de COSO, la compañía debe definir unos objetivos con suficiente claridad para permitir la identificación y evaluación de los riesgos relacionados con la IA. En este sentido, un adecuado marco de control interno diseñado e implementado para abordar los riesgos derivados de la utilización de sistemas de IA debe considerar los siguientes aspectos:

- Una **definición clara y precisa de los objetivos** que permitan la identificación y evaluación de los riesgos expuestos derivados de la implementación de sistemas de IA.
- La evaluación de los riesgos debe abarcar todos los niveles de la entidad, fomentando un ***risk-assessment* transversal** y coordinado entre todas las áreas implicadas (fundamentalmente aquellas con mayor responsa-

bilidad en la implementación y ejecución de los sistemas de IA), así como estableciendo las bases de cómo los riesgos deben ser gestionados y abordados para su mitigación eficiente.

- Las organizaciones deben contar con **mecanismos de evaluación de riesgos** para identificar y evaluar cambios significativos en los modelos de IA que pudieran afectar significativamente las estructuras de control interno.

En la siguiente tabla se describen los principales riesgos vinculados a los sistemas de IA desarrollados por las organizaciones:

RIESGOS	COMENTARIOS
Riesgos de Gobierno	Relacionados con las estructuras internas de las organizaciones, con las políticas, metodologías y la toma de decisiones en los procesos, incluyendo la supervisión a alto nivel. Se debe enfatizar en los riesgos que puedan impactar en la gestión y liderazgo, en la independencia en la toma de decisiones, en el impulso a la transparencia y en la rendición de cuentas.
Riesgos Operacionales y/o de Negocio	Relacionados con los diversos puntos del ciclo de vida del desarrollo e implementación de un sistema de IA. Los más significativos son los que puedan conllevar errores de procesamiento, riesgos de los datos o riesgos en desviaciones, sesgo en resultados o representatividad de los datos.
Riesgos Financieros	Relacionados con la contabilidad de las operaciones y la presentación de la información financiera. Es decir, se deben contemplar las situaciones donde la IA pueda impactar en la información financiera que presenta la compañía, o en los resultados económicos-financieros de la misma.
Riesgos Regulatorios	Vinculados con las áreas legales y de cumplimiento de las regulaciones. Cabe destacar los riesgos de cumplimiento asociados con las regulaciones externas (p.ej. RGPD) o internas (p.ej. Código Ético). Estos riesgos están relacionados con las actividades de los modelos IA, así como las decisiones y acciones relacionados con ella, sean coherentes con los valores y las responsabilidades éticas, sociales y legales de la Compañía.
Riesgos Tecnológicos y de Ciberseguridad	Asociados a los sistemas y la ciberseguridad de los modelos desarrollados de IA. Por ejemplo, se debe evaluar si esos modelos pueden contener datos personales que corren el riesgo de ser accedidos y utilizados por terceros no autorizados.
Riesgo Reputacional	Relacionado con aquellos riesgos derivados por la presencia de sesgos en los modelos, o sanciones impuestas por incumplimiento normativo. También por la exposición a riesgos externos generados por terceros (ver debajo).
Riesgo de Sostenibilidad	El consumo de energía que permite la operatividad y funcionamiento de los sistemas de Inteligencia Artificial (por ejemplo, los modelos de IA Generativa) puede tener un impacto en los modelos y políticas de sostenibilidad de las empresas, por ejemplo, en lo relacionado con los compromisos adquiridos de reducción de gases de efecto invernadero, eficiencia energética o huella de carbono.
Riesgos intrínsecos de los modelos de IA	**Utilización de datos:** Cuando se trabaja en un proyecto relacionado con la Inteligencia Artificial el error más común suele ser utilizar datos no íntegros o inexactos. Normalmente los algoritmos pueden estar alimentados con datos estructurados y/o no estructurados de distinta procedencia, como pueden ser webs, imágenes, redes sociales, etc. Cualquiera utilización de bases de datos errónea puede provocar resultados de los algoritmos inestables y/o incorrectos. La selección incorrecta de datos no solo influye en los resultados del algoritmo, sino que también puede tener repercusiones éticas, ya que se puede estar dejando de lado algún grupo de información, dando una imagen de un modelo discriminatorio o sesgado. **Desarrollo inadecuado de algoritmos de IA que provocan resultados inapropiados:** Errores en la programación y desarrollo de código de los algoritmos de IA, pueden provocar resultados no esperados, inadecuados o alejados de los objetivos establecidos. La fase de desarrollo de implementación de los algoritmos de modelos IA resulta ser la más crítica, especialmente en aquellos sistemas de IA con una mayor sofisticación, por lo que cualquier error de código o programación puede llevar a resultados inapropiados.

RIESGOS	COMENTARIOS
Riesgos intrínsecos de los modelos de IA	**Incapacidad de interpretación o interpretación incorrecta de los outputs de los modelos de IA:** El mayor valor añadido de la utilización de sistemas de IA son los resultados que se obtienen de dichos modelos, no obstante, en ocasiones podemos encontrarnos ante la incapacidad de interpretar o interpretar de forma incorrecta los resultados que se obtienen una vez las bases de datos son escaneados por los algoritmos de IA, tomando a cabo decisiones (o no) con resultados no deseados o inesperados. A modo de ejemplo, algoritmos de Inteligencia Artificial basados en redes neuronales puede contener con mayor profundidad esta tipología de riesgos intrínsecos, debido a que los algoritmos de redes neuronales tienden a presentar un mayor riesgo de incapacidad o interpretar incorrectamente el output obtenido de los sistemas de IA. La IA Generativa crea nuevos riesgos y, a la vez, amplifica otros riesgos existentes. **Riesgos específicos de los modelos de Inteligencia Artificial Generativa y modelos LLM:** • *Hallucination* **(Desvíos de la realidad):** los modelos LLM, al estar basados en sistemas probabilísticos, predicen palabras para un texto previamente facilitado. En ese sentido, los modelos LLM pueden inventar contenido que no existe (por ejemplo, citas a documentos que no existen, o incluso generar un documento formal que no existe), que sea falso (afirmaciones que no reflejan la verdad) o que sea erróneo (proporciona soluciones equivocadas a problemas). • **Relevancia:** en ocasiones el contenido puede no resultar relevante para el contexto. Por ejemplo, porque podría preguntársele por el dolor de una muñeca, y responderte en relación con una muñeca de juguete, y no con la parte anatómica. • **Toxicidad:** al haber sido entrenado con una enorme cantidad de contenido en Internet, el sistema ha aprendido de textos que reflejan odio, envidia, insultos, racismo, violencia, etc. (entre otros sesgos). • **Privacidad y confidencialidad:** dado que se están ofreciendo servicios en la nube que emplean los contenidos y conversaciones proporcionados por los usuarios para entrenar los sistemas, existe el riesgo de que filtre información privada a usuarios diferentes, porque lo ha aprendido durante el uso. • *Jailbreaking:* es posible forzar a que los modelos lleven a cabo acciones para las que no han sido diseñados. Incluso sortear mecanismos de prevención y salvaguardas mediante entradas especialmente manipuladas.
Riesgos externos por la utilización de la IA Generativa por terceros	La capacidad de la IA Generativa para crear información errónea, ataques de *phishing* y ataques de *malware* cada vez más sofisticados supone un riesgo creciente sobre la gestión del riesgo de ciberseguridad y la privacidad de datos personales y estratégicos de las compañías. En relación con lo anterior, también aumenta el riesgo de que la información que utiliza una compañía en la ingesta de sus procesos y algoritmos de IA pueda verse contaminada por acciones deliberadas de terceros. Adicionalmente, existe una tendencia creciente de las compañías a externalizar la operación de determinados procesos en *vendors,* incluyendo la provisión de datos personales y estratégicos, con lo que existe una mayor exposición a la cadena de suministro en este sentido. Esto es así porque aplican los mismos riesgos relativos al uso de IA y las compañías deben protegerse mediante la configuración de un *mix* de aseguramiento (ej. obtención de un SOC 2 del auditor del *vendor* y establecimiento de controles internos que complementen ese control externo). Lógicamente, todo lo anterior puede tener un impacto directo sobre el riesgo reputacional.

