5

Empresas desarrollando cumplimiento de requisitos

Guía 5. Sistema de gestión de riesgos

Reglamento Europeo de Inteligencia Artificial

Financiado por la Unión Europea
NextGenerationEU

GOBIERNO DE ESPAÑA
MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA
SECRETARÍA DE ESTADO DE DIGITALIZACIÓN E INTELIGENCIA ARTIFICIAL

Plan de Recuperación, Transformación y Resiliencia

España | digital 2026

Esta guía ha sido desarrollada en el marco del desarrollo del piloto español de sandbox regulatorio de IA, en colaboración entre los participantes, asistencias técnicas, potenciales autoridades nacionales competentes y el grupo asesor de expertos del sandbox.

La guía tiene como objetivo servir de apoyo introductorio a la normativa europea de Inteligencia Artificial y sus obligaciones aplicables. Si bien **no tiene carácter vinculante ni sustituye ni desarrolla la normativa aplicable, proporciona recomendaciones prácticas** alineadas con los requisitos regulatorios a la espera de que se aprueben las normas armonizadas de aplicación para todos los estados miembros.

El presente documento está sujeto a un **proceso permanente de evaluación y revisión**, con actualizaciones periódicas conforme al desarrollo de los estándares y las distintas directrices publicadas desde la Comisión Europea, y será actualizada una vez se apruebe el Ómnibus digital que modifica el Reglamento de Inteligencia Artificial.

Entre las referencias técnicas relevantes actualmente en desarrollo y aplicables, destacan el **prEN 18228 "Artificial Intelligence – Risk Management"** y la **ISO/IEC 23894 "Information technology – Artificial intelligence – Guidance on risk management"**, que conjuntamente servirán de base para la gestión y evaluación de riesgos a lo largo del ciclo de vida de los sistemas de inteligencia artificial, alineadas con el cumplimiento del Reglamento Europeo de Inteligencia Artificial.

Fecha de versión: 10 de diciembre de 2025

Contenido general

1. Preámbulo 5
2. Introducción 8
3. Reglamento de Inteligencia Artificial 11
4. ¿Qué elementos debo implantar y cómo debo hacerlo para desarrollar un adecuado sistema de gestión de riesgos? 15
5. Otros elementos a considerar 27
6. Documentación técnica 31
7. Cuestionario de autoevaluación 32
8. Anexos 33
9. Referencias, estándares y normas 61

Índice detallado

1. Preámbulo 5
1.1 Objetivo del documento 5
1.2 ¿Cómo leer esta guía? 5
1.3 ¿A quién está dirigido? 6
1.4 Casos de uso y ejemplos dispuestos a lo largo de la guía 6
2. Introducción 8
2.1 ¿Qué es un sistema de gestión de riesgos y cuáles son los elementos principales? 8
2.2 Proporcionalidad en la gestión del riesgo 10
3. Reglamento de Inteligencia Artificial 11
3.1 Análisis previo y relación de los artículos 11
3.2 Contenido del artículo 11
3.3 Correspondencia del articulado con los apartados de la guía 14
4. ¿Qué elementos debo implantar y cómo debo hacerlo para desarrollar un adecuado sistema de gestión de riesgos? 15
4.1 Determinación del apetito al riesgo 15
4.2 Contexto del sistema de IA 17
4.3 Identificación de riesgos 18
4.4 Análisis y evaluación de riesgos 20
4.5 Respuesta al riesgo 22
4.6 Documentación técnica del sistema de gestión de riesgos 24
4.7 Comunicación y consulta 25
4.8 Seguimiento y mejora continua 25
4.9 Liderazgo y compromiso 25
5. Otros elementos a considerar 27
5.1 Procedimientos de prueba 27
5.1.1 Pruebas destinadas a comprobar que los HRAIS funcionan conforme a su finalidad prevista 27
5.1.2 Pruebas destinadas a comprobar que los HRAIS cumplen con los requisitos establecidos 27
5.1.3 Pruebas en condiciones reales 28
5.1.4 Momento adecuado para la realización de las pruebas 28
5.2 Evaluación de riesgos relacionados con el sistema de vigilancia poscomercialización 28
5.3 Acceso e impacto del sistema por menores de 18 años 29
5.4 Entidades sujetas a legislación sectorial 30
6. Documentación técnica 31
7. Cuestionario de autoevaluación 32

8. Anexos 33
8.1 Anexo A - Elementos más relevantes del contexto interno y externo entorno a la IA 33
8.1.1 Anexo A.I - Elementos generales 33
8.1.2 Anexo A.II - Elementos relacionados con la Carta de los Derechos Fundamentales de la UE 34
8.2 ANEXO B - Componentes más comunes de los sistemas de IA 44
8.3 ANEXO C - Tipos de riesgos comunes en el ámbito de la IA 48
8.4 ANEXO D - Ejemplos de controles en el ámbito de la IA 51
8.5 ANEXO E – Ejemplos de indicadores de efectividad 53
8.5.1 ANEXO E.I – En relación con las medidas de gestión de riesgos 53
8.5.2 ANEXO E.II – Con relación a los controles del Anexo D 54
8.6 ANEXO F – Glosario de términos 59
8.7 ANEXO G – Política de Gestión de riesgos de IA 59
9. Referencias, estándares y normas 61

1. Preámbulo

1.1 Objetivo del documento

En esta guía se presentan las medidas organizativas y técnicas que servirán a proveedores y responsables del despliegue para dar cumplimiento con el artículo "Sistema de gestión de riesgos" del Reglamento Europeo de la IA.

Este artículo regula el sistema de gestión de riesgos que deberá incorporar todo sistema de IA de alto riesgo (HRAIS) y ciertos sistemas de IA de propósito general (artículo "Requisitos para sistemas IA de propósito general y obligaciones para proveedores de estos sistemas").

En este sentido, a lo largo de la guía nos referiremos, generalmente, a estos sistemas como "sistema de IA" con el objetivo de simplificar el discurso.

1.2 ¿Cómo leer esta guía?

Si el lector no conoce cómo desarrollar un sistema de gestión de riesgos:

Se recomienda al lector una primera lectura de la guía de la mano de los ejemplos incorporados, que le ayudará a adquirir una visión de los aspectos que han de cubrirse para la implantación de un sistema de gestión de riesgos en el contexto de la IA.

Seguidamente, se recomienda al lector una segunda lectura, siguiendo el detalle incorporado en el Excel mencionado en la sección 1.4, que recoge el proceso de desarrollo de un sistema de gestión de riesgos para 2 casos de uso.

Si el lector conoce cómo desarrollar un sistema de gestión de riesgos:

Se recomienda al lector como mínimo una lectura completa de la guía. Aunque los conceptos de la sección 2 y sección 4, que describen los elementos principales para el desarrollo de un sistema de gestión de riesgos le puedan resultar familiares, no obstante, se recomienda prestar especial atención a:

- Los ejemplos expuestos a lo largo de dichas secciones que tratan de aterrizar estas prácticas al entorno de la IA.
- Los catálogos referenciados en estas secciones (presentes en los anexos):
  - Anexo A: Elementos más relevantes del contexto interno y externo entorno a la IA.
  - Anexo B: Componentes más comunes de los sistemas de IA (todo sistema de gestión de riesgos gira en torno a la identificación y análisis de riesgos y estos riesgos son inherentes a los componentes del sistema en análisis).
  - Anexo C: Fuentes de riesgos comunes en el ámbito de la IA (permitirá al lector identificar los nuevos riesgos connaturales de la IA).
  - Anexo D: Ejemplos de controles en el ámbito de la IA.
  - Anexo E: Indicadores de efectividad.

- Sección 5 desarrollada para cubrir otros aspectos propios de los requisitos del artículo.

1.3 ¿A quién está dirigido?

Los requisitos descritos en el artículo "Sistema de gestión de riesgos" deben cumplimentarse por aquel encargado del desarrollo del sistema, es decir, el proveedor.

En dicho artículo, no se especifican requisitos para el responsable del despliegue del sistema. Si el responsable del despliegue participa en el desarrollo del sistema, deberá aplicar las medidas desarrolladas para el proveedor. A pesar de todo, se interpela al responsable del despliegue a realizar un uso responsable y ético del sistema en todo momento.

Además, implementar los requisitos del artículo "Sistema de gestión de riesgos" no está en las obligaciones que define el artículo "Obligaciones de los responsables del despliegue de los sistemas IA de alto riesgo".

Las medidas detalladas a lo largo de esta guía son orientativas para el proveedor. Son tanto de carácter organizativo como técnico. Si bien atendiendo a la naturaleza del artículo "Sistema de gestión de riesgos", la mayor parte de las medidas son de carácter organizativo.

1.4 Casos de uso y ejemplos dispuestos a lo largo de la guía

Para **facilitar la comprensión de la guía**, se incorpora junto un documento Excel que recoge el proceso de desarrollo de un **sistema de gestión de riesgos** para **diferentes casos de uso**.

Para ello se han seguido los pasos descritos en la sección 4 de la presente guía, donde se describen los elementos a implementar para desarrollar un adecuado sistema de gestión de riesgos.

Estos ejemplos se desarrollan en base a los **casos de uso descritos** en la **Guía práctica y ejemplos para entender el Reglamento IA.**

El documento Excel se complementa con diversas explicaciones, para vincular cada fase del desarrollo del sistema de gestión de riesgos abordadas con los elementos descritos en la guía.

Adicionalmente, para facilitar la lectura de la guía, se han incorporado pequeños ejemplos entre los diferentes apartados de esta.

Finalmente, hay que aclarar que estos ejemplos son meramente ilustrativos. Proveedor y responsable del despliegue han de considerar la aplicación de todas las medidas indicadas en esta guía, según corresponda.

Además, los ejemplos expuestos son específicos de los casos de uso. Esto implica que las propuestas son específicas para los modelos considerados como ejemplo, y no una solución general para otros tipos de modelos, o incluso modelos de la misma tipología.

Cada organización deberá, acorde a esta guía, establecer las medidas oportunas para su tipo de sistema de IA y su finalidad prevista.

2. Introducción

2.1 ¿Qué es un sistema de gestión de riesgos y cuáles son los elementos principales?

Un sistema de gestión de riesgos es un sistema de gestión cuyo objetivo es la identificación y análisis de riesgos y la implementación de medidas mitigadoras.

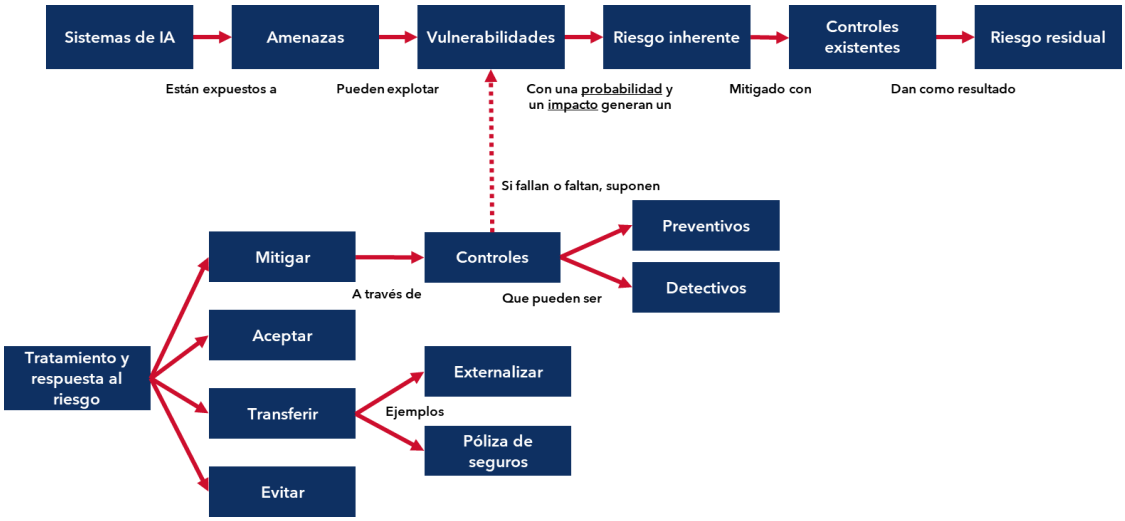
Entendemos por riesgo, cualquier evento con una probabilidad de suceder y un impacto en caso de que suceda. El riesgo, como veremos más adelante en detalle, se calcula como el producto de los factores de riesgo, su probabilidad e impacto.

En el contexto del Reglamento Europeo de la IA, el sistema de gestión de riesgos que se desarrolle deberá prestar especial atención a los **riesgos que puedan afectar a la salud, la seguridad y los derechos fundamentales de las personas**.

Se deberá poner el foco en la identificación y análisis de cualquier riesgo que pueda tener especial afectación en los elementos mencionados. Además, se deberán implementar las medidas adecuadas que permitan su mitigación.

El proceso de gestión de riesgos debe abordarse en todas las etapas del ciclo de vida del sistema de IA, desde el diseño y desarrollo hasta su comercialización y poscomercialización.

A continuación, se detallan los conceptos fundamentales que giran en torno a ecosistema de la gestión de riesgos y se disponen en un esquema que representa las relaciones entre ellos:



- Nuestro **sistema de IA** puede estar expuesto a diferentes **amenazas** que podrían terminar suponiendo un **riesgo** para nuestro sistema y, por consiguiente, para **la salud, la seguridad y los derechos fundamentales de las personas.**

Ejemplo

En el caso de uso de la **promoción de empleados**, supongamos que somos una organización que decide desarrollar y comercializar un **sistema de IA** que analiza los perfiles y el rendimiento de los trabajadores y ayuda a determinar quién merece en mayor medida un ascenso.

Una posible **amenaza** asociada a este sistema sería, por ejemplo, un **agente malicioso externo que trate de contaminar los datos de entrenamiento** que el sistema analiza para perturbar las recomendaciones que éste facilita. Pudiendo dar lugar a una decisión de promoción discriminatoria para algunos empleados.

- Estas **amenazas** pueden materializarse en **riesgos** a través de la explotación de **vulnerabilidades** de componentes de nuestro sistema. Si no tenemos implementadas las **medidas de control oportunas** podemos ser **vulnerables** a estas **amenazas**.

Ejemplo

Una posible **medida de control** sería, por ejemplo, la implementación de una herramienta de identificación de datos adversos en el sistema de IA. Esta herramienta analiza nuestros datos de entrenamiento y trata de determinar si hay datos que han sido modificados o introducidos por un agente externo de forma indeseada.

- Estas **amenazas** que pueden explotar una **vulnerabilidad** de componentes del sistema pueden materializarse con una cierta **probabilidad** y producir un cierto **impacto**.
  El riesgo lo determina el valor del impacto multiplicado por la probabilidad de que se produzca (riesgo = impacto*probabilidad). En la sección 4.4 lo veremos con más detalle.

Ejemplo

Siguiendo nuestro ejemplo, deberemos determinar el **impacto** que supone para **la salud, la seguridad y los derechos fundamentales** de los **empleados**, el hecho de que un **agente externo perturbé los datos (amenaza)** con los que el sistema toma las decisiones sobré quien merece más un ascenso.

También deberemos analizar la **probabilidad** de que esto suceda. Este tipo de **amenaza** podría suponer una promoción discriminatoria de un empleado frente a otro.

Además de mitigarse los riesgos, existen **otras formas de tratar los riesgos**.

Éstos se pueden **asumir** (*aceptando sus consecuencias*), **evitar** (*decidiendo no iniciar o descontinuar la actividad que genera el riesgo*) o **transferir** *(por ejemplo, contratando una póliza de seguros*). En la sección 4.5 se verá con más detalle.

Estos **elementos más en detalle y otros adicionales** que pertenecen al ecosistema de la gestión de riesgos desde una perspectiva más completa, son los que analizaremos en la sección 4.

2.2 Proporcionalidad en la gestión del riesgo

Las medidas descritas a lo largo de este documento pretenden servir como guía para que los proveedores de los sistemas de IA puedan dar cumplimiento a los requisitos exigidos en el artículo "Sistema de gestión de riesgos".

En este sentido, para la correcta implementación de un sistema de gestión de riesgos se recomienda a toda organización abordar las fases dispuestas en esta guía.

Aunque la exhaustividad y profundidad con las que se aborde cada fase podrán depender de las necesidades y recursos de cada organización, los riesgos para **la salud, la seguridad y los derechos fundamentales de las personas** deben ser especialmente considerados.