EVALUACIÓN DE RIESGOS: MEJORES PRÁCTICAS Y RECOMENDACIONES

Existencia de políticas internas de gestión de riesgos, específicas para los sistemas de IA; con la finalidad de que toda la compañía conozca y se implique en la identificación de riesgos relacionados con modelos de IA, de una manera continua. Se deben incluir directrices generales para el diseño e implementación de estrategias de mitigación de riesgos.

Existencia de un Inventario o Mapa de Riesgos con la naturaleza de cada uno de los riesgos identificados, incluyendo su criticidad, y cuando posible, cuantificación de la probabilidad de ocurrencia o potencial impacto financiero, en los sistemas de TI participantes en el proceso, así como otras categorías de riesgos.

3. ACTIVIDADES DE CONTROL

Conforme a lo indicado por COSO, las actividades de control se definen como las acciones establecidas, a través de las políticas y procedimientos, que contribuyen a garantizar que se lleven a cabo las instrucciones de la dirección establecidas para mitigar los riesgos con impacto potencial en los objetivos definidos en la implementación de los modelos de IA.

En este sentido, las organizaciones deben contar con mecanismos de comunicación internos apropiados para expandir y dar a conocer los objetivos y responsabilidades de cada área, así como los responsables involucrados en los modelos de IA desplegados. De igual forma, las organizaciones deben contar con estructuras de control apropiadas que permitan trasladar de forma adecuada los aspectos relacionados con los modelos de IA, a través de los canales de comunicación externos dirigidos, por ejemplo, a reguladores, accionistas y otros grupos de interés.

Más adelante en este documento, se presenta un Programa de Trabajo donde se contempla una serie de actividades de control ilustrativas para el abordaje de los riesgos relevantes, y que permiten prevenir la materialización de riesgos innecesarios y/o la minimización del impacto de las consecuencias de estos.

ACTIVIDADES DE CONTROL: MEJORES PRÁCTICAS Y RECOMENDACIONES
Existencia de procedimientos internos y/o matrices de riesgos / controles, identificando el diseño de las actividades de control, y sus atributos clave; incluyendo la documentación soporte que evidencie la efectividad de los controles, así como los responsables de su ejecución y revisión para el abordaje *end-to-end* de los riesgos en procesos de negocio con sistemas de IA implementados.
El diseño de las actividades de control considera tanto la implementación como los controles recurrentes en sistemas de IA; siendo estos últimos evaluados de forma periódica para identificar riesgos no abordados por las correspondientes actividades de control interno, así como cambios en el diseño de los controles necesarios durante el ciclo de vida de los sistemas de IA.
Diseño y establecimiento de actividades oportunas de "revisión humana" de los comportamientos y resultados de los algoritmos de IA, para asegurar que éstos reflejan el objetivo original y, además, se utilizan de manera legal, ética y responsable.
En relación con el punto anterior, también es una buena práctica el diseño e implementación de alertas y/o indicadores de desviaciones respecto a los objetivos iniciales de los algoritmos de IA.
La compañía mantiene un inventario de sistemas de IA, identifica sinergias y analiza riesgos desde un punto de vista individual y consolidado.
Riesgo de "cadena de suministro" *(supply chain risk).* Es importante tener en cuenta que la existencia de actividades de control externalizadas no exime a la compañía de su responsabilidad última sobre los riesgos que gestionan, por lo que debe de contar con los correspondientes controles internos sobre las actividades externalizadas. Por ejemplo, la evaluación del riesgo de deficiencias de control que reporta el auditor del *Vendor* en su informe ISAE/SOC/SSAE, y el diseño e implementación de actividades de control claves o compensatorias, en caso de que sea necesario; aparte del seguimiento de la remediación, por parte del *Vendor*, de las deficiencias reportadas.

4. INFORMACIÓN Y COMUNICACIÓN

La información generada por los sistemas de IA debe contar con protocolos de comunicación y reporte adecuados

La información generada por los sistemas de IA debe contar con protocolos de comunicación y reporte adecuados, tanto a nivel interno como externo. La sensibilidad de los datos tratados por las estructuras de IA y los impactos en la sociedad que se derivan de la utilización de IA en los procesos de negocio, hacen muy relevante una comunicación transparente y precisa de cómo las organizaciones deben transmitir los principios generales de esta tecnología.

Actualmente, muchas organizaciones pioneras en estas tecnologías hacen públicos los avances adquiridos, no solo como ventaja competitiva en los sectores en los que operan, sino también para dar a conocer, de manera transparente, los principios que gobiernan los valores éticos y morales derivados de la utilización de sistemas de Inteligencia Artificial.

INFORMACIÓN Y COMUNICACIÓN: MEJORES PRÁCTICAS Y RECOMENDACIONES
La organización publica sus mejores prácticas sobre valores éticos y morales en la utilización de sistemas de IA.
Compartir con la opinión pública los principios sobre Inteligencia Artificial, limitando aquellos aspectos que preocupan más a organismos públicos y privados.
Los máximos responsables de la compañía, accionistas y Consejo de Administración son informados de los aspectos relevantes del avance, desempeño real, y de las iniciativas sobre sistemas de IA alcanzados.
Los procedimientos escritos y/o la matriz de riesgos / controles *(end-to-end* de los procesos de negocio asistidos por técnicas de IA) son conocidos y aplicados en la práctica por los correspondientes dueños de los procesos, y se actualizan en función de la etapa de madurez del modelo de IA.
Existencia de Planes de Emergencia para abordar acontecimientos imprevistos derivados del comportamiento no esperado de los modelos.

5. SUPERVISIÓN Y EVALUACIÓN

Las actividades de supervisión y evaluación *(monitoring)* constituyen evaluaciones periódicas o continuas para verificar que cada uno de los cinco componentes del control interno, incluyendo los controles que afectan a los principios dentro de cada componente de COSO, están tanto correctamente diseñados como operando adecuadamente, incluyendo el nivel de precisión establecido en cada caso.

Los modelos de gobierno enfocados a una gestión empresarial adecuada de los sistemas de IA deben incluir actividades de revisión y evaluación continua y/o independientes, con el objeto de determinar si los componentes del sistema de control interno se encuentran presentes y están en funcionamiento durante el ciclo de vida de estos sistemas de IA.

Las compañías deben contar con mecanismos de evaluación de deficiencias de control interno de los sistemas de IA, así como mecanismos de comunicación de esas deficiencias a los responsables de implementar los planes

de acción o medidas correctivas correspondientes, incluyendo, según corresponda, a la Alta Dirección y al Consejo.

En este sentido, la función de Auditoría Interna toma un papel relevante para garantizar que las actividades de supervisión y evaluación continua son las adecuadas para mitigar los riesgos intrínsecos en los modelos de IA, tanto desde un punto de vista de diseño como de efectividad operativa.

SUPERVISIÓN Y EVALUACIÓN: MEJORES PRÁCTICAS Y RECOMENDACIONES
Planificación y ejecución, por parte de Auditoría Interna, de las actividades de supervisión y evaluación de los modelos, así como validación de la idoneidad de los procesos de identificación de riesgos.
Definición de un plan a largo plazo (o estratégico) de auditoría de los modelos de IA, que acompañe la estrategia de la compañía a este respecto.
Definición de pruebas sustantivas o de auditoría del control interno sobre la integridad, precisión y confiabilidad de los datos sobre los que se construyen los algoritmos de IA.
Identificación, evaluación y comunicación, por parte de Auditoría Interna, de las deficiencias de control interno de manera oportuna a los responsables de tomar medidas correctivas en la compañía, incluida la alta dirección y el órgano de administración (o su Comisión de Auditoría), según corresponda.
Supervisión de las actividades de control interno diseñadas y ejecutadas en relación con el esquema de control de los procesos externalizados en terceras partes o *Vendors (supply chain risk).*

6. ROL DE AUDITORÍA INTERNA

Auditoría Interna es experta en evaluar y comprender los riesgos y oportunidades relacionados con la capacidad de una compañía para cumplir con sus objetivos estratégicos, entre ellos aquellos destinados al despliegue de sistemas de IA. Aprovechando esto, los equipos de Auditoría Interna deben ayudar a la compañía a evaluar, comprender y comunicar el grado en que los algoritmos de IA tendrían un efecto (negativo o positivo) sobre la capacidad de la organización para crear valor en el corto, medio o largo plazo.

Auditoría Interna puede participar mediante, al menos, seis actividades críticas distintas, respecto a los procesos afectados por algoritmos de Inteligencia Artificial:

- Incluir los aspectos relevantes de Inteligencia Artificial en su evaluación de la gestión de riesgos relevantes, así como considerar en su Plan de Auditoría, basado en riesgos, la evaluación del diseño e implementación real conforme el modelo de Gobierno de los modelos de Inteligencia Artificial diseñado por la organización.
- Participar activamente en proyectos de IA desde sus comienzos, a través de la ejecución de auditorías de diseño sistemáticas, a medida que estos proyectos van evolucionando, asegurándose así que las deficiencias de control de diseño son reportadas en tiempo y forma, manteniendo tanto la independencia como su objetividad, ya que Auditoría Interna no es responsable de la implementación de procesos, políticas o procedimientos sobre el despliegue de modelos de Inteligencia Artificial.
- Proporcionar aseguramiento o evaluar la gestión de los riesgos relacionados con la

Auditoría Interna debe ayudar a evaluar, comprender y comunicar el efecto que los algoritmos tendrían sobre la capacidad de crear valor

Entre otros, Auditoría Interna aporta aseguramiento sobre la gestión de los riesgos relacionados con la confiabilidad de los datos y los algoritmos

confiabilidad de los algoritmos subyacentes y los datos en los que se basan los algoritmos de IA.

- Asegurar, dentro de las actividades de Auditoría Interna anteriores, que existen (y operan efectivamente) controles internos destinados a identificar asuntos generados por los algoritmos de IA que puedan afectar al Código Ético de las organizaciones.
- Evaluar tanto el diseño como la operatividad de las estructuras de control interno diseñadas e implementadas, como resultado de la aplicación del modelo de Gobierno *end-to-end* establecido en la compañía.
- Supervisar el cumplimiento de la normativa relacionada con ESG, así como las estructuras de control interno diseñadas para conseguir los objetivos publicados en materia de reducción del uso de recursos durante su ciclo de vida, y sobre la eficiencia desde el punto de vista energético de los modelos de inteligencia artificial.
- Reportar al Consejo de Administración (o Comisión de Auditoría) y a la Alta Dirección los principales resultados de sus evaluaciones de riesgos en materia de IA, así como las deficiencias de control de diseño u operación relevantes identificadas en sus auditorías, y recomendar mejores prácticas de modelo de gobierno en base a los hallazgos reportados y otros riesgos identificados.

Programa de trabajo ilustrativo para la auditoría del control interno de la Inteligencia Artificial aplicada en procesos de negocio

1. ESTRATEGIA DE AUDITORÍA PARA SISTEMAS DE INTELIGENCIA ARTIFICIAL

Se expone a continuación, a modo de guía, un programa de trabajo orientado a garantizar unas estructuras de control interno apropiadas en procesos de negocio que cuenten con sistemas de Inteligencia Artificial implementados.

Este programa de trabajo está diseñado para asegurar la correcta mitigación de los riesgos expuestos por la implementación y ejecución de sistemas de IA. Las pruebas contenidas en el programa pretenden ser válidas, sea cual sea el modelo en particular a auditar. Sin em-

bargo, la propia ejecución de estas pruebas dependerá de la complejidad y sofisticación de los algoritmos de Inteligencia Artificial utilizados, así como de otras circunstancias que arrojen riesgos intrínsecos en cualquier fase del proceso de negocio objetivo de revisión.

Antes de abordar este programa de trabajo, el auditor a cargo de un trabajo que implique la revisión de algún modelo de Inteligencia Artificial debe evaluar el riesgo de cumplimiento regulatorio asociado a dicho modelo. Este riesgo será distinto, en función del objeto final del modelo, así como de los datos que hace uso. Habrá que valorar, entre otras cosas, si el sistema diseñado hace uso de datos de carácter personal o no; si está incrustado en un proceso de decisión de *pricing*; si permite realizar operaciones sobre mercado de instrumentos cotizados o si afecta, por ejemplo, a la propia elaboración de los estados financieros.

Este análisis previo permitirá diseñar pruebas adicionales más allá del modelo, que permitan evaluar si existe un control adecuado para mitigar estos riesgos.

Adicionalmente, ha de tenerse en cuenta la existencia de otros modelos de control sobre los procesos en los que opera el modelo de Inteligencia Artificial. Estos modelos de control estarán cubriendo riesgos adicionales, más allá de los riesgos inherentes a la propia operativa del modelo. El auditor debe evaluar qué elementos de dicho sistema de control cubren aspectos relacionados con el modelo de Inteligencia Artificial. Con este análisis se pretende no introducir duplicidades a la hora de hacer el trabajo de auditoría y tampoco dejar espacios sin cubrir.

El programa detallado a continuación da cobertura a elementos relacionados con el *governance* del propio modelo de Inteligencia Artificial, atacando aspectos más técnicos relacionados con la arquitectura de los datos y la infraestructura utilizadas. Adicionalmente, existen pruebas destinadas a evaluar los controles sobre los datos que se usan como fuente de información. Por supuesto, existe un bloque amplio que permite evaluar el propio desempeño del modelo de IA interna, también cuando este se presenta como una caja negra *(black box).*

En los apartados siguientes se describen los 6 ámbitos de control interno esperados que aborden los riesgos intrínsecos de sistemas de IA en procesos de negocio, incluyendo aquellos **objetivos y/o actividades de control relevantes** que debieran formar parte de las estructuras de control interno de los procesos de negocio que cuenten con modelos de Inteligencia Artificial implementados.

Asimismo, se contemplan los **procedimientos de auditoría** sugeridos para cada objetivo o actividades de control. A continuación, enumeramos los mencionados 6 ámbitos de control interno y las secciones siguientes donde se desarrollan en profundidad:

1. Modelo de Gobierno de sistemas de Inteligencia Artificial.
2. Arquitectura de datos y sistemas de TI.
3. Calidad de los datos.
4. Medición del desempeño.
5. El factor "Caja Negra" *(Black Box)* en los sistemas de IA.
6. El factor humano y el sesgo algorítmico.

El auditor debe evaluar qué elementos del sistema de control cubren aspectos relacionados con el modelo de Inteligencia Artificial

2. MODELO DE GOBIERNO DE SISTEMAS DE INTELIGENCIA ARTIFICIAL

Las organizaciones deben contar con un adecuado modelo de gobierno de los sistemas de Inteligencia Artificial, de forma que las estructuras de control interno, los procesos, procedimientos y políticas implementados para dirigir, gestionar y monitorizar estos sistemas sean acordes a los riesgos derivados de su utilización.

El diseño del modelo de gobierno de sistemas de Inteligencia Artificial debería abordar los siguientes aspectos:

1. Establecer con claridad las **áreas de la organización y los responsables** del diseño, implementación, mantenimiento y monitorización de las estrategias y sistemas de Inteligencia Artificial.
2. Garantizar que la organización cuenta con la **experiencia y conocimiento necesario que los sistemas de Inteligencia Artificial** necesitan para su despliegue y mantenimiento adecuado, especialmente en aquellos modelos de Inteligencia Artificial más sofisticados y complejos.
3. Asegurar que las decisiones sobre los casos de uso y objetivos perseguidos con la implementación de los modelos de Inteligencia Artificial son consistentes con **los valores éticos y políticas internas de la organización,** así como asegurar el adecuado **cumplimiento con la regulación y normativas externas** aplicables.
4. **Evaluar/contrastar la existencia de riesgos** y la definición e implementación de las estrategias adoptadas para su abordaje.
5. Medición y gestión del **presupuesto y recursos necesarios,** así como un **análisis ROI** *(return of investment)* de la implementación de modelos de IA.