También las medidas de control que se implementen: cada organización seleccionará las que considere oportunas y necesarias para cumplir con las exigencias del Reglamento para garantizar **la salud, la seguridad y los derechos fundamentales de las personas**.

Es posible que las medidas descritas en esta guía sirvan de inspiración a las organizaciones para la definición e implementación de nuevas medidas o que la organización determine la implementación de otras medidas diferentes.

Las medidas descritas en esta guía podrán ser implementadas por cada organización de forma proporcional a los riesgos a analizar y mitigar, su impacto sobre la organización y el coste de su implementación.

3. Reglamento de Inteligencia Artificial

La puesta en servicio o la utilización de sistemas de IA de alto riesgo debe supeditarse al cumplimiento de determinados requisitos obligatorios, entre los cuales está el de la gestión de riesgos. Estos requisitos tienen como objetivo garantizar que los sistemas de IA de alto riesgo disponibles en la Unión o cuyos resultados de salida se utilicen en la Unión no representen riesgos inaceptables para intereses públicos importantes reconocidos y protegidos por el Derecho de la Unión.

En este apartado se incluye los artículos referentes a la generación de la gestión de riesgos del Reglamento 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024 (Reglamento Europeo de Inteligencia Artificial) y se detalla en que secciones de esta guía se abordan los diferentes elementos de dichos artículos.

3.1 Análisis previo y relación de los artículos

El reglamento aborda los sistemas de gestión de riesgos principalmente en el **artículo 9**, donde se regula la **implementación y mantenimiento** de estos para los sistemas de inteligencia artificial (IA) de alto riesgo, estableciendo un enfoque **iterativo** y continuo que abarca todo el ciclo de vida del sistema.

Su objetivo principal es **identificar, analizar, evaluar y mitigar riesgos** potenciales relacionados con la salud, seguridad y derechos fundamentales, tanto en el uso previsto como en usos razonablemente previsibles. El artículo también incluye mecanismos de vigilancia, pruebas y actualizaciones sistemáticas, con énfasis en la mitigación técnica, el diseño adecuado, y la capacitación de los responsables del despliegue.

Se establece la obligatoriedad de **realizar pruebas y revisiones** regulares, priorizando la minimización de riesgos residuales y adaptando las medidas según la finalidad y el contexto de uso, con atención especial a menores y colectivos vulnerables.

3.2 Contenido del artículo

AI Act

Art.9 – Sistema de gestión de riesgos

1. Se establecerá, implantará, documentará y mantendrá un sistema de gestión de riesgos en relación con los sistemas de IA de alto riesgo.

2. El sistema de gestión de riesgos se entenderá como un proceso iterativo continuo planificado y ejecutado durante todo el ciclo de vida de un sistema

de IA de alto riesgo, que requerirá revisiones y actualizaciones sistemáticas periódicas. Constará de las siguientes etapas:

a) la determinación y el análisis de los riesgos conocidos y previsibles que el sistema de IA de alto riesgo pueda plantear para la salud, la seguridad o los derechos fundamentales cuando el sistema de IA de alto riesgo se utilice de conformidad con su finalidad prevista;

b) la estimación y la evaluación de los riesgos que podrían surgir cuando el sistema de IA de alto riesgo se utilice de conformidad con su finalidad prevista y cuando se le dé un uso indebido razonablemente previsible;

c) la evaluación de otros riesgos que podrían surgir, a partir del análisis de los datos recogidos con el sistema de vigilancia poscomercialización a que se refiere el artículo 72;

d) la adopción de medidas adecuadas y específicas de gestión de riesgos diseñadas para hacer frente a los riesgos detectados con arreglo a la letra a).

3. Los riesgos a que se refiere el presente artículo son únicamente aquellos que pueden mitigarse o eliminarse razonablemente mediante el desarrollo o el diseño del sistema de IA de alto riesgo o el suministro de información técnica adecuada.

4. Las medidas de gestión de riesgos mencionadas en el apartado 2, letra d), tendrán debidamente en cuenta los efectos y la posible interacción derivados de la aplicación combinada de los requisitos establecidos en la presente sección, con vistas a reducir al mínimo los riesgos de manera más eficaz al tiempo que se logra un equilibrio adecuado en la aplicación de las medidas para cumplir dichos requisitos.

5. Las medidas de gestión de riesgos mencionadas en el apartado 2, letra d), considerarán aceptables los riesgos residuales pertinentes asociados a cada peligro, así como el riesgo residual general de los sistemas de IA de alto riesgo.

A la hora de determinar las medidas de gestión de riesgos más adecuadas, se procurará:

a) eliminar o reducir los riesgos detectados y evaluados de conformidad con el apartado 2 en la medida en que sea técnicamente viable mediante un diseño y un desarrollo adecuados del sistema de IA de alto riesgo;

b) implantar, cuando proceda, unas medidas de mitigación y control apropiadas que hagan frente a los riesgos que no puedan eliminarse;

c) proporcionar la información requerida conforme al artículo 13 y, cuando proceda, impartir formación a los responsables del despliegue.

Con vistas a eliminar o reducir los riesgos asociados a la utilización del sistema de IA de alto riesgo, se tendrán debidamente en cuenta los conocimientos técnicos, la experiencia, la educación y la formación que se espera que posea el responsable del despliegue, así como el contexto en el que está previsto que se utilice el sistema.

6. Los sistemas de IA de alto riesgo serán sometidos a pruebas destinadas a determinar cuáles son las medidas de gestión de riesgos más adecuadas y específicas. Dichas pruebas comprobarán que los sistemas de IA de alto riesgo funcionan de manera coherente con su finalidad prevista y cumplen los requisitos establecidos en la presente sección.

7. Los procedimientos de prueba podrán incluir pruebas en condiciones reales de conformidad con el artículo 60.

8. Las pruebas de los sistemas de IA de alto riesgo se realizarán, según proceda, en cualquier momento del proceso de desarrollo y, en todo caso, antes de su introducción en el mercado o puesta en servicio. Las pruebas se realizarán utilizando parámetros y umbrales de probabilidades previamente definidos que sean adecuados para la finalidad prevista del sistema de IA de alto riesgo.

9. Cuando se implante el sistema de gestión de riesgos previsto en los apartados 1 a 7, los proveedores prestarán atención a si, en vista de su finalidad prevista, es probable que el sistema de IA de alto riesgo afecte negativamente a las personas menores de dieciocho años y, en su caso, a otros colectivos vulnerables.

10. En el caso de los proveedores de sistemas de IA de alto riesgo que estén sujetos a requisitos relativos a procesos internos de gestión de riesgos con arreglo a otras disposiciones pertinentes del Derecho de la Unión, los aspectos previstos en los apartados 1 a 9 podrán formar parte de los procedimientos de gestión de riesgos establecidos con arreglo a dicho Derecho, o combinarse con ellos.

3.3 Correspondencia del articulado con los apartados de la guía

En la tabla dispuesta a continuación se detalle en que secciones de esta guía se abordan los diferentes elementos de dicho artículo:

Artículo Reglamento	Requerimiento Reglamento	Sección guía
9.1	¿Qué elementos debo implantar y cómo debo hacerlo para para desarrollar un adecuado sistema de gestión de riesgos?	Apartado 4
9.2.a	Identificación de riesgos Análisis y evaluación de riesgos	Apartado 4.3 y 4.4
9.2.b	Identificación de riesgos Análisis y evaluación de riesgos	Apartado 4.3 y 4.4
9.2.c	Evaluación de riesgos relacionados con el sistema de vigilancia poscomercialización	Apartado 5.2
9.2.d	Respuesta al riesgo	Apartado 5.5
9.3		
9.4		
9.5.a		
9.5.b		
9.5.c		
9.6	Pruebas destinadas a comprobar que los HRAIS funcionan conforme a su finalidad prevista y cumplen con los requisitos de gestión de riesgos	Apartado 5.1.1 y 5.1.2
9.7	Pruebas en condiciones reales	Apartado 5.1.3
9.8	Momento adecuado para la realización de las pruebas	Apartado 5.1.4
9.9	Acceso e impacto del sistema por menores de 18 años	Apartado 5.1
9.10	Entidades sujetas a legislación sectorial	Apartado 5.4

4. ¿Qué elementos debo implantar y cómo debo hacerlo para desarrollar un adecuado sistema de gestión de riesgos?

Un **sistema de gestión de riesgos** se basa en la implementación de **procesos** que permiten **identificar y evaluar los riesgos** y la **definición e implementación** de **medidas de tratamiento** oportunas para **gestionar** el efecto de estos.

Como ya hemos indicado previamente, los riesgos principales que deberemos gestionar son los que puedan afectar a **la salud, la seguridad y los derechos fundamentales de las personas.**

En el esquema dispuesto a continuación, se detallan los **procesos** que completan un **sistema de gestión de riesgos** y, luego, trataremos de explicar en qué consisten y cómo implementarlos.

4.1 Determinación del apetito al riesgo

¿Qué es?

En términos generales, es el nivel de riesgo que se está dispuesto a aceptar.

En el contexto del Reglamento Europeo de la IA, es el nivel de riesgo que estaremos dispuestos a aceptar en relación con el riesgo que el sistema de IA puede suponer para **la salud, la seguridad y los derechos fundamentales de las personas.**

¿Cómo debo abordarlo?

Para ello, deberíamos establecer el apetito al riesgo. Es decir, si nuestro sistema de IA puede tener un impacto crítico sobre la vida de una persona (por ejemplo, el sistema de IA que administra insulina a los pacientes) el apetito al riesgo debería ser muy bajo. En cambio, si nuestro sistema de IA no tuviese ningún impacto sobre **la salud, la seguridad y los derechos fundamentales de las personas** (por ejemplo, un sistema de IA utilizado

para la recomendación de películas o series, en base a nuestros gustos y preferencias) el apetito al riesgo podría ser mucho más alto.

El apetito al riesgo debe definirse de forma cuantitativa. En primer lugar, se establece una escala y después se selecciona el nivel de riesgo en dicha escala.

Generalmente, la escala se define estableciendo un valor mínimo (normalmente igual a 1) y un valor máximo. El valor máximo viene determinado por el valor máximo del riesgo que, como hemos visto en la sección 2.1, es igual al producto del impacto por la probabilidad.

Como allí mencionábamos, el proceso de evaluación del riesgo lo veremos en detalle en la sección 4.4. No obstante, supongamos que seleccionamos una escala de 3 niveles para la probabilidad de que suceda el riesgo y 5 niveles para el impacto que podría generar. En este caso, el valor máximo es igual a 15 y en dicha escala deberemos determinar cuál es el apetito al riesgo (siendo un valor en torno a 12 alto y un valor en torno a 3 bajo).

Ejemplo

En nuestro ejemplo de la sección anterior, decíamos que la inadecuada promoción de unos empleados frente a otros podría suponer un **riesgo** para los **derechos fundamentales** de éstos, en concreto si el sistema de IA diese lugar a promociones basadas en **decisiones discriminatorias**. Supongamos que hemos determinado (esto lo haremos en la sección 4.4) que este riesgo, en caso de suceder, resultaría en un nivel 9 sobre 15 (probabilidad igual a 3 e impacto igual a 3). Si hemos definido un **umbral de apetito al riesgo** igual a 4 (ya que no estamos dispuestos a aceptar mucho riesgo dado el potencial impacto del sistema sobre los **derechos fundamentales** de los trabajadores) significa que no estamos dispuestos a asumir el nivel de riesgo que éste supone así que tendremos que determinar las correspondientes **medidas de control** hasta que ese nivel deje de estar por encima del umbral definido (todo este proceso se abordará con más detalle en la sección 4.5).

Adicionalmente, para obtener más detalle de cómo se aborda este ejercicio y dónde y cómo queda reflejado, se aconseja consultar el documento Excel señalado en la sección 1.4.

¿Por qué es importante?

Es una de las bases de la gestión de riesgos, ya que es la forma que tenemos de cuantificar el nivel de riesgo que aceptaremos tras analizar y evaluar los riesgos que identifiquemos en fases posteriores. Es el elemento que nos permitirá valorar las medidas de tratamiento del riesgo que necesitemos y el punto en el que éstas sean suficientes para garantizar un nivel de riesgo que estemos dispuestos a aceptar. Sin determinar el apetito al riesgo, la gestión de riesgos se reduciría a la etapa de identificación de los riesgos, pues si no podemos determinar cómo puede afectar su impacto, toda la fase de análisis y evaluación de los riesgos perdería sentido y tampoco se podría establecer medidas de tratamiento del riesgo.

4.2 Contexto del sistema de IA

¿Qué es?

Es el entorno externo e interno, donde diseñar, desarrollar y utilizar el sistema de IA. En el Anexo A.I. Se dispone un listado de algunos ejemplos relevantes (*entre ellos se encuentran, los factores económicos y regulatorios, el contexto tecnológico de la organización, el nivel de madurez y complejidad de los sistemas de IA en la organización o la cultura en torno al dato y el uso de datos en la organización*).

Sin embargo, los elementos más importantes de este contexto en el que se diseña, desarrolla y utiliza el sistema de IA son aquellos relacionados con garantizar **la salud, la seguridad y los derechos fundamentales de las personas.**

Por ello, hemos incorporado en el Anexo A.II un listado con los principales derechos de la Carta de los Derechos Fundamentales de la Unión Europea. Además, se recoge una descripción general de dicha Carta y un breve ejemplo para aquellos derechos más relevantes desde la perspectiva del Reglamento Europeo de IA de acuerdo con el considerando 28.

¿Cómo debo abordarlo?

Lo que haremos en esta fase será inventariar en un documento (una hoja Excel, por ejemplo) aquellos elementos del contexto que pudiesen impactar al análisis de riesgos, poniendo especial atención a los relacionados con **la salud, la seguridad y los derechos fundamentales de las personas**. Es importante tener en cuenta que este proceso es un proceso mayoritariamente cualitativo y subjetivo y cuya precisión y completitud dependerá de nuestro conocimiento de esos elementos del entorno.

Una vez inventariados estos elementos, tendremos que ver cómo pueden afectar al análisis de riesgos, es decir, deberemos plantearnos la siguiente pregunta, ¿pueden estos elementos suponer un riesgo adicional que deba incorporar en mi proceso de identificación de riesgos? (señalar de nuevo que este proceso lo veremos más en detalle en la sección 4.3)

Ejemplo

Siguiendo con el ejemplo desarrollado en las secciones anteriores, deberemos identificar los elementos más relevantes del contexto en el que se diseña, desarrolla y utiliza el **sistema de IA de promoción de empleados**. Como hemos visto, deberemos prestar especial atención aquellos que puedan estar relacionados con **garantizar la salud, la seguridad y los derechos fundamentales de las personas.**

En este sentido, uno de los elementos más importantes del contexto de un sistema de IA utilizado para promocionar empleados es el cumplimiento con el **derecho fundamental a la no discriminación**. También lo es, el cumplimiento con el derecho fundamental relacionado con la **protección de datos de carácter personal**. Estos elementos, deberemos identificarlos e inventariarlos en la documentación elaborada para el sistema de gestión de riesgos y son la base que nos ayudará a la adecuada identificación de las principales amenazas como veremos en las siguientes secciones.

Adicionalmente, para obtener más detalle de cómo se aborda este ejercicio y dónde y cómo queda reflejado, se aconseja consultar el documento Excel señalado en la sección 1.4.

¿Por qué es importante?

La comprensión del entorno y el contexto en el que se desarrolla el sistema de IA es la base para la identificación de los riesgos que podrían afectar a **la salud, la seguridad y los derechos fundamentales de las personas.** Por ello, resulta imprescindible analizar este elemento con la debida precisión y profundidad, ya que una inadecuada evaluación del contexto y del entorno resultará en una incompleta identificación de riesgos.

4.3 Identificación de riesgos

¿Qué es?

Es el proceso de descubrimiento, reconocimiento y documentación de los diferentes riesgos que pueden afectar a nuestro sistema de IA y, por consiguiente, a **la salud, la seguridad y los derechos fundamentales de las personas.**

¿Cómo debo abordarlo?

El objetivo de esta fase será completar un inventario (por ejemplo, en un documento Excel) con todos aquellos posibles riesgos que consideremos que pueden afectar a nuestro sistema de IA. Para ello, seguiremos un procedimiento como el descrito a continuación (se incluyen breves ejemplos en cada etapa descrita, si bien, como se indica al final del apartado, el ejercicio completo se dispone en el Excel indicado):

1. En primer lugar, deberemos **identificar los componentes de nuestro sistema de IA**. Estos componentes, definen cada sistema y lo diferencian del resto.
   Para facilitar la identificación de los componentes de un sistema de IA, en el Anexo B se ha incorporado un listado con algunos de los componentes más comunes de los sistemas de IA.
2. El segundo paso consiste en **identificar las amenazas asociadas a dichos componentes** a través del análisis de contexto interno y externo. Los riesgos afectan a nuestro sistema de IA a través de sus componentes (ver ejemplo a continuación).
3. Estas **amenazas pueden materializarse en riesgos a través de la explotación de vulnerabilidades de componentes** de nuestro sistema.

Ejemplo

En nuestro ejemplo, identificamos como **elementos del contexto** la necesidad del cumplimiento con el **derecho fundamental a la no discriminación** y el cumplimiento con el **derecho fundamental** relacionado con la **protección de datos de carácter personal.**

Tal y como indicábamos en la sección anterior, el contexto es la **base** para la identificación de amenazas y en consiguiente, los riesgos. Así, en este ejemplo, vamos a identificar cada uno de los elementos del contexto mencionados con un componente del sistema de IA y con un riesgo asociado:

Sistema	Componente	Amenaza	Riesgo
Sistema de promoción de empleados	Datos de entramiento	Sobrerrepresentación o subrepresentación de un conjunto en la base de datos de entrenamiento	Discriminación de unos empleados frente a otros en la promoción
	Propietario de los datos	Filtración de datos intencionada o no intencionada por parte del propietario	Revelación de datos personales de los empleados

Es importante destacar, una vez más, que el Reglamento Europeo de la IA hace especial énfasis en los **riesgos que puedan afectar a la salud, la seguridad y los derechos fundamentales de las personas.** En este sentido, se deberá poner especial atención a la identificación de cualquier riesgo que pueda afectar a los elementos mencionados. Por ejemplo, en un sistema de IA encargado de administrar de forma automática insulina a un paciente diabético, el riesgo asociado a un fallo en la lectura de los parámetros que permiten determinar la cantidad de insulina que el paciente necesita en un momento dado, tiene un impacto directo sobre la salud y la vida de dicho paciente. En cambio, un riesgo de ese mismo sistema asociado a un fallo a la hora de enviar un reporte mensual de las dosis administradas a lo largo del mes a la aplicación del móvil del paciente no tiene un impacto directo tan representativo como el anterior sobre la salud o la vida del paciente.

Adicionalmente, para obtener más detalle de cómo se aborda este ejercicio y dónde y cómo queda reflejado, se aconseja consultar el documento Excel señalado en la sección 1.4.

¿Por qué es importante?

La identificación de riesgos es importante dado que sólo los riesgos identificados pueden evaluarse y recibir la respuesta adecuada. Cuando una organización no logra identificar el riesgo adecuadamente, éste queda completamente fuera del sistema de gestión de riesgos.

4.4 Análisis y evaluación de riesgos

¿Qué es?

Es el proceso de análisis y evaluación de los riesgos identificados en la etapa predecesora mediante la determinación, para cada uno de ellos, de la probabilidad de la materialización de la amenaza y la magnitud del impacto en el componente.

¿Cómo debo abordarlo?

El objetivo de esta fase es determinar el impacto y la probabilidad de ocurrencia de cada riesgo identificado en la fase anterior. Por ello, tomaremos nuestro documento donde estamos registrando el proceso de gestión de riesgos e incorporaremos dos nuevos atributos, el impacto y la probabilidad (se recomienda consultar y seguir el ejemplo desarrollado de forma detallada en el documento Excel señalado en la sección 1.4).

Ejemplo

En los apartados anteriores, hemos utilizado como ejemplo el **sistema de IA de promoción y ascenso** de los empleados y hemos identificado como posibles riesgos **la discriminación de unos empleados frente a otros en la promoción y la revelación de datos personales de los empleados**.

Lo que tenemos que hacer ahora, es decidir qué **magnitud de impacto le damos a este riesgo**, para ello determinaremos una escala de forma similar a cómo lo hicimos en la sección 4.1 cuando definimos el apetito al riesgo. Y haremos lo mismo con la **probabilidad**.

Llegados a este punto, es probable que nos surjan dudas como, ¿bajo qué criterio decido yo el nivel de impacto que tiene un riesgo?, ¿cómo puedo determinar la probabilidad que existe de que se materialice?, ¿cómo estableceré el valor del riesgo final?, ¿cómo se relaciona con la probabilidad y el impacto?, ¿y con el apetito al riesgo? A continuación, trataremos de dar respuesta a estas cuestiones.

Hay que recordar que la evaluación del riesgo es un ejercicio cualitativo y subjetivo, que debe facilitar conocer los sistemas de IA que se tienen y los principales riesgos que podrían afectarles y, por tanto, **a la salud, la seguridad y los derechos fundamentales de las personas.**

El valor o nivel de riesgo lo calcularemos como el producto del impacto en el componente por la probabilidad de que la amenaza suceda. No se trata de cuantificar el impacto y la probabilidad de que se materialice una amenaza con precisión numérica, sino valorarlos en una escala conceptual, y ser coherente en su aplicación a todos los casos. Así, por ejemplo, podemos definir una escala del 1 al 5 para categorizar los riesgos en 5 niveles de impacto (en orden ascendente, muy bajo o inexistente, bajo, medio, alto, muy alto o crítico) y una escala del 1 al 3 para categorizar la probabilidad de que éstos sucedan (en orden ascendente, improbable, probable, muy probable). Dicho esto, el máximo nivel de riesgo que me podría encontrar sería igual a 15. Debemos considerar que esta escala debe ser la

misma que la definida para el apetito al riesgo, y lo que haremos a continuación es compararlos.

Una vez establecidos los niveles de impacto y probabilidad y calculado el valor final del riesgo, debemos compararlo con el apetito al riesgo determinado inicialmente.

Ejemplo

Supongamos que, para el **riesgo** relacionado con **revelación de datos personales de los empleados**, hemos determinado que tiene una **probabilidad** de suceder de 3 en la escala del 1 al 3 definida. ¿Por qué 3?, por ejemplo, porque no tenemos **ningún control de acceso** a la base de datos que contiene el conjunto de datos de entrenamiento **ni controles de ciberseguridad** que la protejan de agentes maliciosos externos.

En caso de suceder, determinamos que éste tendría un **impacto grave sobre los empleados**, ya que atentaría contra uno de los **derechos fundamentales**, la **protección de datos de carácter personal**. Digamos, entonces, que un impacto 4 en la escala definida del 1 al 5.

El **resultado del análisis de este riesgo** sería igual a 12 (producto de 3 por 4). Si habíamos definido un **apetito al riesgo** igual a 4, estaríamos por encima del umbral y por ende no estaríamos dispuestos a asumir este riesgo.

¿Qué debemos hacer ahora?, decidir qué tipo de respuesta al riesgo es la que mejor se ajusta a este escenario (evitar, transferir, mitigar, aceptar). Por ejemplo, mitigarlo estableciendo medidas (como controles de acceso o ciberseguridad) que lo reduzcan hasta que baje por debajo del umbral de apetito definido. En la sección 4.5 ahondaremos en las diferentes medidas y tratamientos del riesgo que podemos abordar.

Adicionalmente, para obtener más detalle de cómo se aborda este ejercicio y dónde y cómo queda reflejado, se aconseja consultar el documento Excel señalado en la sección 1.4.

¿Por qué es importante?

El análisis y evaluación de riesgos permite cuantificar el nivel de riesgo para nuestro sistema y, por consiguiente, **a la salud, la seguridad y los derechos fundamentales de las personas,** los riesgos identificados en la etapa anterior.

Sin este proceso de evaluación de los riesgos no tendríamos la capacidad de determinar, cuantitativamente si los aceptamos o, por el contrario, necesitásemos definir e implementar medidas de tratamiento para gestionarlos adecuadamente.

4.5 Respuesta al riesgo

¿Qué es?

Es el proceso de selección e implementación de medidas para abordar los riesgos identificados y analizados en las etapas anteriores.

¿Cómo debo abordarlo?

En esta fase vamos a definir las medidas de respuesta al riesgo que incorporaremos en nuestro documento donde estamos registrando el proceso de gestión de riesgos. Para ello seguiremos un procedimiento como el descrito a continuación (se incluyen breves ejemplos en cada etapa descrita, si bien, como se indica al final del apartado, el ejercicio completo se dispone en el Excel indicado en la sección 1.4):

1. Primero debemos **determinar qué tipo de medidas de tratamiento de riesgos** seleccionaremos para abordar cada uno de los riesgos identificados y analizados. Las opciones principales son las que indicamos a continuación:

    a. **Implementar medidas de control para mitigar el riesgo:** Normalmente la primera opción que nos planteemos a la hora de tratar un riesgo cuyo valor supera nuestro apetito al riesgo será implementar controles adicionales para mitigar su probabilidad o impacto. En el Anexo D se disponen algunos ejemplos de controles específicos del ámbito de la IA.

Ejemplo

Siguiendo con nuestro ejemplo del **sistema de IA** para la **promoción de empleados** y el **riesgo** identificado asociado con la **revelación de datos personales de los empleados**, una medida que mitigaría este riesgo es la implementación de un sistema RBAC (*Role Based Access Control*) o sistema de control de accesos y administración de los derechos sobre los datos.

Con este control conseguiremos reducir la probabilidad de que suceda el riesgo. En el ejemplo más detallado de la sección 1.4 veremos cómo se reduce el valor del riesgo mediante la implementación de estos controles, pero supongamos por ahora que dicho valor se reduce del valor original 12 a un valor igual a 4, en la misma escala del 1 al 15.

    b. **Asumir el riesgo**, aceptando las consecuencias que éste tendría en caso de suceder.

Ejemplo

En el apartado anterior veíamos como, tras implementar una **medida de mitigación del riesgo**, éste se reducía de un valor 12 a un valor 4. Inicialmente habíamos definido un **apetito al riesgo** igual a 4, ahora nos encontramos por debajo de dicho umbral.

En este escenario, podríamos implementar más medidas de control para seguir reduciendo el riesgo. Pero también podríamos determinar que, dado que el valor del **riesgo actual** ya está **por debajo** de nuestro **apetito al riesgo, no implementar medidas de control adicionales** y aceptar el nivel de riesgo actual.

En el ejemplo más detallado de la sección 1.4 veremos casos representando este escenario y cómo se releja en el documento que recoge nuestro sistema de gestión de riesgos.

c. **Evitar el riesgo** mediante la decisión de no iniciar o descontinuar la actividad que genera el riesgo.

Ejemplo

Siguiendo con nuestro ejemplo, tras analizar los riesgos asociados a nuestro sistema de IA para la promoción de empleados, podemos encontrarnos en un **escenario donde el número de riesgos**, su probabilidad de suceder y sus respectivos impactos sobre **la salud, la seguridad y los derechos fundamentales de las personas** sean tan elevados, que nos hagan **plantearnos** que quizá no nos interese **seguir utilizando el sistema** o descontinuarlo.

d. **Transferir el riesgo**, por ejemplo, a través de contratos, como la compra de seguros.

Ejemplo

En el contexto del Reglamento de la IA, cabe destacar que la gestión del riesgo, como ya hemos comentado a lo largo de la guía, se centra en la gestión aquellos riesgos que puedan afectar a **la salud, la seguridad y los derechos fundamentales de las personas**.

La transferencia del riesgo es una medida, por su naturaleza, aplicable en un contexto de gestión de riesgos para la organización. Por ejemplo, supongamos que el **riesgo** asociado a la **una posible pérdida de información de la organización** es **tan elevado** que decidimos **contratar** los servicios de una **compañía externa** que velará por **garantizar** que nuestros **datos no son manipulados** y que llegamos a un acuerdo con dicha compañía según el cual en caso de que éstos **sufran una manipulación** indeseada la compañía nos **compensará** con una cuantía económica acordada.

2. En segundo lugar, y una vez definidas las medidas de tratamiento del riesgo debemos establecer una **planificación para su implementación**. Para ello, por ejemplo, podemos establecer mediante un *diagrama de Gantt* la secuencia de implementación de las medidas definidas en función la prioridad y las fechas de compromiso para implementarlas.
   La priorización de la implementación de las medidas se puede establecer, por ejemplo, en función de su impacto de mitigación, dando especial prioridad a aquellas que mitiguen los riesgos más relevantes.
3. En tercer lugar, deberemos **documentar e informar los riesgos residuales.** Deberemos informar en la documentación asociada a nuestro sistema aquellos riesgos identificados y analizados. Para la adecuada elaboración de esta documentación se deberá consultar la guía de *transparencia y provisión de información a los usuarios* dispuesta para ello, tal y como se indica en el Reglamento, en el apartado 5c del artículo: "Sistema de gestión de riesgos").
4. Finalmente, estableceremos **periodos de revisión y seguimiento** del sistema de gestión de riesgos (se deben incorporar todos aquellos nuevos riesgos que se identifiquen, así como evaluar y definir las medidas de tratamiento oportunas). También deberemos comunicar los resultados y actividades de la gestión del riesgo, así como las actualizaciones derivadas a lo largo de la organización (especialmente a las partes interesadas y/o impactadas).

Adicionalmente, para obtener más detalle de cómo se aborda este ejercicio y dónde y cómo queda reflejado, se aconseja consultar el documento Excel señalado en la sección 1.4.

¿Por qué es importante?

Es la fase donde, tras identificar, analizar y evaluar los riesgos que amenazan a nuestro sistema de IA, establecemos las medias oportunas para gestionarlos.

Sin implementar medidas de respuesta al riesgo, el sistema de gestión de riesgos se quedaría en una fase de conocimiento de los riesgos amenazantes y, en el peor de los casos, de conocimiento de que estos superan el nivel de riesgo que estamos dispuestos a asumir. Esta fase es la que nos permite, una vez determinado todo ello, enfrentarnos a los riesgos y tratar de reducir o mitigar su impacto, en definitiva, darles una respuesta.

4.6 Documentación técnica del sistema de gestión de riesgos

Para la adecuada documentación del sistema de gestión de riesgos, deberemos seguir los pasos descritos en la sección 4 y reflejarlos en un documento igual o similar al facilitado como ejemplo en la sección 1.4. Adicionalmente, poner especial atención a la documentación de los riesgos residuales, tal y como se explica en la sección 4.5 apartado 3.

Para más detalle, ver sección 6.

4.7 Comunicación y consulta

Los actores involucrados en el diseño, desarrollo y comercialización del sistema de IA deben conocer el sistema de gestión de riesgos del sistema de IA y deberán colaborar en su completitud y actualizaciones.

La comunicación busca promover la toma de conciencia y la comprensión del riesgo.

La consulta implica obtener retroalimentación e información para apoyar la toma de decisiones.

La comunicación y consulta con las partes interesadas apropiadas, tanto externas como internas, se deberían realizar en todas las etapas del proceso de la gestión del riesgo.

En el caso de las organizaciones que desarrollan o utilizan sistemas de IA, éstas deben identificar qué partes de la organización están involucradas. La organización debe ser consciente de que la aplicación de las tecnologías de IA puede tener un mayor impacto que la aplicación de otras tecnologías y prestar especial atención a aquellos impactos sobre **la salud, la seguridad y los derechos fundamentales de las personas.**

4.8 Seguimiento y mejora continua

Su finalidad es asegurar y mejorar la calidad y la eficacia del sistema de gestión de riesgos. Es imprescindible desarrollar un seguimiento y establecer unos periodos determinados de revisión y actualización del sistema de gestión de riesgos. De este modo, se deberá revisar el inventario de riesgos e incorporar todos aquellos adicionales que puedan identificarse. También se deberá abordar y revisar el análisis y evaluación de los nuevos riesgos y de los ya existentes. Por último, las medidas de mitigación oportunas también deberán revisarse y actualizarse en consecuencia.

Como comentábamos al inicio de la guía, el proceso de gestión de riesgos debe abordarse en todas las etapas del ciclo de vida del sistema de IA, desde el diseño y desarrollo hasta su comercialización y poscomercialización. Por tanto, el seguimiento y actualización de los riesgos deberá realizarse también en cada etapa.

4.9 Liderazgo y compromiso

La gestión de riesgos es el proceso por el cual la dirección de una organización: supervisa, lidera, se compromete con el desarrollo del sistema de gestión de riesgos y garantiza la integración de los diferentes sistemas de gestión de riesgos de la organización (por ejemplo, el de la gestión de riesgos de los sistemas de IA con el resto).