OBJETIVOS DE CONTROL O ACTIVIDADES DE CONTROL INTERNO GENERALES	PROCEDIMIENTOS DE AUDITORÍA INTERNA[25]
Implementación de un modelo de gobierno adecuado a la complejidad y riesgos de los sistemas de IA utilizados, desde un punto de vista de diseño y de operatividad del modelo de gobierno, incluyendo análisis ROI durante la fase de diseño de modelos de IA.	• **Revisión de la estructura organizativa y modelo de gobierno.** Determinar si el diseño del modelo de gobierno es adecuado y/o suficiente, y si opera tal y como fue diseñado, así como coincidente con los valores éticos y otras políticas internas de la organización. • **Revisión de los análisis ROI realizados,** con las metodologías de cálculo y justificativa de conclusión sobre la implementación de modelos de IA.
Existencia de una adecuada **segregación de funciones** sobre el sistema de IA.	• **Revisión de una matriz de segregación de funciones** que manifieste las responsabilidades y funciones conflictivas sobre el uso y gestión de la IA

25. Auditoría Interna tendrá que determinar si el enfoque de estos procedimientos se debe realizar **desde una perspectiva eminentemente sustantiva o de auditoría de control interno.** En el segundo de los casos, es necesario ampliar el programa de trabajo para establecer los atributos necesarios que cada objetivo de control debe cubrir, y verificar con documentación soporte interna de la Compañía, que existe documentación que haga evidente la actividad de control interno realmente ejecutada. Adicionalmente, es necesario analizar y discernir en las conclusiones de Auditoría Interna si las deficiencias de control identificadas suponen una deficiencia en la operatividad del control interno o en el diseño de la actividad de control.

Definición de los principales roles y responsabilidades relativas al desarrollo y gestión de entornos avanzados que incluyen modelos de Inteligencia Artificial, teniendo en cuenta los departamentos y equipos de trabajo involucrados. Las personas definidas para llevar a cabo estas tareas cuentan con el **conocimiento técnico** (recursos internos o externos) y de **negocio suficiente,** así como de los recursos necesarios, para llevar a cabo su labor.	• **Verificar la existencia de personal cualificado** (interno o externo) dedicado a la organización y gestión de los sistemas de IA. • **Revisión del *job description*** y experiencia profesional *(curriculum vitae)* del personal técnico y responsables clave al cargo de los modelos de IA, incluyendo la **verificación** de sus cualificaciones. • Revisión de la **gestión de presupuestos** (incluyendo análisis recurrente de desviaciones) y **recursos** necesarios en la implementación de modelos de IA.
Definición de políticas y procedimientos internos suficientes y adecuados destinados al buen gobierno de los sistemas de IA. Dichas políticas y procedimientos son accesibles, conocidas y aplicadas por toda la organización, incluyendo campañas de concienciación y formación específica.	• **Revisión de las políticas internas y procedimientos.** Valorar si abordan los aspectos mínimos y riesgos intrínsecos de los modelos de IA, incluyendo (lista no exhaustiva); identificación roles y responsabilidades, arquitectura de TI y datos, estrategia y objetivos de los modelos de IA, medición de desempeño y métricas de los sistemas de IA.
Análisis del impacto de las **normativas y regulaciones externas** aplicables (p.ej. Normativas de regulación europea y/o local, incluyendo RGPD u otra regulación) e implementación de un adecuado sistema de cumplimiento regulatorio, incluyendo eventuales evoluciones futuras de las normativas y regulaciones externas.	• **Revisión de *checklist* de verificación** de la normativa de aplicación al Sistema de IA (AI Act, Reglamento de Protección de Datos, Normativa medioambiental, etc.). • **Evaluar si los sistemas de cumplimiento regulatorio** son suficientes y adecuados para atender los requerimientos de las normativas externas aplicables.
Definición de las características de los algoritmos sujetos a un análisis de riesgos y modelo de gobierno, garantizando un inventario centralizado y actualizado de los mismos. Es importante que existan actividades de control que permitan separar los casos de uso de la IA Generativa, del resto de tipologías de IA, para realizar una evaluación de riesgos enfocada a las características del caso de IA Generativa.	• **Revisión del inventario actualizado** de modelos y de su documentación establecida en los procedimientos definidos.
Implementación de un modelo de evaluación y estrategias de mitigación de riesgos, considerando un *risk-assessment* continuo y revisable en el tiempo, con la utilización de herramientas GRC *(Governance, Risk and Compliance),* que permita la monitorización continua de riesgos en un inventario centralizado de sistemas de IA. La evaluación de riesgos incluye, entre otros, aspectos éticos, sociales, tecnológicos y de ciberseguridad.	• Revisión del proceso para la identificación y construcción de un **inventario y/o mapa de riesgos,** incluyendo la idoneidad de las estrategias planteadas para su mitigación. • Valorar una adecuada **segregación de funciones** para asegurar las métricas de **impacto de los riesgos** (métricas y/o factores cuantitativos y cualitativos) identificados de forma continua en el tiempo.
Establecimiento de mecanismos de control para identificar aquellos modelos de IA que, de forma directa o indirecta, **pudieran tener impacto en la información financiera y no financiera** (así como cualquier otro tipo de información que pueda afectar a los procesos de decisión de la organización), implicando a los responsables de los procesos afectados (Dirección o Alta Dirección, cuando sea aplicable). Dichos mecanismos de control garantizan que los datos de salida de los modelos de IA utilizados para el registro contable de cualquier tipo de transacción, cuentan con los pertinentes controles de revisión, previos a su contabilización.	• **Revisión del inventario de sistemas de IA** de la compañía para garantizar la correcta identificación de aquellos modelos de IA, que pudieran servir como base de registros contables u otro tipo de información no financiera divulgada al mercado. • Análisis de los **datos utilizados por el control *owner* para el registro contable,** así como verificación de la documentación soporte y evidencias de revisión previa a la contabilización.

Requisitos específicos para la contratación de proveedores de servicios o aplicaciones de IA (incluyendo servicios de IA Generativa).	• **Evaluar las condiciones de contratación** para proveedores y servicios de IA • **Verificar la inclusión de proveedores de IA** en el inventario general de modelos de IA. • **Revisar los indicadores de servicio** y la documentación sobre datos de entrenamiento realizados por terceros.
Supervisar el impacto ambiental de los modelos de IA implantados, con el objetivo de monitorizar eventuales impactos en **indicadores no financieros clave de la compañía** (por ejemplo, Emisiones de Gases en Efecto Invernadero) y en los **compromisos de sostenibilidad comunicados** al mercado (por ejemplo, compromisos de reducción de Emisiones).	Evaluar que existe un **análisis del potencial impacto ambiental (indicadores no financieros y compromisos de materia de sostenibilidad)** del entrenamiento y uso de los modelos de Inteligencia Artificial y que se encuentra alineado con las políticas de la organización.

3. ARQUITECTURA DE DATOS Y SISTEMAS DE TI

La arquitectura de datos y TI utilizados son críticos para un adecuado gobierno de los sistemas de IA utilizados por las organizaciones, siendo relevantes los siguientes aspectos desde un punto de vista de control interno:

1. **Acceso a los datos generados** por la organización que son utilizados por los sistemas de IA (metadatos, taxonomía, entre otros).
2. **Aseguramiento de la privacidad, seguridad y tratamiento** adecuado a lo largo de todo el ciclo de vida de los datos, considerando las distintas fases de recolección, uso, almacenamiento y destrucción de los datos utilizados por los sistemas de Inteligencia Artificial.
3. Definición precisa del **rol y responsabilidades** en lo referido a la propiedad y responsabilidad de los usuarios sobre el ciclo de vida de los datos.