Formalizar el liderazgo y compromiso con la gestión de riesgos es otro de los pilares fundamentales en el desarrollo de un sistema de gestión de riesgos.

La alta dirección y los órganos de supervisión, siempre que apliquen, deberán asegurar que la gestión del riesgo del ciclo de vida de los sistemas de IA está integrada con la gestión del riesgo del resto de la organización. Esto lo pueden llevar a cabo mediante, por ejemplo:

a) **La publicación de una declaración o política** que establezca un enfoque, plan o línea de acción para la gestión del riesgo descrita en el ANEXO G.
b) **El aseguramiento que existen recursos suficientes** asignados a la gestión del riesgo.
c) **La asignación de autoridad, responsabilidad y rendición de cuentas** en los niveles apropiados dentro de la organización.

5. Otros elementos a considerar

5.1 Procedimientos de prueba

El Reglamento, al regular el sistema de gestión de riesgos, establece que hay que probar el sistema para asegurar que cumple su propósito y que se cumplen los requisitos establecidos para los HRAIS. A este efecto, permite que se hagan pruebas en condiciones reales.

Estas pruebas deberán diseñarse en función de parámetros, métricas y umbrales definidos de acuerdo con la finalidad prevista del sistema IA.

5.1.1 Pruebas destinadas a comprobar que los HRAIS funcionan conforme a su finalidad prevista

Para garantizar que el sistema de IA funciona conforme a su finalidad prevista, se deberán consultar las *guías de precisión y solidez* desarrolladas para facilitar el cumplimiento con del artículo "Precisión, solidez y ciberseguridad" del Reglamento.

Estas guías definen medidas y métricas que permiten valorar los parámetros necesarios para asegurar que el sistema está funcionando de la forma prevista y continuará haciéndolo a lo largo del tiempo.

5.1.2 Pruebas destinadas a comprobar que los HRAIS cumplen con los requisitos establecidos

Para garantizar que las medidas desarrolladas e implantadas son las adecuadas, se podrán establecer indicadores y métricas que permitan la medición y validación de la efectividad de estas medidas.

Una forma de abordar este aspecto es mediante la definición de **indicadores de efectividad**. Un indicador de efectividad es una métrica que facilita la medición del grado en el que se ha cumplido o alcanzado un objetivo determinado.

Para desarrollar las pruebas que ayuden a garantizar que los requisitos del capítulo 2 se cumplimenten adecuadamente, se deberá comprender y abordar lo establecido en cada uno de ellos. Se deberán consultar las guías elaboradas para ese fin (*datos y gobierno de datos, documentación técnica, conservación de registros, transparencia y provisión de información a usuarios, supervisión humana, precisión, solidez y ciberseguridad).*

Adicionalmente, en estas guías se incorporará un inventario de pruebas o indicaciones que ayudarán al lector a verificar que ha cubierto los aspectos establecidos en el artículo correspondiente.

Para guiar al lector en la definición de dichos indicadores, en el Anexo E se han incorporado algunos ejemplos de indicadores de efectividad con relación a las medidas de gestión de

riesgos descritas en la presente guía y otras adicionales relacionadas con el resto de los requisitos.

5.1.3 Pruebas en condiciones reales

Los procedimientos de prueba que se diseñen podrán incluir pruebas en condiciones reales, en caso de que la organización lo considere oportuno o necesario, siguiendo lo establecido en el artículo "*Pruebas de sistemas de IA de alto riesgo en condiciones reales fuera de los espacios controlados de pruebas para la IA*".

En dicho artículo se disponen las condiciones bajo las que los proveedores de los HRAIS del Anexo III del Reglamento Europeo de la IA podrán realizar pruebas en condiciones reales.

En los puntos listados a continuación se disponen estas condiciones de forma sintetizada:

- Las pruebas en condiciones reales deberán realizarse siempre antes de la comercialización o puesta en servicio del sistema.
- Cualquier sujeto podrá retirarse de la prueba en cualquier momento revocando su consentimiento informado sin tener que dar ninguna justificación.
- Cualquier incidente grave detectado en el transcurso de las pruebas se comunicará a las autoridades oportunas indicadas en el apartado 6 del artículo.
- El proveedor y el posible proveedor serán responsables, en virtud de la legislación de la Unión y de los Estados miembros en materia de responsabilidad, de cualquier daño causado a los sujetos por su participación en las pruebas en condiciones reales.

5.1.4 Momento adecuado para la realización de las pruebas

Las pruebas diseñadas e implementadas para el desarrollo de los sistemas de IA pueden realizarse en cualquier momento de su desarrollo, pero siempre antes de su comercialización o puesta en servicio.

No obstante, se deberá considerar que las pruebas se deben planificar durante el diseño, llevarse a cabo durante la implementación y estar alineadas con la finalidad prevista y los riesgos identificados antes de la puesta a producción.

5.2 Evaluación de riesgos relacionados con el sistema de vigilancia poscomercialización

El desarrollo del sistema de gestión de riesgos deberá tener en consideración la evaluación de otros riesgos que puedan surgir basándose en el análisis de los datos recogidos en el sistema de seguimiento poscomercialización mencionado en el artículo "*Vigilancia poscomercialización por los proveedores y plan de vigilancia poscomercialización para sistemas IA de alto riesgo*".

En este artículo se describe el seguimiento que los proveedores tendrán que realizar tras la comercialización del sistema IA y los elementos principales que deberán contener el plan de seguimiento tras la comercialización que deberán elaborar para dichos sistemas.

Para la adecuada comprensión y entendimiento de este proceso, se ha elaborado la guía que describe el plan de vigilancia poscomercialización a elaborar por los proveedores (*guía para la elaboración de un sistema de vigilancia poscomercialización*).

Los proveedores de los sistemas de IA deberán evaluar otros riesgos que pudieran surgir a partir del análisis de los datos recogidos en el sistema de seguimiento posterior a la comercialización.

Para llevar a cabo esta evaluación de riesgos se deberán seguir los pasos descritos en la sección 4, de forma que:

a) **En el proceso de identificación** se incluyen los riesgos que pudieran surgir a partir del análisis de los datos recogidos en el sistema de seguimiento posterior a la comercialización. Deberá considerar y analizar el contexto y los entornos en los que el sistema podrá utilizarse.
   Quien comercialice un sistema de IA encargado de administrar de forma automática insulina a un paciente diabético, deberá tener en consideración los diferentes riesgos que puedan originarse durante el funcionamiento y uso del sistema por parte del paciente.
b) **En el proceso de análisis y evaluación** de riesgos se incorporen los riesgos identificados en el punto anterior.
c) **En el proceso de adopción de las medidas y controles** para la gestión de los riesgos se tengan en consideración los riesgos identificados, analizados y evaluados en los puntos anteriores.

5.3 Acceso e impacto del sistema por menores de 18 años

Durante el desarrollo del sistema de gestión de riesgos se deberá determinar si es probable que menores de 18 años accedan al sistema, o éste tenga un impacto sobre ellos.

En este contexto, se deberán analizar e incorporar al sistema de gestión de riesgos diseñado los posibles riesgos derivados que puedan tener un impacto sobre dichos menores de 18 años. En este sentido, y de forma similar al apartado anterior, deberemos:

a) **En el proceso de identificación** de riesgos, incluir aquellos que pudieran derivarse del uso o impacto sobre menores de 18 años.
   Por ejemplo, si tenemos un sistema de IA que ayuda en procesos de concesión de ayudas sociales, deberemos considerar si dentro del conjunto de afectados por dicho sistema hay menores de 18 años. Otro ejemplo de sistema de IA que podría afectar a menores de 18 años sería un sistema de IA utilizado para para determinar el acceso o la admisión de personas físicas a programas o centros educativos.

   Según explica la guía, este proceso de identificación de riesgos vendrá determinado por el conocimiento de la organización que desarrolla el sistema de

gestión de riesgos para su sistema de IA del contexto y su entorno en el que se desarrolla.

b) **En el proceso de análisis y evaluación** de riesgos, incorporar los riesgos identificados en el punto anterior.

c) **En el proceso de adopción de las medidas y controles** para la gestión de los riesgos, considerar los riesgos identificados, analizados y evaluados en los puntos anteriores.

5.4 Entidades sujetas a legislación sectorial

Los proveedores de sistemas de IA sujetos a requisitos de los procesos internos de gestión de riesgos según la legislación sectorial pertinente de la Unión, podrán incorporar las medidas descritas de esta guía como parte de los procedimientos de gestión de riesgos de dicha legislación.

6. Documentación técnica

El Artículo 9 establece que los sistemas de inteligencia artificial deberán someterse a un proceso continuo e iterativo de identificación, análisis, evaluación y mitigación de riesgos a lo largo de todo su ciclo de vida. La documentación asociada a dicho proceso deberá reflejar de forma clara y completa cómo se han aplicado estas fases, proporcionando la información necesaria para demostrar la conformidad con los requisitos del reglamento.

De acuerdo con lo anterior, y en coherencia con los apartados pertinentes del Anexo IV, la documentación del sistema de gestión de riesgos deberá incluir los elementos que resulten relevantes para describir adecuadamente el marco, las metodologías empleadas, las decisiones adoptadas y los resultados obtenidos.

Para la adecuada documentación del sistema de gestión de riesgos, deberemos seguir los pasos descritos en la sección 4 y reflejarlos en un documento igual o similar al facilitado como ejemplo en la sección 1.4. Adicionalmente, se deberá prestar especial atención a la documentación de los riesgos residuales, tal y como se explica en la sección 4.5, apartado 3.

Con el objetivo de facilitar la comprensión del desarrollo de un sistema de gestión de riesgos y su documentación, se adjunta el documento "Ejemplo ilustrativo del desarrollo de un sistema de gestión de riesgos.xlsx", que muestra un ejemplo práctico para varios casos de uso. En este documento se detalla el desarrollo para los casos de uso descritos en la sección 4.3, siguiendo cada una de las fases de la guía, desde la definición del apetito al riesgo, hasta la selección de medidas de respuesta.

El documento incorpora explicaciones que vinculan la guía y las fases descritas con el desarrollo del sistema de gestión de riesgos en cada ejemplo.

7. Cuestionario de autoevaluación

Para realizar una autoevaluación del cumplimiento de los requisitos del Reglamento de Inteligencia Artificial referidos en esta guía, se ha generado un cuestionario de autoevaluación global con una serie de preguntas con los puntos clave a tener en cuenta respecto a las obligaciones que dictaminan los artículos del Reglamento de IA mencionados en esta guía.

Será necesario referirse a ese documento para realizar el apartado del cuestionario de autoevaluación correspondiente a esta guía.

8. Anexos

8.1 Anexo A - Elementos más relevantes del contexto interno y externo entorno a la IA

8.1.1 Anexo A.I - Elementos generales

A continuación, se dispone un listado detallado de los elementos más relevantes que completan este entorno [1] [2]:

a) **Los factores sociales, culturales, políticos, legales, regulatorios, financieros, tecnológicos, económicos y ambientales**. Puede ser útil usar directrices y guías sobre cuestiones éticas en la IA, como las guías para una IA ética y confiable publicadas por la CE en 2019.

b) **Las principales tendencias tecnológicas, avances en áreas de la IA y las implicaciones sociales y políticas** del despliegue de estas tecnologías que pueden afectar a nuestro sistema y, en consecuencia, **a la salud, la seguridad y los derechos fundamentales de las personas.**

c) **Las principales partes interesadas, sus percepciones, valores, necesidades y expectativas.** Éstas pueden verse afectadas por cuestiones como la falta de transparencia de los sistemas IA o por los sistemas IA sesgados.

d) **La complejidad de las redes y sus dependencias**, que pueden aumentar con el uso de tecnologías de IA.

e) **Los factores internos organizacionales** que giran en torno a la visión, la misión, los valores, la cultura, la estrategia, el modelo de gobierno, las políticas, normas y procedimientos adoptadas y las relaciones contractuales y compromisos.

f) **La cultura de la organización**, así como guías, modelos y normas adoptadas.

g) **Las capacidades de la organización, recursos y conocimientos** en relación con la IA, es importante tener en consideración problemas de transparencia de los sistemas de IA, la variación de los recursos necesarios y los requisitos de conocimientos específicos en tecnologías de IA y ciencia de datos, éstos pueden ser causas de riesgos adicionales.

h) **El uso de datos y flujos de información**. Los sistemas de IA pueden utilizarse para automatizar, optimizar y mejorar el tratamiento de los datos.

i) **Las relaciones con las partes interesadas internas**, teniendo en cuenta sus percepciones y valores. La percepción de las partes interesadas puede verse afectada por cuestiones como la falta de transparencia de los sistemas de IA o los

sistemas de IA sesgados. Las partes interesadas deben recibir información sobre las capacidades, los modos de fallo y la mitigación de los fallos de los sistemas de IA. Si bien esta vertiente incluye a toda la organización a nivel interno, debe considerarse especialmente en el ámbito de los sistemas de IA, los departamentos y profesionales involucrados en la concepción, implementación, explotación y evolución de los sistemas de IA.

8.1.2 Anexo A.II - Elementos relacionados con la Carta de los Derechos Fundamentales de la UE

Las consecuencias adversas que puede generar un sistema IA de alto riesgo para los derechos protegidos por la Carta de Derechos Fundamentales de la Unión Europea [11] pueden ser realmente graves.

Por eso, durante el ciclo de vida de los sistemas de Inteligencia Artificial es necesario adoptar un enfoque que trate de asegurar un nivel elevado de protección de estos derechos.

Concretamente, el considerando 28 del Reglamento Europeo de la IA menciona aquellos derechos fundamentales de la Carta que pueden resultar especialmente afectados por el desarrollo y despliegue de los sistemas de IA de Inteligencia Artificial.

En esta sección se menciona cada uno de los derechos y principios reconocidos por la Carta de Derechos Fundamentales de la Unión Europea que expresamente aparecen mencionados en el Considerando 28 del Reglamento Europeo de la IA.

Adicionalmente, se incorpora un ejemplo de los posibles impactos que puede generar el uso de un sistema de inteligencia artificial en cada uno de estos derechos fundamentales[1]:

1) Derecho a la dignidad humana: Artículo 1 de la Carta.

La dignidad de la persona humana no es sólo un derecho fundamental en sí mismo, sino que constituye la base real o presupuesto de los derechos fundamentales.

No se puede utilizar el ejercicio de los derechos para lesionar la dignidad de otra persona. Asimismo, las restricciones necesarias a los derechos fundamentales deben respetar la dignidad.

Los sistemas de IA de alto riesgo pueden impactar no sólo en derechos fundamentales específicos, sino de modo conjunto en varios e incluso en muchos de ellos. En estos casos, una visión desde la dignidad permite aunar estos riesgos variados. Asimismo, una visión

[1] Para un análisis exhaustivo de cada uno de los derechos fundamentales reconocidos por la Carta de los Derechos Fundamentales se recomienda visitar la página web de la Agencia de los Derechos Fundamentales de la Unión Europea. http://fra.europa.eu/es

Los textos jurídicos que se mencionan en esta sección son los siguientes:

Carta de Derechos Fundamentales de la Unión Europea. (2000)

Observación general n.º 25 del Comité de los Derechos del Niño relativa a los derechos de los niños en relación con el entorno digital. (2021)

Carta de Derechos Digitales Española. (2021)

de la dignidad permite elevar el análisis de riesgos a una visión que trasciende a las posibles afectaciones en derechos individuales de personas específicas, para abordar el impacto que el uso de la IA de alto riesgo puede implicar en una visión en conjunto sobre los derechos de grandes colectivos de personas e incluso de toda la sociedad.

2) Respeto de la vida privada y familiar. Artículo 7 de la Carta.

Las tecnologías de la información y comunicación desde hace décadas vienen impactando especialmente en la vida privada y familiar, del domicilio y de las comunicaciones. En buena medida la vida privada en general y, como se verá, la protección de datos en particular puede quedar especialmente afectada cuando un sistema de IA de alto riesgo se emplea respecto personas concretas, lo cual es lo más habitual.

Posible Vulneración: Anexo III. 1.a) del Reglamento

Especialmente conflictivos son los sistemas de alto riesgo de identificación biométrica que no estén directamente prohibidos. Y, sobre todo impactan en la vida privada los sistemas HRAIS que incorporan un sistema de reconocimiento de emociones para detectar o deducir los estados mentales, las emociones o las intenciones de las personas físicas a partir de sus datos biométricos; o los sistemas que asignan a personas a categorías concretas en función de sus datos biométricos. Pese a que no están prohibidos, la elección del uso de estos sistemas debe estar muy identificada y el análisis de riesgos debe justificar muy bien la necesidad y legitimación de uso de este tipo de sistemas siempre que no haya otra alternativa que impacte menos en la vida privada.

Los ejemplos son muy abundantes. Antes de la regulación del AIA diversos supuestos son ya incompatibles con diversos derechos fundamentales, otros, son dudosamente legales.

En Reino Unido, desde 2017 la Policía de Gales del Sur llevó a cabo un proyecto "AFR Locate" para eventos como la final de la Champions League, partidos internacionales de rugby, conciertos, un día de Navidad en una concurrida calle comercial de Cardiff, etc. Se captaban las imágenes, se procesaban automatizadamente y se contrastaban con personas en listas de vigilancia. Su uso era anunciado con carteles y avisado en redes sociales. En 2019 los tribunales lo dieron por bueno.

En Alemania, en Hamburgo con motivo de una reunión del G20 en 2017 se implantó un sistema de reconocimiento facial a partir de grabaciones para la detección e investigación de delitos. La autoridad de datos de Hamburgo lo consideró inadmisible, pero un tribunal anuló la suspensión.

En Suecia en 2021 la autoridad de datos sancionó que unos policías decidieran por su cuenta utilizar para fines policiales y de investigación un software de reconocimiento facial con imágenes de redes sociales y páginas web.

En Buenos Aires se empezó a implementar en 2019 un sistema con 300 cámaras activas que permitía la identificación de prófugos y generó unos 10 millones de consultas sobre personas no prófugas. Guillermo Federico Ibarrola fue erróneamente identificado como prófugo y estuvo detenido 6 días. El 12 de abril de 2022 el sistema fue suspendido por un juez y finalmente ha sido anulado por sentencia de 7 de septiembre de 2022.

En España, Mercadona fue fuertemente sancionada en 2021 por implantar un sistema inteligente biométrico que controlaba si quienes accedían a algunos establecimientos estaban en sus listas de búsqueda por motivos judiciales previos.

En Brasil, el Metro de São Paulo implantó un sistema de control de seguridad de 4 millones de usuarios diarios. El 7 mayo 2021 el Tribunal de Justicia de **São Paulo** prohibió a la concesionaria del Metro de Sao Paulo que utilizara el "Sistema Digital Interactivo de Puertas" (DID) con reconocimiento facial. El sistema infería emociones, género y edad de las personas para personalizar la publicidad. Finalmente, fue suspendido judicialmente 22 de marzo de 2022.

En EEUU el Agente Virtual Automatizado para la Evaluación de la Verdad en Tiempo Real (AVATAR) analiza el comportamiento no verbal y verbal de los viajeros y al parecer, el sistema también se ha probado en el aeropuerto de **Bucarest**. La Comisión Europea financió el proyecto "Intelligent Portable Control System" (iBorderCtrl), con herramientas de detección del engaño y de evaluación basada en el riesgo que ha generado una relativa reacción desde la sociedad civil. El mismo ha generado una relativa reacción desde la sociedad civil, una iniciativa ciudadana europea y la campaña reclaimyourface.eu.

Los sistemas inteligentes biométricos permiten el reconocimiento de emociones, categorizar a las personas, detectar comportamientos, pensamientos o valorar la personalidad.

Lo cierto es que estos sistemas impactan no sólo en la privacidad y protección de datos de las personas, sino que, como luego se indica, en muchas otras libertades, especialmente por su uso en espacios públicos y el efecto inhibidor que provocan.

Posible Vulneración: Anexo II del Reglamento

Los sistemas de IA de alto riesgo que son productos o componentes de seguridad del Anexo II, como por ejemplo juguetes o embarcaciones de recreo es posible que tengan la capacidad de captar datos de su entorno de naturaleza diversa (imágenes, sonidos, geolocalización, etc.). En el análisis de riesgos habrá de velarse especialmente porque esta captación sea la mínima imprescindible para la funcionalidad y que, en cualquier caso, la información que se extraiga no sea utilizada para cualquier otra finalidad y especialmente para hacer perfilados o evaluaciones de la personalidad.

Dispositivos y sistemas del Anexo II, como los juguetes, ya han generado polémicas por su conexión a internet y extracción de datos incluso imágenes desproporcionada[2].

[2] Véase:

https://www.lavozdegalicia.es/noticia/sociedad/2017/12/21/lista-juguetes-espia-crece-tras-analisis-advierte-riesgos-intolerables-dos-robots/0003_201712G21P28991.htm

La AEPD ha señalado las directrices básicas que deben cumplirse en estos supuestos[3].

También el INCIBE ha señalado los elementos básicos de seguridad a cumplir por los juguetes para evitar ser hackeados[4].

3) Protección de datos de carácter personal. Artículo 8 de la Carta.

Sobre la base de la vida privada y como uno de sus componentes, en las últimas décadas se ha venido reconociendo y desarrollando específicamente el derecho a la protección de datos personales. Todos los contenidos que se derivan de este derecho y su regulación europea y nacional deben quedar garantizados por el sistema de IA de alto riesgo del que se trate. Siempre que un sistema HRAIS trate datos de personas identificadas o identificables en cualquiera de las fases de su ciclo de vida afectará a la protección de datos. La técnica de la gestión de riesgos y evaluación de impacto está especialmente desarrollada en el ámbito de la protección de datos y cabe remitir a los muchos instrumentos para realizar la misma en general y en el caso de sistemas de IA en particular. Debe especialmente llamarse la atención de que, si el sistema de IA de alto riesgo trata datos personales, muy posiblemente será obligatorio realizar no sólo un general análisis de riesgos, sino un especial y exhaustivo estudio de impacto en materia de protección de datos (Artículo 35 RGPD). De igual modo muy posiblemente habrá que tener específicamente en cuenta las garantías específicas relativas al artículo 22 RGPD sobre decisiones automatizadas.

Asimismo, hay que advertir que los análisis de riesgo de protección de datos en los últimos años incorporan no sólo la afectación que tiene un tratamiento de datos personales a la vida privada o familiar o la intimidad, sino que también incluyen análisis de riesgos respecto de otros derechos en juego, como por ejemplo la no discriminación en el tratamiento de datos. De este modo, a la hora de realizar un análisis de riesgos o evaluación de impacto de un sistema de IA de alto riesgo, será especialmente importante para la organización tener en cuenta los instrumentos que se hayan desarrollado para el cumplimiento de la protección de datos. Es más, la mejor práctica será desarrollar conjuntamente estos análisis de riesgos de impacto en derechos y, para ello, integrar, cooperar o coordinar con los sujetos que en la organización tengan especiales atribuciones en la materia, como pueda ser el delegado de protección de datos.

Posible Vulneración: Anexo II y III del Reglamento

Son plenamente válidos los ejemplos indicados previamente mencionados en el derecho al respeto a la vida privada y familiar.

4) Libertades de expresión y de información, de reunión y de asociación. Artículo 11 y 12 de la Carta.

[3] Más información en: https://www.lavozdegalicia.es/noticia/sociedad/2017/12/21/lista-juguetes-espia-crece-tras-analisis-advierte-riesgos-intolerables-dos-robots/0003_201712G21P28991.htm

https://www.aepd.es/es/node/824

[4] Más información en:

https://www.incibe.es/incibe/informacion-corporativa/con-quien-trabajamos/proyectos-europeos/is4k

Los sistemas de IA pueden impactar de muy variadas formas en la libertad de expresión e información. De modo más claro puede apreciarse en los sistemas autónomos para la gestión, control o moderación de contenidos de plataformas. En muchos casos el impacto se puede producir de modo conjunto a las libertades de reunión y de asociación.

Posible Vulneración: Anexo III. 4.a) y Anexo III. 3. b)

Esta afectación puede suceder en general por la interferencia que en su ejercicio puede suponer el uso de sistemas de IA de alto riesgo. En muchos casos, la mera existencia del sistema de IA de alto riesgo puede generar por el efecto inhibidor del ejercicio de estas libertades, esto es, la persona que sabe que hay un sistema que puede captar y evaluar sus pensamientos, manifestaciones y expresiones de los mismos es muy posible que decida no expresarse o participación en actividades o reuniones asociativas, sindicales o reivindicativas.

Así puede suceder por ejemplo en el caso de uso de sistemas de reconocimiento de emociones o evaluación de personalidad incorporado a un sistema de IA de alto riesgo para evaluación de la educación o la selección de empleo. Ello sucederá particularmente en determinados contextos y lugares de uso de estos sistemas o con relación a determinados sujetos (conformadores de opinión en redes, líderes sindicales, periodistas, profesores, investigadores, etc.). Por ello habrá de analizarse específicamente en estos contextos o con estos sujetos el posible impacto y, si el sistema de IA de alto riesgo ha de emplearse, cabrá especialmente controlar y mitigar y disminuir estos riesgos con las garantías y técnicas adecuadas.

5) La no discriminación. Artículo 21 de la Carta.

Cualquier uso de sistemas de IA habitualmente genera errores y sesgos, esto es una anomalía en la salida del sistema debido entre otras razones a: los prejuicios o suposiciones erróneas realizadas durante el proceso de diseño del sistema, prejuicios en los datos de entrenamiento, el propio desarrollo autónomo que ha derivado el despliegue del sistema de IA de alto riesgo y su interacción con el entorno donde se implementa. En algunos casos estos sesgos pueden generar tratos diferentes a personas que deberían ser tratadas igualmente. Es más, los errores o sesgos pueden llevar a tratar de modo a conjuntos de personas a los que está especialmente prohibido discriminar: nacimiento, origen racial o étnico, sexo, religión, convicción u opinión, edad, discapacidad, orientación o identidad sexual, expresión de género, enfermedad o condición de salud, estado serológico y/o predisposición genética a sufrir patologías y trastornos, lengua, situación socioeconómica (art. 2. 1º, Ley 15/2022).

Posible Vulneración: Anexo III del Reglamento.

Los supuestos de discriminación algorítmica son de lo más variado, muchos de ellos involuntarios. Escáneres corporales de aeropuertos en EE.UU. (TSA) ha marcado a los viajeros transgénero como más probables sospechosos para efectuar un control particular. En el **ámbito de la salud**, un sistema aplicado aproximadamente 200 millones de estadounidenses penalizó indirectamente a personas negras, que eran postergados por el sistema respecto de los blancos con enfermedades y necesidades similares. Bajo el sistema COMPAS "Correctional Offender Management Profiling for Alternative Sanctions", las personas negras tienen casi el doble de probabilidades que los blancos de ser etiquetados como de mayor riesgo de reincidir.

En el **ámbito educativo**, en Francia el **sistema de admisión Parcoursup** distribuye estudiantes en establecimientos de educación superior con base en criterios previstos legalmente. Tribunales y autoridades han exigido una especial transparencia. En 2020 en Reino Unido, por el Covid no se realizaron los exámenes A-Level de acceso a la universidad, el organismo regulador (Ofqual) generó un sistema basado en un algoritmo que afectó a casi un millón de personas. La crítica más habitual es que se penalizó a los alumnos de colegios públicos.

6) La protección de los consumidores. Artículo 38 de la Carta.

Afirma la Carta de los Derechos Fundamentales de la UE que "*Las políticas de la Unión garantizarán un alto nivel de protección de los consumidores*", ello deriva en una normativa muy amplia en la UE y los Estados Miembros en este ámbito. Los sistemas de IA de alto riesgo, especialmente vinculados a los productos del anexo II del Reglamento Europeo de la IA, en muy buena medida pueden afectar a los derechos de los consumidores. De igual modo, cabe tener en cuenta las obligaciones que impone especialmente la Ley de mercados digitales, el Reglamento Europeo de la IA que también afecta al uso de sistemas de IA con relación a los consumidores.

Posible Vulneración: Anexo III 4.a) del Reglamento

Especialmente los sistemas de IA de alto riesgo del Anexo II que son componentes de seguridad de productos o que son en sí mismos productos pueden afectar a los consumidores.

También los sistemas de IA de altor riesgo del anexo III relativos a servicios de educación, solvencia de personas físicas o **establecer** su calificación crediticia, fijación de precios, ofrecimiento de servicios, trabajo o productos mediante anuncios personalizados o seguros de vida y de salud. En estos contextos habrá que tener especialmente en cuenta los derechos y garantías de los consumidores.

Los **anuncios personalizados** pueden ofrecer a sus destinatarios una ventaja al informarles sobre determinados servicios, productos u ofertas de trabajo. Sin embargo, estos mismos anuncios pueden afectar a otros colectivos que normalmente han sido discriminados. Por ejemplo, un estudio ha demostrado que **los anuncios de trabajo personalizados** en Facebook **pueden reforzar los estereotipos y prejuicios raciales** y de género en el trabajo. Así, para puestos de cajera de supermercado se mostraban a una audiencia compuesta por un 85% de mujeres.[5]

7) Los derechos de los trabajadores: Artículos 27 a 33 de la Carta

Los derechos de los trabajadores se encuentran expresamente reconocidos en los artículos 27 a 33 de la Carta [11]. El contenido de estos derechos engloba: el derecho de los trabajadores a ser informados y consultados de las decisiones más relevantes que adopte la empresa y que les afecte, el derecho a la negociación y acción colectiva, el derecho de huelga, el derecho a la conciliación laboral y personal, el derecho a un trabajo en

[5] Más información en:

https://arxiv.org/abs/1904.02095

condiciones justas y equitativa, el derecho a la seguridad social o el derecho a obtener una protección adecuada en caso de despido injustificado.

Posible Vulneración: Anexo III. 4.b) del Reglamento

Una posible vulneración del artículo 27 de la Carta (Información y consulta) y la legislación de desarrollo sería que los sistemas de inteligencia artificial utilizados por las empresas no fueran lo suficientemente transparentes para que los empleadores pudieran informar de forma adecuada a los trabajadores o a sus representantes de las reglas e instrucciones en las que se basan los algoritmos cuando estos se utilicen para el despido de un trabajador.

Posible vulneración: Anexo III. 4b) del Reglamento.

Una posible vulneración del artículo 31 de la Carta (Condiciones justas y equitativas) por parte del desarrollador del sistema de IA es el valor o puntuación que establezca para las distintas variables de su algoritmo. En un caso real, un juzgado consideró que existía discriminación por parte de una empresa debido a que el algoritmo que utilizaba puntuaba negativamente de forma similar una ausencia o falta de puntualidad del trabajador con una falta de asistencia por huelga, enfermedad o cuidado de menores a cargo [11]. En este caso, el algoritmo sigue las instrucciones marcadas por la empresa e implementadas en algoritmo. Instrucciones que causan una situación discriminatoria a esos trabajadores.

Posible Vulneración: Anexo III. 4.b) del Reglamento

Una posible vulneración del artículo 31 de la Carta (Condiciones justas y equitativas) es que la empresa establezca un sistema de monitorización constante hacia el trabajador que le obligue a este último a adoptar comportamientos que no son naturales en el centro de trabajo (sonreír continuamente, estar activo en todo momento, estar atento, etc.). Sobre todo, cuando esa observación permanente pueda venir combinada con un despido, una sanción laboral, una reducción de sueldo, etc.

8) Los derechos de las personas discapacitadas: artículo 26 Carta

En buena medida cabe remitir a los derechos fundamentales de los que son titulares las personas discapacitadas y a los posibles impactos que se produzcan por los sistemas de IA de alto riesgo.

Posible vulneración. Anexo III. 1. a) y Anexo III. 5. a) del Reglamento

Una posible vulneración del artículo 26 de Carta puede ocurrir si una organización utiliza un sistema de IA en el cual, durante su diseño no se ha tenido en cuenta que este sistema puede arrojar más imprecisiones para determinadas personas. Por ejemplo, una organización utiliza un sistema de IA para realizar entrevistas de trabajo por video que analiza los patrones del habla de los solicitantes para llegar a conclusiones de su futuro rendimiento en el trabajo. Aquellos solicitantes de ese empleo que tengan una determinada discapacidad para hablar, el sistema posiblemente les otorgará una calificación más baja o inaceptable para el puesto.