OBJETIVOS DE CONTROL O ACTIVIDADES DE CONTROL INTERNO GENERALES	PROCEDIMIENTOS DE AUDITORÍA INTERNA
Existencia de procesos y procedimientos formalizados sobre los **perfiles y roles de accesos** de los usuarios de los sistemas de IA. **Disponibilidad de controles de accesos** de usuarios a los sistemas de IA de acuerdo con los perfiles y roles predefinidos en los procedimientos de gestión de sistemas de IA, incluyendo: a) **Protocolo de autenticación de acceso** a los sistemas de IA (longitud mínima, caducidad predefinida contraseñas, entre otros aspectos de autenticación). b) **Gestión de cuentas de acceso y permisos de acceso** a los sistemas de IA.	• **Evidenciar la existencia y verificar la idoneidad de mecanismos de control** para la autenticación de los usuarios a los sistemas de IA. • **Revisión de los procedimientos de gestión altas/bajas** de usuarios y del listado de personas autorizadas a acceder a los sistemas de IA. • **Evidenciar una adecuada segregación de funciones** ente los diferentes roles (operadores, desarrolladores de algoritmos, propietarios de los datos, entre otros). • **Evidenciar la revisión periódica** de los permisos de acceso y gestión de accesos de usuarios con privilegios.

Existencia de procedimientos para la **gestión de cambios de arquitectura y datos** de los sistemas de TI. Este procedimiento incluye la operativa en los procesos de actualización de softwares, migraciones de entorno, entre otros. Existencia de un **proceso de gestión del cambio adecuado, que defina tres entornos separados** (de desarrollo, test y producción) y que se encuentren documentadas sus directrices de utilización, así como la gestión de los entornos y el acceso.	**Validar que los procedimientos de gestión del cambio** de arquitectura de TI y datos se realiza de forma adecuada, **evidenciando:** a) Las solicitudes para cambios en la arquitectura y desarrollos de los sistemas de IA son aprobados por los responsables autorizados. b) Evidenciar que la gestión de cambios incluye pruebas de aceptación de usuarios y puesta en producción de nuevos desarrollos de los sistemas de IA. c) Los cambios a los parámetros clave de configuración de sistemas de IA son monitorizados y revisados periódicamente. d) El equipo de gestión de los sistemas de IA se encarga de identificar, evaluar, priorizar e implantar los parches y nuevas versiones de software.
Definición de una política de respaldo para los sistemas implicados en el modelo, que alojen información de estos. Además, existencia de un **plan de continuidad ante incidencias y un plan de recuperación.**	• **Evidenciar** la existencia de copias de respaldo de los datos. • **Revisión de los planes de continuidad y recuperación,** así como evidenciar la implementación de estos mediante eventos ocurridos o probados.
Asegurar que los sistemas de IA implementados en la organización se encuentran **protegidos de ciber-incidentes** y se encuentran dentro de las políticas de ciberseguridad de la organización.	• **Evaluar** si los sistemas de IA se encuentran integrados dentro de la estrategia de ciberseguridad de la compañía, adecuadamente bastionados y sujetos a evaluaciones periódicas de seguridad.
Los datos que utilizan los sistemas de TI se encuentran protegidos con los estándares necesarios para atender los requerimientos de la normativa aplicable sobre protección de datos **(Reglamento General de Protección de Datos).**	• **Revisión** de la adecuada implementación de las políticas de protección sobre el universo de datos utilizados por los sistemas de IA, especialmente aquellos datos susceptibles y/o sensibles de acuerdo con las políticas internas y regulación (ej. RGPD).
Actividades de control para garantizar que los **datos confidenciales o sensibles** utilizados en el entrenamiento de modelos de IA Generativa dentro del contexto de la organización, y aquellos utilizados como parte del input en el comando de *prompt,* no son incorporados al entrenamiento del modelo de propósito general (externo a la compañía, y por lo tanto evitar riesgos de salida indebida de datos).	• **Evaluación** del sistema de entrenamiento del modelo de Inteligencia Artificial Generativa. • **Revisión de la licencia utilizada evidenciando de manera técnica** que el modelo sólo incorpora la información aportada a la rama o capa contratada por la organización, y no al modelo de propósito general.
Los **modelos entrenados por terceros** no vulneran derechos de privacidad o propiedad intelectual.	**Revisión** de los mecanismos de monitorización y seguimiento de los modelos entrenados por terceros (incluyendo modelos de propósito general).

4.CALIDAD DE LOS DATOS

Garantizar la **integridad, exactitud y confiabilidad de los datos** que alimentan los algoritmos de IA es crítico desde un punto de vista del buen gobierno de los modelos de IA. Las organizaciones deben de estar preparadas para la gestión de volúmenes ingentes de datos necesarios para un adecuado desempeño de los algoritmos construidos en las estructuras de IA.

En este sentido los procedimientos centrados en garantizar una adecuada calidad de los datos deben ser prioritarios para las organizaciones con la finalidad de garantizar el desempeño de los algoritmos de IA construidos sobre los datos.

OBJETIVOS DE CONTROL O ACTIVIDADES DE CONTROL INTERNO GENERALES	PROCEDIMIENTOS DE AUDITORÍA INTERNA
Definición y documentación de un **proceso de lectura de los datos de entrada,** asegurando su integridad y exactitud.	• **Revisión** de los procedimientos del proceso de lectura de datos utilizado. • **Obtener** una muestra de los datos de entrada y verificar que la organización ha incorporado protocolos de lectura adecuados, garantizando la integridad y exactitud de los datos añadidos al modelo. • **Revisión de *log* de errores de entrada de datos** y validar que son revisados y resueltos antes de la ejecución de los modelos de IA.
Existencia de un proceso de **testeo de la calidad de los *inputs*** empleados por el modelo de Inteligencia Artificial, así como de las transformaciones realizadas (por ejemplo, normalización).	• **Obtener las evidencias** de que se han definido valores máximos y mínimos sobre variables cuantitativas, y que existen controles que detectan la presencia de valores anómalos. • **Revisión de los controles** sobre variables con valores nulos, o sobre variables de control *checksum* o similar.
Las **fuentes y repositorios de datos** (p.ej. *data lake)* así como los cambios que les afectan, son supervisados y monitorizados de forma continua.	• **Evidenciar la supervisión y documentación,** por parte de los usuarios adecuados, de las fuentes y repositorios de datos (internas o externas) que alimentan los sistemas de IA. • **Obtener las evidencias** de aprobación de los cambios de las fuentes y repositorios de datos, incluyendo evaluación de riesgos y la calidad de los datos, derivada de dichos cambios.
Los modelos de Inteligencia Artificial cuentan con **actividades de control para la medición de la integridad, exactitud y confiabilidad de los datos,** los cuales son monitorizados con métricas o reportes de excepción para su análisis y resolución por los usuarios u *owners* de los sistemas de IA.	• **Revisión** de los reportes de excepciones y métricas sobre la calidad de los datos. • **Evidenciar las actividades** tomadas a cabo por los *onwers* de los sistemas de IA para la resolución de excepciones y análisis de las métricas de calidad de los datos.
En el caso de IA Generativa, establecimiento de **mecanismos de control y registro de los *prompt* de entrada, y su salida,** que restrinjan la utilización de datos personales o confidenciales o la generación de contenido ofensivo.	**Revisión** de los controles de entrada y salida sobre modelos de IA Generativa para garantizar la adecuada aplicación de las políticas de la compañía sobre protección de datos.

5. MEDICIÓN DEL DESEMPEÑO

Las organizaciones que abordan la implementación de sistemas de AI en sus procesos de negocio, deben incorporar en sus actividades la definición de métricas de desempeños de los algoritmos con el objetivo de asegurar que el comportamiento de los sistemas de AI atienden (y en qué medida) los objetivos de negocio para los cuales fueron creados, todo ello garantizando **medidas apropiadas de supervisión humana para minimizar riesgos.**

A modo de ejemplo ilustrativo, de métricas utilizadas para la medición del desempeño de sistemas de AI se encuentran el *error cuadrático medio,* el *coeficiente R2* y el *error absoluto medio*[26], para sistemas de regresión; o matrices de confusión para algoritmos de aprendizajes supervisados.