En estos supuestos, la vulneración no tiene por qué generarse durante el diseño de la aplicación por parte del proveedor del sistema de IA, es mucho más probable que la vulneración venga de la mano de la organización que utiliza el sistema al no tener en cuenta que ese sistema puede afectar de forma injustificada a determinadas personas con ciertas

discapacitadas. La supervisión humana de estos procesos por parte de la organización usuaria puede compensar esta situación inicial. El proveedor del sistema de IA debería también avisar de las limitaciones que este sistema tiene respecto de determinadas personas con ciertas discapacidades.

9) El derecho a la tutela judicial efectiva y a un juez imparcial. Artículo 47 de la Carta

Los derechos y garantías en el ámbito judicial son los más comprometidos por el uso de sistemas IA de alto riesgo debido a la relevancia de los mismos. Entre otras garantías, este derecho reconoce de forma genérica el acceso a la justicia y el derecho que tiene toda persona a que sus pretensiones sean tenidas en cuenta por los órganos jurisdiccionales. Todo ello en condiciones de igualdad.

Posible vulneración. Anexo III. 8. a) del Reglamento

Una posible vulneración del derecho a la tutela judicial efectiva por parte de un sistema de inteligencia artificial podría suceder si un juez utiliza un sistema de inteligencia artificial para aplicarlo a un caso judicial específico donde el sistema tiene que interpretar una serie de hechos y proponer una solución al caso. Si el sistema no es lo suficiente transparente para facilitar posteriormente la motivación de la decisión judicial que adoptará el juez, es posible que exista dicha vulneración.

Posible vulneración. Anexo III. 8. a) del Reglamento

Una posible vulneración del derecho a la tutela judicial efectiva por parte de un sistema de inteligencia artificial podría suceder este sistema se introduce en la Administración de Justicia para que lo utilicen todos los jueces de un país cuyo objetivo es interpretar una serie de hechos y proponer una solución al caso. Si el sistema ha sido entrenado con sentencias judiciales muy antiguas o sentencias que sólo recogen resoluciones de algunas provincias o regiones y no tienen en cuenta el conjunto de sentencias de todo el país, es posible que no generalice adecuadamente el entorno donde este sistema luego se utilizará.

10) Los derechos de la defensa y la presunción de inocencia: Artículo 28 de la Carta.

Al igual que ocurre con los derechos mencionados en el apartado anterior, los derechos de defensa y presunción de inocencia pueden verse claramente afectados por el uso de sistemas de IA de alto riesgo. La afectación de estos derechos implica importantes riesgos para el propio sistema judicial y policial.

Posible vulneración. Anexo III. 6. a) del Reglamento

Cualquier sistema de IA que tenga como objetivo predecir que una persona presenta un riesgo alto para cometer un delito penal afecta de forma grave a los derechos de defensa y presunción de inocencia. Estos derechos pueden verse vulnerados en la medida que la implantación de dicho sistema de IA se utilice por las autoridades competentes y no se hayan previsto suficientes medidas de rendición de cuentas, así como cierto grado de transparencia en cuanto a la decisión adoptada.

11) El derecho a una buena administración: Artículo 41 de la Carta.

En buena medida, muchos de los derechos y garantías propios del debido proceso quedan integrados para el ámbito de la actuación pública en el derecho a una buena administración.

Posible vulneración. Anexo III. 3. a) y Anexo III. 5. a) del Reglamento

Una posible vulneración del derecho a la buena administración se daría por ejemplo si el uso de sistemas de inteligencia artificial de alto riesgo por parte de las Administraciones Públicas no permite una motivación en lenguaje natural de las decisiones que fundamente y, por ende, el control adecuado de la actividad administrativa basada en estas decisiones. Ello puede darse por falta de transparencia o porque no haya mecanismos de rendición de cuentas adecuados.

Pueden existir determinadas variables que durante el diseño de un sistema de IA a priori no afecten a determinados colectivos, pero una vez se pone en marcha el sistema se aprecia esa afectación. Ello ha ocurrido en EE. UU. con un sistema que evalúa el riesgo de un prisionero de volver a cometer delitos y, en función del resultado obtener beneficios para salir de manera anticipada de la cárcel[6]

12) El derecho a una buena administración: Artículo 41 de la Carta.

En buena medida, muchos de los derechos y garantías propios del debido proceso quedan integrados para el ámbito de la actuación pública en el derecho a una buena administración [11].

Por un lado, constituye un deber y exigencia a las Administraciones Públicas que debe estar presente en sus actuaciones. De manera que han de actuar con la debida diligencia y en tiempo.

Por otro lado, de este principio se derivan toda una serie de derechos en favor de la ciudadanía (audiencia, resolución en plazo, motivación, tratamiento eficaz y equitativo de los asuntos, buena fe) que deben plasmarse de forma efectiva y diligente.

Posible vulneración. Anexo III. 5. a) del Reglamento

Una posible vulneración del derecho a la buena administración se daría si el uso de sistemas de inteligencia artificial de alto riesgo por parte de las Administraciones Públicas no permite una motivación, cuando resulte obligatorio por la legislación, en lenguaje natural de las decisiones que fundamente y, por ende, el control adecuado de la actividad administrativa basada en estas decisiones. Ello puede darse por falta de transparencia o porque no haya mecanismos de rendición de cuentas adecuados. En EE. UU., la sentencia *K.W. V. Armstrong* 89 F.3D 962, 976 (9TH Cir. 2015) revisó el sistema algorítmico del *Department of Health and Welfare* que otorgaba beneficios de seguridad social a personas con necesidades especiales en el Estado de Idaho. En el procedimiento que llevaba a cabo esta Administración Pública para presupuestar la ayuda pública no se contemplaba una

[6] Más información en: https://www.npr.org/2022/04/19/1093538706/justice-department-works-to-curb-racial-bias-in-deciding-whos-released-from-pris

mínima motivación sobre las razones o causas por las que el sistema había llegado a tal presupuesto, concretamente cuando éste se reducía.[7]

13) Derechos de los menores

De acuerdo con el Considerando 28 del Reglamento, a la hora de valorar la afectación que puede generar un sistema de IA de alto riesgo en los menores, se han de tener en cuenta los derechos de los menores previstos en el artículo 24 de la Carta de Derechos Fundamentales de la UE [11] y la Observación general n.º 25 del Comité de los Derechos del Niño relativa a los derechos de los niños en relación con el entorno digital.

Posible vulneración. Anexo III. 3. a) del Reglamento

Una posible vulneración de los derechos de los menores puede estar presente cuando un sistema de inteligencia artificial se utilice con el objetivo de determinar la admisión a un centro educativo y determinadas variables para predecir un resultado no puedan ingresarse porque los menores no disponen de esos datos. Por ejemplo, en Reino Unido, a causa de la pandemia de coronavirus, se decidió por parte de las autoridades públicas la no realización de los exámenes. Concretamente las calificaciones que marcan el final de la educación escolar obligatoria, la cual se requiere para la mayoría de acceso a los cursos universitarios. Esas calificaciones se otorgaron tomando como referencia una nota ponderada de los profesores de esos estudiantes y las predicciones de un algoritmo. Entre otros problemas, se detectó que algunas de las variables que se debían tener en cuenta para favorecer la precisión del sistema no podían incorporarse al no existir cierta información de diversos estudiantes, por ejemplo, listado de calificaciones anteriores[8].

14) El derecho fundamental a un nivel elevado de protección del medio ambiente. Artículo 37 de la Carta

Los sistemas de IA, y también los de alto riesgo, han de respetar este derecho. Como se ha sostenido en la Carta de Derechos Digitales de España (Artículo XXII), "*El desarrollo de la tecnología y de los entornos digitales deberá perseguir la sostenibilidad medioambiental y el compromiso con las generaciones futuras, y es por ello, que los poderes públicos impulsarán políticas ordenadas a la consecución de tales objetivos con particular atención a la sostenibilidad, durabilidad, reparabilidad y retrocompatibilidad de los dispositivos y sistemas evitando las políticas de sustitución integral y de obsolescencia programada*." De igual modo, se ha de promover "la *eficiencia energética en el entorno digital, favoreciendo la minimización del consumo de energía y la utilización de energías renovables y limpias*." Pues bien, un sistema de evaluación de riesgos del sistema debe tener en cuenta estos elementos por cuanto un desconocimiento muy evidente de los mismos podría llevar a la vulneración de este derecho.

[7] Más información en: https://casetext.com/case/kw-ex-rel-dw-v-armstrong-5?q=

[8] Más información en:

https://blog.container-solutions.com/what-can-we-learn-from-the-ofqual-algorithm-debacle

https://www.engadget.com/uk-algorithm-a-levels-gcse-results-143503870.html

Posible vulneración. Anexo III. 2.a) del Reglamento

Una posible vulneración del derecho a la protección del medio ambiente puede suceder cuando un sistema de IA que se utiliza como componente de seguridad de una red de suministro gas tenga un error crítico que ponga en peligro dicha instalación generando una fuga relevante de estos recursos. La afectación al medio ambiente vendría tanto de la pérdida de ese recurso que no es infinito como de los impactos que en el medio ambiente puede generar la presencia de ese recurso.

8.2 ANEXO B - Componentes más comunes de los sistemas de IA

En este apartado se incorporan algunos de los componentes más comunes de los sistemas de IA categorizados en los siguientes grupos [4]:

1. <u>Actores principales:</u>
    a. **Propietario de datos:** Es el responsable en la organización del dato, es el encargado de su definición, clasificación, protección, uso y calidad.
    b. **Propietario del sistema:** Es el responsable en la organización que solicita el estudio de una solución de IA, es el responsable del sistema de IA.
    c. **Científicos de datos:** Profesionales que aplican la estadística, el aprendizaje automático y los enfoques analíticos para analizar diferentes conjuntos de datos de distintos tamaños y formas y resolver problemas complejos y críticos.
    d. **Ingenieros de datos:** Profesionales que preparan la infraestructura computacional y se centran en el diseño, la gestión y la optimización del flujo de datos.
    e. **Usuarios finales:** Aquellos dentro de una organización que utilizan y se benefician de los resultados proporcionados por el sistema de IA.
    f. **Proveedor de datos:** Terceros que proporcionan datos para su uso en el desarrollo del sistema de IA.
    g. **Proveedor de la nube:** Terceros que ofrecen plataformas informáticas, e incluso en algunos casos tienden a ofrecer algunas capacidades de análisis de datos o "aprendizaje automático como servicio".
    h. **Proveedor del sistema:** Ver definición en glosario global (definición del AI Act).
    i. **Responsables del despliegue:** Ver definición en glosario global (definición del AI Act).
2. <u>Datos:</u>
    a. **Datos sin procesar:** Información recopilada con fines de análisis de IA, posiblemente después de la limpieza, pero antes de que se transforme o analice de alguna manera.
    b. **Datos etiquetados:** Conjunto de elementos de datos escalares o multidimensionales etiquetados con una o más etiquetas informativas, para entrenar un sistema de IA de aprendizaje supervisado.
    c. **Datos públicos:** Información que puede ser utilizada, reutilizada y redistribuida libremente por cualquier persona sin que existan restricciones legales locales, nacionales o internacionales de acceso o uso.

d. **Datos de entrenamiento:** Datos iniciales que se utilizan para desarrollar un sistema de IA, a partir de los cuales el sistema adapta sus parámetros internos para refinar sus reglas.
e. **Datos aumentados:** Conjunto de datos (normalmente etiquetados) que se ha aumentado añadiendo datos producidos por transformaciones o por sistemas generativos. En el reconocimiento de imágenes, las técnicas de aumento de datos incluyen el recorte, el relleno y el volteo horizontal.
f. **Datos de prueba:** Conjunto de datos utilizado para proporcionar una evaluación no sesgada de un sistema de IA ajustado al conjunto de datos de entrenamiento. Los datos de prueba se utilizan para probar el sistema.
g. **Datos de validación:** Conjuntos de datos etiquetados, que difieren de los datos etiquetados ordinarios sólo en su uso y, normalmente, en sus circunstancias de recogida. Para evaluar un sistema de IA en el entrenamiento se usan datos de validación.
h. **Datos de evaluación:** Los datos de evaluación se utilizan para evaluar la calidad predictiva del sistema entrenado. El sistema de IA evalúa el rendimiento predictivo comparando las predicciones en el conjunto de datos de evaluación con los valores reales (conocidos como ground truth) utilizando una serie de métricas.
i. **Datos preprocesados:** Los datos preprocesados antes de introducirlos en el sistema de IA.
j. **Datos para métricas:** El tipo de números que recogemos cuando medimos algo. Los datos métricos pueden ser de escala de proporción, de escala de intervalo, de escala de números enteros y de números cardinales.
k. **Parámetros del sistema de IA:** Un parámetro del sistema de IA es una variable de configuración que es interna del sistema de IA y cuyo valor puede estimarse a partir de los datos dados.
l. **Parámetros de entrenamiento:** Los parámetros de entrenamiento del sistema de IA son cantidades ajustadas por el proceso de aprendizaje mediante la aplicación de algoritmos de entrenamiento basados en los datos de entrenamiento. Los valores de los parámetros de entrenamiento determinan la función real de clasificación, predicción o detección calculada por el sistema de IA.
m. **Hiperparámetros:** Los hiperparámetros definen conceptos de alto nivel sobre los sistemas de IA, como la frecuencia de ajuste de los parámetros internos por parte del algoritmo de entrenamiento. No pueden aprenderse a partir de los datos de entrada, sino que deben establecerse por ensayo y error utilizando técnicas de búsqueda en el espacio del sistema de IA.

3. <u>Entornos y herramientas:</u>
   a. **Listas de control de acceso:** Una lista de control de acceso (LCA) es una tabla que representa qué derechos de acceso tiene cada usuario a un recurso concreto, como un directorio de archivos o un archivo individual.
   b. **Nube:** Es la disponibilidad bajo demanda de los recursos del sistema informático, especialmente el almacenamiento de datos (almacenamiento en la nube) y la potencia de cálculo, sin una gestión activa directa por parte del usuario.

c. **Redes de comunicación:** Redes con conectividad a Internet para fines de comunicación.
d. **Librerías:** Programas prescritos que implementan sistemas listos para ser utilizados, para: cálculo científico, datos tabulares, análisis de series temporales, modelado y preprocesamiento de datos, aprendizaje profundo, entre otros.
e. **Plataformas de ingesta de datos:** Es la plataforma donde se realiza la ingesta de datos.
f. **Sistema de Archivos Distribuidos:** La distribución del sistema de archivos es un método para almacenar y acceder a los archivos, que permite que varios usuarios accedan y compartan archivos desde varias máquinas, o varios hosts, a través de una red informática.
g. **Protocolos de comunicación:** El protocolo de comunicación es un sistema de reglas que permite a dos o más entidades de un sistema de comunicaciones transmitir información a través de cualquier tipo de variación de una cantidad física.
h. **Sistema gestor de base de datos (SGBD):** Es el software que gestiona el almacenamiento, la recuperación y la actualización de datos en un sistema informático.
i. **Herramientas de exploración de datos**: Las herramientas utilizadas para la exploración de datos. Herramientas como las de visualización y los gráficos se utilizan con frecuencia para crear una visión más directa de los conjuntos de datos que el simple examen de miles de números o nombres individuales.
j. **Herramientas de monitoreo:** Las herramientas que se usan para seguir el estado del sistema en uso, para que antes se avisen y mejoren fallos, defectos o problemas.
k. **Sistema operativo:** Gestiona el hardware de los ordenadores, los recursos de software y proporciona servicios comunes para los programas informáticos.
l. **Técnicas de optimización:** Técnicas utilizadas para la optimización en el ajuste de sistemas, como la búsqueda en cuadrícula, la búsqueda aleatoria y la optimización bayesiana.
m. **Plataformas de aprendizaje automático:** Proporciona un ecosistema de herramientas, bibliotecas y recursos que apoyan el desarrollo de aplicaciones de aprendizaje automático.
n. **Procesadores:** Un procesador es la parte de un ordenador que interpreta las órdenes y realiza los procesos que el usuario ha solicitado.
o. **Herramientas de retención y borrado de datos:** Herramientas de retención que proporcionan ayuda para la implementación de la retención, archivado, bloqueo, anonimización y borrado de los datos de manera segura según los periodos definidos.
p. **Herramientas de retención y borrado de sistemas:** Herramientas de retención que proporcionan ayuda para la implementación de la retención, archivado, bloqueo, anonimización y borrado de los sistemas de manera segura según los periodos definidos.

4. <u>Procesos:</u>
    a. **Transferencia de aprendizaje:** Capacidad de reutilizar los conocimientos previamente aprendidos para resolver nuevos problemas con mayor rapidez.
    b. **Preprocesamiento de datos:** Comprensión, preparación y limpieza de los datos.
    c. **Almacenamiento de datos:** Los datos pueden almacenarse localmente, en un sistema de archivos distribuido o en la nube.
    d. **Comprensión de datos:** Conocimiento que se tiene sobre los datos, los activos de datos, las necesidades que los datos van a satisfacer, su contenido y su ubicación.
    e. **Selección de características:** Durante este proceso se reduce el número de dimensiones o características del vector de entrada, identificando aquellas que son más significativas para el sistema de IA.
    f. **Ingesta de datos:** Proceso relacionado con el transporte de datos desde múltiples fuentes para componer puntos de datos multidimensionales. Los datos pueden colocarse en un medio de almacenamiento en el que se pueda acceder a ellos, utilizarlos y analizarlos, o bien, el flujo de datos puede utilizarse directamente por el sistema de IA.
    g. **Etiquetado de datos:** Es el proceso de detección y etiquetado de las muestras de datos. El proceso puede ser manual y lento y estar asistido por software.
    h. **Aumento de datos:** Técnicas utilizadas para aumentar la cantidad de datos añadiendo copias ligeramente modificadas de datos ya existentes o datos sintéticos recién creados a partir de datos existentes. Ayuda a reducir el sobreajuste cuando se entrena un aprendizaje automático.
    i. **Ajuste del sistema:** El ajuste se centra en el establecimiento de parámetros especiales, a menudo llamados hiperparámetros. Este proceso puede realizarse de forma manual o automática buscando en el espacio de los parámetros del sistema, mediante la llamada optimización de hiperparámetros.
    j. **Técnica de discretización:** Es el proceso de convertir un atributo numérico en un atributo simbólico mediante la partición del dominio del atributo.
    k. **Mantenimiento del sistema:** Tras la implantación, es necesario supervisar la precisión de la predicción para detectar posibles cambios o desviaciones de los conceptos. Un descenso en el rendimiento del sistema podría superarse volviendo a entrenarlo con datos recientes y luego volver a desplegarlo en producción.
    l. **Retención y borrado de datos:** Proceso de definición e implementación de los periodos de retención y borrado de los datos en función de su tipología.
    m. **Retención y borrado de sistemas:** Proceso de definición e implementación de los periodos de retención y borrado de los sistemas en función de su tipología.

8.3 ANEXO C - Tipos de riesgos comunes en el ámbito de la IA

A continuación, se detallan algunas de los tipos de riesgos más comunes en el ámbito de la IA [2] [3]:

1. **Falta de transparencia:**

La transparencia consiste en comunicar las actividades y decisiones de una organización (políticas, procedimientos, etc.) e información adecuada sobre un sistema de IA (capacidades, rendimiento, limitaciones, opciones de diseño, algoritmos, datos de entrenamiento, etc.) a las partes interesadas. Si las organizaciones no pueden proporcionar la información adecuada a las partes interesadas, tendrá un impacto negativo en la confiabilidad y la responsabilidad de la organización y el sistema de IA. Esto podría derivar, por ejemplo, en riesgos relacionados con una identificación deficiente e ineficaz de responsabilidades. Por ejemplo, en el sistema de promoción de empleados, si no somos capaces de establecer y comunicar adecuadamente las limitaciones del sistema, podría derivar en una confianza excesiva por las partes involucradas en el proceso de toma de decisiones y esto en potenciales decisiones de promoción erróneas, perjudicando a otros empleados y perdiendo la confianza en el sistema.

2. **Falta de explicabilidad:**

La explicabilidad es la propiedad de un sistema de IA de que los factores que influyen en una decisión se pueden expresar de una manera que los humanos puedan entender. Si no se pueden explicar estos factores, la validación del sistema de IA y la confianza en el sistema se ven afectados negativamente, ya que no está claro por qué el sistema ha tomado una decisión y si tomará la decisión correcta en todos los casos. Esta incertidumbre puede dar lugar a muchos riesgos y tener un fuerte impacto en objetivos generales como la confiabilidad y la rendición de cuentas y objetivos específicos como la seguridad, la protección, la equidad y la robustez. Por ejemplo, en el sistema de promoción de empleados, si ante una decisión del sistema de promocionar a un empleado frente a otro el empleado perjudicado solicita una revisión del proceso de decisión, debemos ser capaces de dar una explicación de cómo el sistema ha tomado la decisión, por ejemplo, identificando aquellas variables del sistema que han sido las más determinantes en el proceso.

3. **Nivel de automatización:**

Los sistemas de IA pueden operar con diferentes niveles de automatización. Este nivel de automatización puede ser muy bajo, en el caso donde un operador controla el sistema, o muy alto, como los sistemas de acción autónoma. Dependiendo del caso de uso específico, las decisiones automatizadas de dichos sistemas pueden tener un impacto en diversas áreas de preocupación, como la seguridad o la equidad. Por ejemplo, en el sistema de IA de promoción de empleados, si no existe un responsable de revisar las decisiones que propone el sistema y se ejecutan las promociones de forma directa, existe el riesgo de que en caso de haber algún error en los datos de entrada se produzca una promoción errónea, perjudicando a otros empleados y perdiendo la confianza en el sistema.

4. **Fuentes de riesgo relacionadas con el aprendizaje automático:**

El comportamiento de los sistemas de IA depende, no solo de los algoritmos en uso, sino también de los datos con los que se entrenan los sistemas. Existen diversos riesgos derivados del uso de los datos. Por ejemplo:

- La calidad inadecuada de los datos podría afectar a varios objetivos como la equidad, la seguridad y la solidez.
- Los datos pueden dejar de ser representativos del dominio de aplicación, lo que conlleva riesgos para los objetivos del negocio.
- El recabado y almacenamiento de datos puede incurrir en riesgos éticos y legales significativos.

En el ejemplo del sistema de promoción de empleados, un riesgo legal (y además ético) sería, como hemos analizado a lo largo de la guía, una promoción discriminatoria de unos empleados frente a otros incumpliendo así con el derecho fundamental a la no discriminación.

5. **Problemas de hardware del sistema:**

Las fuentes de riesgo relacionadas con problemas de hardware incluyen, por ejemplo, errores basados en componentes defectuosos (cortocircuitos, interrupciones, líneas de bus defectuosas, etc.). El desarrollo de sistemas de IA podría verse limitado debido a las diferentes capacidades de hardware de los sistemas en términos de potencia de procesamiento, memoria y la disponibilidad de aceleradores de hardware de IA dedicados.

En un ejemplo como el de la promoción de los empleados, quizá una limitación de hardware y un retraso en el proceso de decisión del sistema por la indisponibilidad derivada, no supone un riesgo alto, pero en cambio, en un sistema de IA encargado de administrar de forma automática insulina a un paciente diabético, un fallo en el hardware podría suponer un riesgo crítico para la vida del paciente.

6. **Problemas del ciclo de vida del sistema:**

Los métodos, procesos y también el uso inadecuado o insuficiente de un sistema de IA a lo largo de su ciclo de vida pueden generar riesgos. Por ejemplo, un proceso de diseño defectuoso puede no anticipar los contextos en los que se utilizará el sistema de IA, lo que hace que falle inesperadamente cuando se usa en estos contextos.

En el ejemplo del sistema de administración automática de insulina, deberemos analizar y anticipar los contextos en los que se utilizará el sistema a lo largo de su ciclo de vida. Definir los ciclos de actualización y revisión a los que deberá someterse para mitigar posibles fallos en su funcionamiento.

7. **Preparación tecnológica:**

La preparación tecnológica indica cuán madura es una tecnología dada en un contexto de aplicación determinado.

Las tecnologías menos maduras utilizadas en el desarrollo y la aplicación de sistemas de IA pueden añadir riesgos desconocidos para la organización o difíciles de evaluar.

Para las tecnologías maduras, se puede disponer de una mayor variedad de datos de experiencia, lo que facilita la identificación y la evaluación de los riesgos. Por ejemplo, en

el sistema de promoción de empleados un riesgo asociado a la preparación tecnológica podría ser el relacionado con la falta de experiencia en el funcionamiento de estos sistemas que podría derivar en retrasos y aumento de los costes si en un momento dado se requiere actualizar o modificar los parámetros del sistema y no se dispone del conocimiento necesario.

8. **Complejidad del entorno:**

La complejidad del entorno de un sistema de IA determina la gama de situaciones que un sistema de IA puede soportar en su contexto operativo. Uno de los riesgos más relevantes, por ejemplo, es el relacionado con el grado de comprensión del entorno del sistema de IA.

Una comprensión parcial del entorno dará lugar a un nivel de incertidumbre que es una fuente de riesgo especialmente relevante en la fase del diseño de los sistemas de IA.

En el ejemplo del sistema de promoción de empleados, deberemos analizar la complejidad del entorno y las diferentes situaciones en las que se utilizará el sistema para identificar adecuadamente los posibles riesgos derivados y mitigar en la mayor medida posibles fallos en su funcionamiento.

9. **Otras potenciales fuentes de riesgos:**
    a. Dificultad en la identificación de responsabilidades y rendición de cuentas.
    b. Uso inadecuado, incorrecto o fraudulento del sistema de IA.
    c. Exceso de confianza en las decisiones del sistema de IA (automatización de las decisiones sin ningún tipo de supervisión).
    d. Amenazas de seguridad y ciberseguridad (se recomienda especialmente en este punto consultar la guía de ciberseguridad que incorpora un inventario adicional de riesgos y amenazas de los sistemas de IA en el contexto de la ciberseguridad).
    e. Posibles amenazas a la privacidad de las personas si no se desarrolla un adecuado gobierno de los datos (se recomienda consultar la guía de gobierno de datos).
    f. Posibles perturbaciones o manipulaciones en los datos no deseadas (se recomienda consultar la guía de gobierno de datos y la guía de ciberseguridad).
    g. Otros usos inadecuados derivados de una especificación deficiente del sistema de IA.
    h. Los posibles sesgos en los datos también pueden suponer una amenaza para el adecuado uso del sistema de IA (se recomienda consultar la guía de gobierno de datos).

8.4 ANEXO D - Ejemplos de controles en el ámbito de la IA

Nota importante: *Este anexo, tal y como se indica en la sección 4.5 de la guía, pretende incorporar un listado de algunos ejemplos de posibles controles para que sirva como orientación y referencia al lector cuando se enfrente al proceso de definición de medidas de control [2]. En este sentido, no se espera que el lector implemente todas las medidas de control listadas o que se limite únicamente a éstas, este proceso dependerá del contexto y de los riesgos identificados, analizados y evaluados a lo largo del desarrollo del sistema de gestión de riesgos.*

1. **Medidas de gobierno:**
    a. Establecer y definir las normas profesionales y éticas en el ámbito de las tecnologías de la información para desarrollar sistemas de IA siguiendo los estándares de referencia.
    b. Facilitar documentación para los responsables del despliegue que incluye el contexto apropiado y las limitaciones conocidas de los sistemas de IA.
    c. Definir mecanismos de reparación si las personas se ven afectadas negativamente por las decisiones del sistema de IA.
2. **Medidas de inclusión:**
    a. Incorporar en todo el ciclo de vida de los sistemas de IA la contribución de expertos en tecnologías de la información, con conocimientos técnicos que incluyan un conocimiento profundo de las tecnologías de inteligencia artificial, datos y computación de datos.
    b. Incluir a los usuarios y partes interesadas, en la medida de lo posible, en el proceso de desarrollo del sistema de IA (revisión de las especificaciones, participación en las pruebas, etc.).
    c. Recoger, analizar e incorporar las opiniones y el feedback de los usuarios y las partes interesadas (por ejemplo, a través de encuestas).
    d. Consolidar, en la medida de lo posible, equipos de desarrollo y mantenimiento con diversidad de opiniones, orígenes y pensamientos.
    e. Evaluar el sesgo de la interacción tras la retroalimentación recogida de todas las partes interesadas.
3. **Transparencia y explicabilidad:**
    a. Implementar técnicas de aproximación de los modelos incorporados en el sistema de IA (como la técnica de explicación diagnóstica del modelo local interpretable).
    b. Implementar métodos de diagnóstico de modelos (como el análisis de componentes principales).
    c. Implementar métodos de imitación de modelos o destilación (por ejemplo, transferencia de conocimientos de redes neuronales a árboles de decisión).
4. **Capacidad de control:**
    a. Establecer puntos de control en el ciclo de vida de desarrollo del sistema de IA (normalmente cuantos más puntos de control mejor, pero este es un factor por evaluar en cada contexto y escenario y depende de factores como nivel de riesgo a gestionar, recursos disponibles, necesidades a cubrir, etc.) y un mecanismo de transferencia del conocimiento.
    b. Instaurar mecanismos de para que los usuarios informen de cualquier tipo de incidencia del sistema de IA.

c. Implementar mecanismos para que los usuarios, en la medida de lo posible, ajusten el sistema decisión del sistema de IA.

5. **Controles de seguridad y ciberseguridad:**
   a. Implementar mecanismos para la detección de posibles ataques de *fuzzing* (técnica de testeo automatizado mediante la que se introducen datos inválidos, aleatorios o inesperados a un sistema informático) o manipulación de entrada a voz, video y gestos.
   b. Desarrollar mecanismos para la oportuna identificación de datos de entrenamiento maliciosos (por ejemplo, sistemas de detección de cambios no autorizados, sistemas de detección de intrusiones o sistemas de monitorización del comportamiento del sistema de IA).
   c. Identificar a los usuarios que actúan de forma anómala (de manera coordinada y diferente a la ordinaria), pues éstos pueden ser una potencial fuente de ataque malicioso.
   d. Implantar instalaciones de auditoría o rastreo de eventos para examinar los estados de decisión, entrenamiento o detección de los sistemas de IA.
6. **Controles de privacidad de los datos**
   a. Implementar mecanismos para garantizar el adecuado uso y tratamiento de datos sensibles.
   b. Facilitar formación a los usuarios que trabajen con datos sensibles para garantizar que lo hacen de forma adecuada, con discreción y precaución para evitar, por ejemplo, revelar información crítica del sistema de IA a conocidos o familiares.
   c. Obtener el consentimiento de individuos o grupos de individuos para el tratamiento de sus datos en el sistema de IA.
   d. Capacitar a los sistemas de IA para identificar sesgos o desviaciones indeseadas.
7. **Controles y medidas sobre los datos**
   a. Seleccionar adecuadamente los conjuntos de datos de entrenamiento y prueba del sistema de IA para que sean coherentes con el entorno que pretenden representar.
   b. Identificar los atributos confidenciales que se requieren para el comportamiento adecuado del sistema de IA.
   c. Implementar medidas de control y validación de conjuntos de datos (por ejemplo, métodos de retención de datos, K-fold cross validation, leave-one-out of cross validation (LOO), remuestreo jackknife, muestreo estratificado, etc.).
8. **Rendimiento del diseño y la implementación del modelo**
   a. Desarrollar una especificación detallada del diseño, funcionalidades e implementación del sistema de IA.
   b. Detallar las técnicas de "feature engineering" utilizadas, como la selección, extracción y/o regulación de las características (features).
   c. Especificar el modelo, algoritmos, hiperparámetros y la topología por escenario utilizados.
   d. Desarrollar una API para verificar las métricas de rendimiento del sistema de IA.

e. Desarrollar un análisis contrafactual para comprender mejor el comportamiento del sistema y evitar resultados no deseados.

9. **Robustez, fiabilidad y resiliencia**
   a. Implementar técnicas de entrenamiento adversario, como inyectar ejemplos contradictorios (generados por diferentes estrategias de ataque) en los datos de entrenamiento).
   b. Implementar mecanismos para reconstruir entradas dañadas, es decir, eliminar ruido (por ejemplo, denoising autoencoder).
   c. Agregar un sistema crítico para detectar cuándo el sistema de IA tiene demasiada confianza a pesar de una entrada ruidosa.
   d. Desarrollar pruebas en entornos simulados y/o pruebas de campo y hacerlo periódicamente incluso después del despliegue del sistema de IA.