OBJETIVOS DE CONTROL O ACTIVIDADES DE CONTROL INTERNO GENERALES	PROCEDIMIENTOS DE AUDITORÍA INTERNA
Existencia de un **procedimiento implementado destinado a la medición continua del desempeño** de los modelos de IA considerando las actividades, parámetros, informes, métricas (entre otros), a tener en cuenta en la monitorización del desempeño de los sistemas de AI utilizados, de acuerdo con los objetivos de negocio perseguidos. Adicionalmente, se ha establecido la **frecuencia de medición y/o las situaciones en las que las desviaciones y/o incidencias identificadas** pueden requerir de calibración, desarrollo y/o mejoras de los algoritmos de IA.	• **Revisión** de los procedimientos y protocolo para la medición del desempeño de los sistemas de IA garantizando una **adecuada supervisión humana de los sistemas de IA.** • **Evaluar** la idoneidad y suficiencia de las métricas definidas e implementadas. • **Revisión** de las actividades de resolución llevadas a cabo antes desviaciones o excepciones identificadas, incluyendo los desarrollos o mejoras de los modelos, en su caso. • **Evaluar la frecuencia** establecida de reevaluación, reajuste o reinicio del componente para ajustarlo a desviaciones en los datos de entrada o cambios en los criterios de toma de decisiones.
Existencia de un **procedimiento documentado de interpretación de los resultados del algoritmo,** que recoja márgenes de tolerancia, umbrales o cualquier otro tipo de estadísticas de análisis requeridas para el entendimiento e interpretación de los datos de salida de los sistemas de IA. El procedimiento debe incluir la definición de las **acciones a tomar a cabo** (correctivas de los modelos o de negocio) en caso de que los resultados obtenidos sean no esperados o alejados de los márgenes o umbrales predefinidos. (Ej. casos de estudio por geografía, tipo de producto, periodo, monitorización transacciones atípicas).	• **Revisión** del procedimiento de interpretación de resultados del algoritmo. **Verificar** la idoneidad y requisitos mínimos para la interpretación de resultados. • **Asegurar que los *owners* de los sistemas de IA** cuentan con la experiencia y conocimiento suficiente (técnico y de negocio) para desarrollar adecuadamente las actividades de interpretación de los resultados del algoritmo. • **Evidenciar las acciones ejecutadas** en situaciones de variaciones significativas en los resultados obtenidos respecto de las salidas esperadas o superado los umbrales predefinidos. • Existen procedimientos para detectar si la respuesta del sistema de IA a los **datos de entrada es errónea o supera un umbral de error determinado.** • Se ha evaluado el comportamiento del sistema IA ante **casos de uso o entornos imprevistos.**

26. Ver definición en Glosario de Términos del Anexo II.

Existencia de un **proceso de *backtesting*** que permite medir la precisión del modelo y su rendimiento, de forma que puedan replicarse resultados del pasado con datos históricos de un periodo concreto.	• **Revisar** los procedimientos *backtesting* de la compañía y seleccionar aleatoriamente una ejecución anterior para el *reperformace* independiente por parte de Auditoría Interna, con el objetivo de comparar los datos de salida del modelo de IA.
Implementación de pruebas de *stress-testing* que permita medir los resultados esperados de los datos de salida del modelo de IA. (Ejemplo de un *stress-testing,* en un modelo de IA destinado a la identificación de transacciones anómalas o atípicas, el *stress-testing* supondría manipular los datos de entrada con datos atípicos para medir si el modelo de IA tendría la capacidad de capturarlo y arrojarlo como dato atípico en sus datos de salida).	• **Revisión de los procedimientos de *stress-testing*** realizados por los *owners* de los sistemas de IA durante la implementación y mantenimiento de los resultados de salida de los algoritmos. • ***Reperformance*** por parte del equipo de auditoría de una ejecución aplicando *stress-testing* en los datos de entrada al modelo de IA.
Implementación de un proceso de evaluación de la fiabilidad de las respuestas aportadas por el **modelo de IA Generativa** si éstas son utilizadas en algún proceso crítico (especialmente en procesos para la generación de Información Financiera y No Financiera clave para la compañía), a partir de la monitorización del **nivel de alucinación** de los modelos utilizados.	• **Revisar** que está documentada y monitoreándose la tasa de alucinación *(hallucination rate* en inglés), entre otros parámetros de desempeño de los sistemas de IA. • Revisión de la existencia de un **procedimiento de evaluación** de la fiabilidad de las respuestas de la IA Generativa, especialmente en aquellos casos en los que la IA Generativa aporte información (financiera *vs.* no financiera) dentro de un proceso crítico o relevante para la organización.
Para aquellos **modelos de IA Generativa** que obtengan como resultado un elemento audiovisual (imágenes, audios, vídeos) comprobar que se ha implementado un proceso que asegura que no se vulneran los **derechos de propiedad intelectual.** Adicionalmente, dicho proceso debe permitir identificar con marca de agua o similar aquellos elementos audiovisuales generados por IA Generativa con el objetivo de reducir el **riesgo de desinformación o uso fraudulento de contenido,** entre otros riesgos.	Revisar el proceso de generación de elementos audiovisuales y que incluye y se llevan a cabo los siguientes elementos: • Se regula su uso con el objetivo de no vulnerar los **derechos de propiedad intelectual.** • Se ha configurado un componente que distingue **unívocamente (por ejemplo, marcas de agua)** que el elemento audiovisual ha sido generado por IA.

6. EL FACTOR "CAJA NEGRA" (BLACK BOX) EN LOS SISTEMAS DE IA

En términos de la ciencia de los datos, el factor "caja negra" o *black box* en su terminología inglesa, se refiere a aquellos algoritmos de Inteligencia Artificial que, por su complejidad y/o sofisticación, los mecanismos internos de ejecución ente los datos de entrada y los de salida son difícilmente entendibles o explicables.

Aquellas organizaciones con iniciativas en complejos o sofisticados algoritmos de Inteligencia Artificial ven incrementarse los riesgos derivados del factor caja negra con mayor intensidad. En este sentido, las estructuras de control interno para reducir los riesgos intrínsecos para asumir los resultados de algoritmos con un alto factor caja negra se vuelven más críticos y relevantes. Como, por ejemplo, los algoritmos de recomendación o visualización de *Feed* de las principales redes sociales y plataformas de contenidos en *streaming.*

OBJETIVOS DE CONTROL O ACTIVIDADES DE CONTROL INTERNO GENERALES	PROCEDIMIENTOS DE AUDITORÍA INTERNA
Existencia un procedimiento documentado de análisis de sensibilidad del modelo ante fluctuaciones en los datos de entrada al modelo, y la interpretación de los resultados, con la finalidad de reducir el riesgo del factor *black box,* considerando la definición de: a) Métricas de precisión, exactitud y rendimiento de los sistemas de IA. b) Valores predefinidos de falsos positivos y falsos negativos. c) Mecanismos de supervisión para la adaptabilidad de sistemas de IA no supervisados (o de aprendizaje continuo) a nuevos datos y supervisión de la idoneidad de las conclusiones sostenibles en el tiempo con el aprendizaje continuo.	• **Evaluar el establecimiento de las métricas** o conjunto de métricas agregadas para determinar en los sistemas de IA su precisión, exactitud, sensibilidad u otro parámetro de rendimiento relativo a la aplicación del principio de exactitud de los datos. • **Evidenciar** los análisis realizados e interpretados sobre los valores de las tasas de falsos positivos y falsos negativos que arroja el componente IA de cara a determinar la precisión, la especificidad y la sensibilidad del comportamiento de los sistemas de IA. • **Evidenciar** la supervisión realizada del grado de adaptabilidad a nuevos datos o tipos de datos de entrada para el caso de modelos de IA no supervisados. • **Validar** los mecanismos de supervisión continua de modelos de aprendizaje continuo, con el objetivo de **verificar** que las conclusiones extraídas siguen siendo válidas, el componente es capaz de adquirir nuevo conocimiento y no se está produciendo una pérdida de las asociaciones previamente aprendidas durante el aprendizaje inicial.
Controles implementados para tener una **seguridad razonable de que se proporcionaron datos suficientes** para que el modelo genere resultados precisos.	**Revisión** de las actividades ejecutadas por el *owner* de los sistemas de IA para garantizar de que se proporcionaron datos suficientes (p. ej., que cubren un período de tiempo o variaciones suficientes en la población) para permitir que el modelo genere resultados precisos y reducir el factor *black box* de los algoritmos.
Controles implementados de monitorización continua destinados a la supervisión sobre los **datos de entrenamiento para evitar sesgos.** Adicionalmente, durante la operativa de los modelos de IA se **evalúa de manera continua la existencia de posibles sesgos** de la IA en lo relativo a componentes éticos / políticos / étnicos / raciales / de género / culturales, etc.	• **Evidenciar** la ejecución apropiada y oportuna de los procesos de supervisión de sesgos y evaluar la idoneidad de las actividades correctivas en los algoritmos en los casos identificados de sesgos ocurridos. • **Obtener** el histórico de revisiones y evaluar la existencia, frecuencia y numero evaluaciones llevadas a cabo, con especial atención a aquellos sesgos, que, de forma reiterada, se han puesto de manifiesto.
Existe un marco/mecanismo general de monitoreo/alerta **en tiempo real para detectar cualquier anomalía** en la operación de extremo a extremo de los procesos, controles, sistemas y/o datos de IA.	• **Revisión** del registro de indicadores clave de supervisión e histórico de alerta y **evaluación** de las medidas tomadas a cabo para la corrección de anomalías.
Implementación de mecanismos de monitorización continua para la identificación de **procesos ineficaces de los sistemas de IA** (es decir, ocurre un incidente importante o la solución ha evolucionado/aprendido de manera inapropiada). Ante ineficiencias de los sistemas de IA, existen **mecanismos de reversión para la corrección de algoritmos y acceso disponible a datos "limpios",** con la finalidad de alcanzar la eficacia de los modelos de IA de forma oportuna en el tiempo.	• **Revisión y valoración** de la idoneidad de los procesos para la identificación de sistemas de IA ineficaces, • **Evidenciar** cómo las actividades de reversión implementadas históricas, consiguieron abordar ejecuciones de los sistemas de IA ineficaces.

En el caso, de modelos de IA Generativa, **existe documentación funcional / técnica** sobre la contextualización, e información asociada a la versión del modelo utilizado.

- **Revisión de la documentación funcional / técnica** asociada a la aplicación de IA Generativa indicando las funcionalidades técnicas de la versión o modelo implantada.
- **Verificar que el modelo de IA Generativa** no vulnera derechos de propiedad intelectual en su entrenamiento.

7. EL FACTOR HUMANO Y EL SESGO ALGORÍTMICO

El **factor humano en el diseño, implementación y mantenimiento** de sistemas de IA es uno de los aspectos a considerar más relevantes, especialmente ante modelos de IA de autoaprendizaje no supervisado, con potencial impacto adverso o no deseado en la sociedad y en los procesos de negocio de las organizaciones. En este sentido, el factor humano incluye aspectos a considerar como; valores éticos y morales y sesgos de los algoritmos *(algorithm bias)* los cuales se desarrollan a continuación:

1. **Valores éticos y morales.** Los algoritmos son desarrollados por humanos, por lo tanto, cualquier error (intencionado o no), tendrá un impacto directo en el desempeño y resultado de los sistemas de IA. En este sentido, debemos plantearnos las implicaciones éticas y morales de los resultados obtenidos de los sistemas de IA; tal y como se describe a continuación:

   a) Los resultados obtenidos de los modelos de IA son utilizados de forma legal, ética y responsable,

   b) Los sistemas de IA son testeados durante las fases de despliegue, estabilización y madurez de forma que se asegure que siguen atendiendo los objetivos para los cuales fueron diseñados, y no existen desviaciones que comprometan, por ejemplo, los principios de negocio responsable o políticas internas de una organización.

   c) Existencia de controles que aborden los riesgos de errores, intencionados o no, en la construcción de los modelos de IA por los humanos.

2. **El sesgo algorítmico *(algorithm bias)*.** En los sistemas de IA, el sesgo algorítmico ocurre cuando los valores de los humanos que lo diseñan y desarrollan terminan, de alguna forma intencionada o no, en los algoritmos de IA que desarrollan. Estos modelos de IA creados por humanos, pueden terminar adquiriendo comportamientos sexistas, racistas, homófobos o de otra índole, a semejanza de los humanos que los crearon. De la misma forma, la utilización de datos históricos podría inferir sesgo también en los modelos. Si un modelo de riesgo de crédito se alimenta de datos históricos y un determinado perfil de personas (por ejemplo, mujeres) tienen un histórico mayor de impagados, el modelo estará sesgado y podría sugerir otorgar menos créditos a mujeres. En ese sentido, la ingesta de datos a los modelos es considerada también crítica desde un punto de vista de sesgo algorítmico.

   En otras palabras, derivado de los valores humanos de aquellos que diseñan los modelos o los datos históricos utilizados, los algoritmos de IA pueden adquirir un sesgo.

OBJETIVOS DE CONTROL O ACTIVIDADES DE CONTROL INTERNO GENERALES	PROCEDIMIENTOS DE AUDITORÍA INTERNA
Actividades de control diseñadas con el objetivo de impedir que los **resultados de los sistemas de IA sean utilizados de forma ilegal o delictiva, o incumpliendo cualquier regulación externa o política empresarial interna.**	• **Revisión** de los objetivos o estrategia de implantación de los sistemas de IA, **e identificar** cualquier brecha legal, o en la regulación externa o políticas internas. • **Revisión** de los resultados de los sistemas de IA, **y asegurar** que los mismos son utilizados sin intenciones ilícitas o legales, o en contra de la regulación externa o políticas internas de la compañía.
Asegurar que los resultados de los modelos de IA se encuentran **libres de sesgos algorítmicos (modelos de IA generativos o no generativos),** intencionados o no.	• **Revisión** de los objetivos de los sistemas de IA para descartar cualquier tipo de sesgo (intencionado o no) en la fase de diseño de los sistemas de IA. • **Revisión** de los resultados perseguidos por los sistemas de IA, y **compararlos** con los objetivos para identificar cualquier desviación y **determinar** si la causa fue un sesgo algorítmico. • **Revisión de la existencia de procedimientos y/o protocolos** para identificar sesgos algorítmicos motivados por datos históricos con sesgo.

Anexo I: Bibliografía

- **PwC,** *22nd Annual Global Survey.* 2019
- **Stanford University,** *Artificial Intelligence Index Report.* 2022
- **Committee of Sponsoring Organizations of the Treadway Commission (COSO),** *Realize the full potential of artificial intelligence.* 2021.
- **Comisión Europea.** *Reglamento del Parlamento y del Consejo Europeo sobre normas armonizadas en materia de Inteligencia Artificial.* [https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=OJ:L_202401689]. 2024.
- **Alan Mathison Turing.** *Computing machinery and intelligence.* 1950.
- **McKinsey Digital, Global Survey:** *The state of AI in 2020.*
- **Google,** *AI Principles 2020 update.* 2020.
- **Lapam, Maxim.** *Deep Reinforcement Learning Hands-On.* Birmingham (UK): Packt Publishing Ltd. 2018. ISBN 978-1-78883-424-7.
- **Recuero, P.** *Los 2 tipos de aprendizaje en Machine Learning: supervisado y no supervisado.* 2017 [Blog] http://data-speaks.luca-d3.com/2017/11/que-algoritmo-elegir-en-ml-aprendizaje.html
- **EY,** *Building the right governance model for AI/ML.* 2019
- **Comisión Nacional del Mercado de Valores (CNMV),** *Código del buen gobierno de las sociedades cotizadas.* 2020.
- **Committee of Sponsoring Organizations of the Treadway Commission (COSO),** *Internal Control Integrated Framework,* 2013
- **Committee of Sponsoring Organizations of the Treadway Commission (COSO),** *Enterprise Risk Management.* 2017.
- **Agencia Española de Protección de Datos.** *Requisitos para auditorías de tratamientos de datos personales que incluyan Inteligencia Artificial.* 2021.
- **The Institute of Internal Auditors.** *The IIA´s Artificial Intelligence Auditing Framework (Part I, II and III).* 2017.