8.5 ANEXO E – Ejemplos de indicadores de efectividad

8.5.1 ANEXO E.I – En relación con las medidas de gestión de riesgos

Medidas a evaluar	Indicadores de efectividad
Análisis y definición del contexto interno y externo.	• Número de factores evaluados y considerados en el desarrollo del sistema de gestión de riesgos
Definición y actualización del apetito al riesgo.	• Elementos considerados en la definición • Frecuencia de actualización definida
Inventariado de componentes del sistema de IA.	• Número de componentes del sistema de IA identificados e inventariados
Identificación de las fuentes de riesgo principales.	• Número de fuentes de riesgo identificadas
Evaluación de las fuentes de riesgo.	• Número de fuentes de riesgos evaluadas
Identificación y análisis del impacto de los efectos y su probabilidad de ocurrencia.	• % de riesgos categorizados en función de su impacto y probabilidad de ocurrencia
Reporting de los resultados.	• Nivel de madurez de la metodología de reporting de resultados definida
Análisis de datos recogidos poscomercialización.	• Metodología implementada de recabado de información poscomercialización • Frecuencia definida de análisis de la información recabada
Identificación de nuevos posibles riesgos.	• Número de riesgos identificados en el proceso de poscomercialización • % de estos riesgos incorporado al sistema de gestión de riesgos (con los correspondientes análisis, evaluación e implementación de controles)

Definición y selección de las opciones para el tratamiento del riesgo.	• Número de opciones de tratamiento valoradas y evaluadas • Distribución de las medidas de tratamiento aplicadas en función del tipo de opción
Planificar e implementar el tratamiento del riesgo.	• Nivel de madurez del plan de tratamiento definido • Número de partes interesadas consultadas para su elaboración
Evaluar la eficacia de cada tratamiento.	• % de riesgos que quedan como aceptables (riesgo por debajo del apetito definido) tras la aplicación de los tratamientos
Determinar si el riesgo residual es aceptable.	• % de riesgos que han necesitado medidas adicionales • Número de medidas adicionales que han requerido cada uno de estos riesgos
Documentar e informar los riesgos residuales.	• % de riesgos que quedan como residuales • % de riesgos residuales que han sido adecuadamente documentados e informados

8.5.2 ANEXO E.II – Con relación a los controles del Anexo D

1. **Medidas de gobierno**

Control / medida técnica	Indicador efectividad
Establecer y definir las normas profesionales y éticas para el uso de la IA siguiendo los estándares de referencia.	• Número de normas o estándares profesionales establecidos y aplicados
Facilitar documentación para los usuarios que incluye el contexto apropiado y las limitaciones conocidas de los sistemas de IA.	• % de usuarios que han leído y aplicado la documentación
Definir mecanismos de reparación si las personas se ven afectadas negativamente por las decisiones del sistema de IA.	• Número de mecanismos definidos y establecidos

2. Medidas de inclusión

Control / medida técnica	Indicador efectividad
Incorporar la contribución de diferentes expertos en el campo de la IA para el desarrollo de los sistemas de IA y a lo largo de todo el ciclo de vida.	• Número de expertos involucrados
Incluir a los responsables del despliegue y partes interesadas, en la medida de lo posible, en el proceso de desarrollo del sistema de IA (revisión de las especificaciones, participación en las pruebas, etc.).	• Número de responsables del despliegue e interesados involucrados en las pruebas
Recoger, analizar e incorporar las opiniones y el *feedback* de los usuarios y las partes interesadas (por ejemplo, a través de encuestas).	• Número de opiniones de los interesados recabadas
Consolidar, en la medida de lo posible, equipos de desarrollo y mantenimiento con diversidad de opiniones, orígenes y pensamientos	• Diversidad de opiniones incorporadas en la consolidación de los equipos
Evaluar el sesgo de la interacción tras la retroalimentación recogida de todas las partes interesadas.	• Número de elementos de sesgo identificados

3. Transparencia y explicabilidad

Control / medida técnica	Indicador efectividad
Implementar técnicas de aproximación de los modelos incorporados en el sistema de IA (como la técnica de explicación diagnóstica del modelo local interpretable).	• Precisión del modelo de explicación implementado • Número de técnicas diferentes implementadas
Implementar métodos de diagnóstico de modelos (como el análisis de componentes principales).	• Número de métodos de diagnóstico implementados
Implementar métodos de imitación de modelos o destilación (por ejemplo, transferencia de conocimientos de redes neuronales a árboles de decisión).	• Precisión del modelo de imitación implementado • Número de técnicas diferentes implementadas

4. Capacidad de control

Control / medida técnica	Indicador efectividad
Establecer puntos de control en el ciclo de vida de desarrollo del sistema de IA y un mecanismo de transferencia del conocimiento.	• Número de puntos de control implementados.
Instaurar mecanismos de comunicación (por ejemplo, habilitando una dirección de correo electrónico) para que los responsables del despliegue informen de cualquier tipo de incidencia del sistema de IA.	• Número de reacciones de los responsables del despliegue informadas
Implementar mecanismos para que los responsables del despliegue, en la medida de lo posible, ajusten el sistema decisión del sistema de IA.	• Número de ajustes abordados por los responsables del despliegue

5. Controles de seguridad y ciberseguridad

Control / medida técnica	Indicador efectividad
Implementar mecanismos para la detección de posibles ataques de fuzzing (técnica de testeo automatizado mediante la que se introducen datos inválidos, aleatorios o inesperados a un sistema informático) o manipulación de entrada a voz, video y gestos.	• Número de ataques de fuzzing detectados en las diferentes entradas.
Desarrollar mecanismos para la oportuna identificación de datos de entrenamiento maliciosos (por ejemplo, sistemas de detección de cambios no autorizados, sistemas de detección de intrusiones o sistemas de monitorización del comportamiento del sistema de IA).	• Número de detecciones de datos de entrenamiento maliciosos.
Identificar a los usuarios que actúan de forma anómala (de manera coordinada y diferente a la ordinaria), pues éstos pueden ser una potencial fuente de ataque malicioso.	• Número de usuarios maliciosos detectados
Implantar instalaciones de auditoría o rastreo de eventos para examinar los estados de decisión, entrenamiento o detección de los sistemas de IA.	• Número de instalaciones de auditoría o rastreo de eventos implantados

6. Controles de privacidad de los datos

Control / medida técnica	Indicador efectividad
Implementar mecanismos para garantizar el adecuado uso y tratamiento de datos sensibles.	• Número de atributos sensibles identificados y medidas implementadas.
Facilitar formación a los usuarios que trabajen con datos sensibles para garantizar que lo hacen de forma adecuada, con discreción y precaución para evitar, por ejemplo, revelar información crítica del sistema de IA a conocidos o familiares.	• % de trabajadores conocedores de los requisitos de discreción y precaución establecidos
Obtener el consentimiento de individuos o grupos de individuos para el tratamiento de sus datos en el sistema de IA.	• % de tratamientos de datos con consentimiento aprobado
Capacitar a los sistemas de IA para identificar sesgos o desviaciones indeseadas.	• Número de fuentes de sesgo identificadas

7. Controles y medidas sobre los datos

Control / medida técnica	Indicador efectividad
Seleccionar adecuadamente los conjuntos de datos de entrenamiento y prueba del sistema de IA para que sean coherentes con el entorno que pretenden representar.	• Observación cualitativa de que los datos representan el entorno y realidad que pretender reproducir
Identificar los atributos confidenciales que se requieren para el comportamiento adecuado del sistema de IA.	• Número de atributos confidenciales identificados
Implementar medidas de control y validación de conjuntos de datos (por ejemplo, métodos de retención de datos, K-fold cross validation, leave-one-out of cross validation (LOO), remuestreo jackknife, muestreo estratificado, etc.).	• Observación cualitativa de cada una de las medidas de validación de datos

8. Rendimiento del diseño y la implementación del modelo

Control / medida técnica	Indicador efectividad
Desarrollar una especificación detallada del diseño, funcionalidades e implementación del sistema de IA.	• Grado de comprensión y conocimiento del diseño, funcionalidades y

	funcionamiento por parte de los responsables del despliegue.
Detallar las técnicas de "feature engineering" utilizadas, como la selección, extracción y/o regulación de las características (features).	• Número de técnicas implementadas documentadas
Especificar el modelo, algoritmos, hiperparámetros y la topología por escenario utilizados.	• Nivel de madurez de la documentación del modelo, algoritmos, hiperparámetros y la topología por escenario utilizados.
Desarrollar una API para verificar las métricas de rendimiento del sistema de IA.	• % de responsables del despliegue que conoce y utiliza las herramientas para verificar el rendimiento
Desarrollar un análisis contrafactual para comprender mejor el comportamiento del sistema y evitar resultados no deseados.	• Número de errores futuros predichos y corregidos

9. Robustez, fiabilidad y resiliencia

Control / medida técnica	Indicador efectividad
Implementar técnicas de entrenamiento adversario, como inyectar ejemplos contradictorios (generados por diferentes estrategias de ataque) en los datos de entrenamiento).	• Número de técnicas implementadas • Análisis cualitativo de la suficiencia de las medidas implementadas para garantizar la robustez, fiabilidad y resiliencia del modelo
Implementar mecanismos para reconstruir entradas dañadas, es decir, eliminar ruido (por ejemplo, denoising autoencoder).	• Número de entradas dañadas reconstruidas
Agregar un sistema crítico para detectar cuándo el sistema de IA tiene demasiada confianza a pesar de una entrada ruidosa.	• Número de detecciones de exceso de confianza del sistema identificadas
Desarrollar pruebas en entornos simulados y/o pruebas de campo y hacerlo periódicamente incluso después del despliegue del sistema de IA.	• Número de pruebas implementadas

8.6 ANEXO F – Glosario de términos

Esta guía se ha desarrollado con un enfoque que trata de explicar cada concepto presente en la guía cuando se expone, no obstante, se han recogido ciertos términos específicos en esta sección como aclaración adicional:

1. **Amenaza:** peligros a los que está expuesto el sistema que pueden terminar materializándose en un riesgo. Las amenazas de un sistema provienen principalmente de ataques externos (como un ciberataque), de no cumplir las políticas de seguridad (conectar dispositivos no autorizados a la red o utilizar contraseñas débiles) y de sucesos inesperados (como incendios o robos físicos, por ejemplo).
2. **Vulnerabilidad:** debilidad propia de un sistema que permite ser atacado y recibir un daño. Las vulnerabilidades se producen de forma habitual por una baja protección contra ataques externos.
3. **Riesgo:** posibilidad de que un sistema sufra un incidente y que una amenaza se materialice causando daños. El riesgo es, por lo tanto, la probabilidad de que la amenaza se materialice aprovechando una vulnerabilidad existente.
4. **Medidas de control:** en el contexto de la gestión de riesgos, son las medidas que deben tomarse para proteger el sistema de las amenazas, haciéndolo menos vulnerable y reduciendo la probabilidad de que el riesgo se materialice o en su lugar reduciendo el impacto que éste tendría en mi sistema.
5. **Apetito al riesgo:** es el volumen de riesgo que nuestra organización está dispuesta a aceptar en la búsqueda de lograr su misión.
6. **Riesgo inherente:** es el riesgo intrínseco de cada actividad, sin tener en cuenta las medidas de control que puedan implantarse.
7. **Riesgo residual:** es aquel riesgo que subsiste, después de haber implementado controles.
8. **Ataque de fuzzing:** técnica de testeo automatizado mediante la que se introducen datos inválidos, aleatorios o inesperados a un sistema informático.
9. **HRAIS:** los sistemas de alto riesgo basados en Inteligencia Artificial son aquellos con un impacto significativo en la vida de las personas y en sus derechos fundamentales.
   Estos sistemas se utilizan en áreas críticas como la biometría, la educación, el empleo, la aplicación de la ley, la gestión de infraestructuras críticas, y otros sectores donde su mal uso podría causar daños considerables.
   El reglamento en el artículo 6 y el Anexo III establecen los requisitos para el desarrollo, implementación y supervisión para garantizar su seguridad y fiabilidad.

8.7 ANEXO G – Política de Gestión de riesgos de IA

Este anexo, tal y como se indica en la sección 4.9, tiene como finalidad presentar una propuesta de Política de Gestión de Riesgos en Inteligencia Artificial (IA) que sirva como referencia práctica para las organizaciones en el proceso de definición y desarrollo de su propio marco de gestión de riesgos. En este sentido, el contenido aquí descrito busca ofrecer una orientación general sobre los elementos y principios que debería contemplar una política de esta naturaleza.

No se espera que esta política sea adoptada de forma literal o completa por el lector, sino que actúe como un ejemplo de estructura y contenido que pueda adaptarse al contexto, naturaleza y nivel de madurez de cada organización. Su aplicación deberá ajustarse a los riesgos específicos identificados, analizados y evaluados a lo largo del proceso de gestión de riesgos en sistemas de IA, de acuerdo con las necesidades y particularidades de cada caso.

1. Resumen

Esta política establece un marco integral para la gestión de riesgos y gobernanza de los sistemas de Inteligencia Artificial (IA) en la organización, alineado con el Reglamento Europeo de IA (UE 2024/1689) y las recomendaciones de la ISO/IEC 23894. Su propósito es proporcionar a la organización una guía clara para manejar de forma ética, segura y responsable los sistemas de IA, especialmente aquellos considerados de alto riesgo. La política define los objetivos y metas que orientan la identificación, evaluación, mitigación y seguimiento de riesgos, asegurando que las decisiones sobre IA se tomen de manera transparente y consistente.

Se establece la audiencia a la que aplica, incluyendo todo el personal que interviene en cualquier fase del ciclo de vida de los sistemas de IA, así como proveedores y colaboradores externos que puedan influir en su desempeño o cumplimiento normativo. Se incluyen también los conceptos clave, ofreciendo un lenguaje común sobre sistemas de alto riesgo, riesgos operativos, éticos y legales, cumplimiento normativo y ciclo de vida de los sistemas.

La política asigna roles y responsabilidades específicas a cada área, asegurando que todos los actores conozcan su papel en la gestión de riesgos y supervisión de los sistemas de IA. Se define la estrategia de riesgo, que abarca la evaluación de riesgos internos y de terceros, la identificación de riesgos residuales y la supervisión continua, consolidando un marco de gobernanza robusto, coherente y alineado con la normativa europea.

2. Política de uso de IA

La política de uso de IA complementa la gestión de riesgos, proporcionando directrices prácticas para el uso responsable de los sistemas de Inteligencia Artificial en la organización. Establece programas de entrenamiento y concienciación para todo el personal involucrado con sistemas de IA, asegurando que comprendan los riesgos operativos, legales y éticos, y fomentando un uso informado y responsable de la tecnología.

Se definen guías internas de uso, que indican los usos aceptables de la IA y las restricciones necesarias para proteger la seguridad, la privacidad y los derechos fundamentales. Además, se incluyen procesos de preaprobación, de manera que cualquier nuevo sistema o modificación significativa pase por una revisión formal de riesgos antes de su implementación. Por último, la política establece un inventario de sistemas de IA, manteniendo un registro actualizado de todos los sistemas utilizados, sus responsables, nivel de riesgo y estado de cumplimiento, garantizando trazabilidad, control y supervisión continua en toda la organización.

9. Referencias, estándares y normas

Para el desarrollo de esta guía se han consultado y utilizado especialmente las normas y estándares siguientes:

[1] ISO 31000:2018 - Risk management – Guidelines
[2] ISO/IEC 23894 - Information technology – Artificial intelligence – Guidance on risk management
[3] ISO/IEC 42001 Information technology – Artificial intelligence – Management system
[4] ENISA Report – AI Cybersecurity Challenges - Threat Landscape
[5] ISO/IEC 27001:2022 - Information security, cybersecurity, and privacy protection – Information security management systems – Requirements
[6] ISO/IEC 22989:2022 - Information technology – Artificial intelligence – Artificial intelligence concepts and terminology
[7] ISO/IEC 23053:2022 - Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)
[8] ISO/IEC 5259-1 - Artificial intelligence – Data quality for analytics and machine learning (ML) – Part 1: Overview, terminology, and examples
[9] NIST - AI Risk Management Framework
[10] ENISA Report - Securing Machine Learning Algorithms
[11] Carta de Derechos Fundamentales de la Unión Europea (2000)
[12] prEN 18228 AI Risk Managment

La presente guía toma como referencia el Reglamento 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024 (Reglamento Europeo de Inteligencia Artificial)

Financiado por
la Unión Europea
NextGenerationEU

GOBIERNO DE ESPAÑA
MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA
SECRETARÍA DE ESTADO DE DIGITALIZACIÓN E INTELIGENCIA ARTIFICIAL

Plan de
Recuperación,
Transformación
y Resiliencia

España | digital 2026