Anexo II
Glosario de términos

Regresión lineal

La regresión lineal es una técnica de modelado estadístico que se emplea para describir una variable de respuesta continua como una función de una o varias variables predictoras. Puede ayudar a comprender y predecir el comportamiento de sistemas complejos o a analizar datos experimentales, financieros y biológicos, entre otros.

Regresión logística

La regresión logística es un tipo de análisis de regresión utilizado para predecir el resultado de una variable categórica (una variable que puede adoptar un número limitado de categorías) en función de las variables independientes o predictoras. Es útil para modelar la probabilidad de un evento ocurriendo en función de otros factores.

Clustering* (o Análisis *Cluster)

Es una técnica estadística multivariante que busca agrupar elementos (o variables) tratando de lograr la máxima homogeneidad en cada grupo y la mayor diferencia entre los grupos.

Análisis factorial

Análisis factorial es una técnica estadística de reducción de datos usada para explicar las correlaciones entre las variables observadas en términos de un número menor de variables no observadas llamadas factores.

Series temporales

Una serie temporal es una sucesión de observaciones de una variable tomadas en el transcurso del tiempo, de manera que los valores que toma la variable aparecen ordenados cronológicamente.

Conexión API *(Application Programming Interface)*

Las API son mecanismos que permiten a dos componentes de software comunicarse entre sí mediante un conjunto de definiciones y protocolos.

Datos estructurados

Información que se encuentra almacenada habitualmente en bases de datos relacionales, cuyos datos se encuentran organizados en registros (filas) y columnas (atributos), de manera que se estructuran en formato tabla. Los datos estructurados se usan de manera habitual en la mayor parte de las bases de datos relacionales. El lenguaje de programación mediante el cual se gestionan comúnmente bases de datos relacionales es el *Structured Query Language* o SQL, desarrollado por IBM al comienzo de la década de 1970.

Datos no estructurados

Son generalmente datos binarios que no tienen estructura interna identificable. Es un conglomerado masivo y desorganizado de varios objetos que no tienen valor hasta que se identifican y almacenan de manera organizada. Una vez que se organizan, los elementos que conforman su contenido pueden ser buscados y categorizados (al menos hasta cierto punto) para obtener información.

Árboles de Decisión

Un árbol de decisión en *Machine Learning* es una estructura de árbol similar a un diagrama de flujo donde un nodo interno representa una característica (o atributo), la rama representa una regla de decisión y cada nodo hoja representa el resultado. Dado un conjunto de datos se fabrican diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que ocurren de forma sucesiva, para la resolución de un problema.

Gradient Boosting

Gradient boosting o Potenciación del gradiente, es una técnica de aprendizaje automático utilizado para el análisis de la regresión y para problemas de clasificación estadística, el cual produce un mo-

delo predictivo en forma de un conjunto de modelos de predicción débiles, típicamente árboles de decisión.

Random Forest

Es una combinación de árboles predictores tal que cada árbol depende de los valores de un vector aleatorio probado independientemente y con la misma distribución para cada uno de estos.

Support Vector Machines

Es un algoritmo de aprendizaje supervisado que se utiliza en muchos problemas de clasificación y regresión, incluidas aplicaciones médicas de procesamiento de señales, procesamiento del lenguaje natural y reconocimiento de imágenes y voz, entre otros.

Naive Bayes

Son algoritmos de aprendizaje automático que se basan en una técnica de clasificación estadística llamada "teorema de Bayes". En ellos se asume que las variables predictoras son independientes entre sí, y, por lo tanto, la presencia de una cierta característica en un conjunto de datos no está en absoluto relacionada con la presencia de cualquier otra característica.

Error cuadrático medio, Coeficiente R2, Error absoluto medio en sistemas de regresión

En modelos de regresión, predecimos o estimamos el valor numérico de una cantidad desconocida, de acuerdo con unas características dadas. La diferencia entre la predicción y el valor real es el error, que es una variable aleatoria utilizada para medir el desempeño de los modelos de Inteligencia Artificial. Algunos ejemplos para medir el rendimiento de los sistemas de regresión son:

a. **El error cuadrático medio** representa la raíz cuadrada de la distancia cuadrada promedio entre el valor real y el valor pronosticado.

b. **El error absoluto medio** es el promedio de la diferencia absoluta entre el valor observado y los valores predichos.

c. **El coeficiente R2** indica la bondad o la aptitud del modelo. A menudo se utiliza con fines descriptivos y muestra que también las variables independientes seleccionadas explican la variabilidad en sus variables dependientes.

Matrices de confusión

Una matriz de confusión es una herramienta que permite la visualización del desempeño de un algoritmo que emplea en aprendizaje supervisado. En una matriz de confusión, las columnas representan el número de predicciones de cada clase, mientras que las filas representan las instancias reales. Uno de los beneficios de las matrices de confusión es que facilitan la visualización del desempeño de los algoritmos de aprendizaje supervisados; qué tipo de aciertos y errores está teniendo el modelo con el procesamiento de los datos.

Instituto de Auditores Internos de España

Santa Cruz de Marcenado, 33 · 28015 Madrid · Tel.: 91 593 23 45 · Fax: 91 593 29 32 · www.auditoresinternos.es

ISBN: 978-84-126682-9-2

Maquetación: desdecero, estudio gráfico

OTRAS PRODUCCIONES DE LA FÁBRICA DE PENSAMIENTO

NUEVAS FORMAS DE TRABAJO EN REMOTO DE AUDITORÍA INTERNA

Esta guía de buenas prácticas analiza todos los aspectos necesarios para desarrollar el trabajo en remoto en la Dirección de Auditoría Interna, incluyendo sus implicaciones en las relaciones con los stakeholders y los retos y limitaciones –y cómo hacerles frente– de esta forma de trabajar.

AUDITORÍA INTERNA DE LA GESTIÓN DE CRISIS Y RESILIENCIA DEL NEGOCIO

Abarca el rol de Auditoría Interna en la supervisión de los mecanismos de gestión de crisis y la resiliencia del negocio, así como el papel que asume en la fase previa, durante y después de que se produzca una crisis, e identifica las mejores prácticas relacionadas con la actuación de Auditoría Interna en este tipo de trabajos.

GESTIÓN ESTRATÉGICA DEL TALENTO EN AUDITORÍA INTERNA

La gestión del talento es fundamental para la consecución de los objetivos de la compañía y de cada uno de los departamentos que la integran. Este documento abarca distintas dimensiones de la gestión del tallento desde la óptica de la consecución de los objetivos de la Dirección de Auditoría Interna y en el ámbito del *Marco Internacional para la Práctica Profesional de la Auditoría Interna*.

PUESTA EN MARCHA DE UN DEPARTAMENTO DE AUDITORÍA INTERNA

Este documento ofrece las pautas y estrategias básicas para que el Director de Auditoría Interna pueda poner en marcha el departamento de manera eficiente, teniendo en cuenta las expectativas de los principales *stakeholders*. Se abordan las diferentes etapas a seguir y los principales hitos que deben ser considerados desde las perspectivas estratégicas, funcionales y metodológicas..

LA FÁBRICA DE PENSAMIENTO
INSTITUTO DE AUDITORES INTERNOS DE ESPAÑA

Este documento actualizado aborda los casos de uso más comunes de la Inteligencia Artificial (IA) en procesos empresariales y la regulación aplicable promovida por los legisladores. Además, describe los principales modelos y tipologías de IA, incluyendo la nueva perspectiva de la Inteligencia Artificial generativa, que la industria está desarrollando.

También se describe el marco de control interno general esperado y los riesgos relacionados para las organizaciones que utilizan tecnología basada en IA. Finalmente, se propone un programa de trabajo para llevar a cabo la auditoría de las estructuras de control interno diseñadas e implementadas en procesos de negocio con presencia de IA, incluyendo los principales procedimientos de auditoría sugeridos para su revisión.