La gobernanza de datos y de IA como herramientas para gestionar los riesgos derivados de los sesgos

Data Governance and AI Governance as Tools to Manage the Risks Arising from Biases

Eduard Chaveli Donet
Especialista en Gobernanza, riesgos y cumplimiento en protección de datos e inteligencia artificial

Resumen:
En este artículo se aborda el concepto de sesgos, se diferencia de otros afines con los que tiene relación y a veces se confunden y se describen las etapas del ciclo de vida de los sistemas de IA identificando cómo pueden penetrar los sesgos en cada una de ellas para, a partir de ahí, poder gestionar los riesgos derivados de los mismos.

Tras una visión general de las obligaciones que para los diferentes sujetos dispone el RIA y que pueden permitir la gestión de los riesgos para los derechos fundamentales, nos hemos centrado en dos concretas. Primero, la gobernanza de los datos, que se refiere a cómo se gestionan los datos utilizados por los sistemas de IA de alto riesgo, y que tiene un impacto directo en los sesgos. Segundo, la gobernanza de la IA que permite ordenar las diversas obligaciones en torno a un sistema de gestión de tal forma que se planifiquen, operen, monitoricen y se proceda a una mejora continua, para lo cual se apuesta por utilizar un estándar como la BS ISO/IEC 42001:2023 (Information technology — Artificial intelligence — Management system, ahora ya norma española mediante la UNE-ISO/IEC: 2025), por los beneficios que se mencionan en el artículo.

Palabras clave:
Gobernanza de datos; Gobernanza de la IA; Sesgos en IA; Gestión de riesgos e IA; ISO 42001.

Abstract:
This article addresses the concept of bias, distinguishing it from other related concepts with which it is sometimes confused. It describes the stages of the AI systems lifecycle, identifying how bias can penetrate each stage in order to manage the resulting risks.

After an overview of the obligations imposed on different stakeholders by the RIA and which can enable the management of risks to fundamental rights, we have focused on two specific aspects. First, data governance, which refers to how the data used by high-risk AI systems is managed, and which has a direct impact on bias. Second, AI governance allows the various obligations to be organized around a management system in such a way that they are planned, operated, monitored and continuously improved, for which it is recommended to use a standard such as BS ISO/IEC 42001:2023 (Information technology — Artificial intelligence — Management system, now a Spanish standard through UNE-ISO/IEC: 2025), due to the benefits mentioned in the article.

Keywords:
Data Governance; AI Governance; Bias in AI; Risk Management and AI; ISO 42001.

1. Introducción

Existe actualmente una conciencia ampliamente extendida de las oportunidades que la IA está suponiendo y va a suponer para la sociedad y la economía. Su impacto posee una amplitud y profundidad superior al del software tradicional, atribuible a su capacidad de aprendizaje autónomo, su escala de despliegue y, en ocasiones, su inherente opacidad. Su progresiva penetración en todos los sectores y contextos (bien sea de forma directa o a través de procesos instrumentales) va a suponer un efecto multiplicador de las oportunidades, pero también un aumento de los sesgos y los posibles riesgos asociados, especialmente aquellos que inciden sobre los derechos fundamentales. Para ello realizaremos una aproximación mostrando no solo los riesgos que pueden suponer y cómo penetran en el ciclo de vida de la IA, sino aportando una visión de la posible oportunidad que ello también puede suponer para gestionarlos. Las tecnologías de IA no sólo reflejan los valores y decisiones de quienes las crean y utilizan, sino que pueden ser un propagador de los sesgos que conviven con nosotros "desde siempre". Sin embargo, el propio proceso de su diseño y el hecho de tenerlos identificados constituye el primer paso para su correcta gestión.

Asimismo, también se profundiza en algunos de los "controles" que el RIA contempla para su gestión, focalizando el análisis tanto en la gobernanza de los datos como de la IA, y planteando la regulación no como un elemento restrictivo, sino como un vector de oportunidad para la generación de confianza.

Pero antes de adentrarnos en la identificación de los riesgos y en su gestión, resulta imprescindible partir de una aproximación conceptual precisa del término sesgo y de su diferenciación respecto de otros afines con los que frecuentemente se confunde, tales como el error, la discriminación y la exclusión.

La RAE define sesgado/da como relacionado con información "tendenciosa" y ésta a su vez como *"que manifiesta parcialidad, obedeciendo a una tendencia o idea determinadas"*.

Por su parte, la Organización Internacional de Normalización (ISO) define sesgos como "diferencia sistemática de trato de determinados objetos, personas o grupos en comparación con otros"[1].

Por tanto, un resultado inexacto puede derivarse bien de un sesgo o bien de un error. La distinción fundamental reside en que los errores suelen ser aleatorios, no predecibles y mitigables mediante un incremento en el volumen de datos o mejoras en el hardware. En cambio, los sesgos en Inteligencia Artificial no son simples errores aleatorios, sino que obedecen a patrones sistemáticos.

A veces también se confunden los sesgos con las posibles consecuencias negativas que – en ocasiones, pero no siempre[2] – puedan tener los mismos.

[1] ISO/IEC 22989:2022(E) cláusula 3.5.4.

[2] La desviación de la verdad que se produce en los sesgos puede contribuir a resultados diversos: perjudiciales o discriminatorios, ser neutra, o incluso puede ser beneficiosa. Por ejemplo, en el ámbito de la selección de personas, por ejemplo, los sistemas de IA pueden tener muchos beneficios, pero los sesgos, pueden suponer consecuencias que pueden ser tanto negativas, positivas como neutras. Un caso de sesgo negativo que podría llegar a discriminación por razón del sexo podría ser el siguiente: un sistema de IA en el que se hayan utilizado datos de entrenamiento sesgados (por ejemplo, la búsqueda de un perfil que ha desempeñado históricamente mayormente un sexo) utilizará ese sesgo en la fase de inferencia y – por tanto – producirá un resultado discriminatorio hacia ese sexo puesto que el hecho de que históricamente hayan desempeñado ese rol en un sexo no significa que lo vayan o deban desempeñar mejor en el futuro las personas de ese sexo. Pero es posible también que dicho sesgo genere un resultado positivo. Por ejemplo, si el sistema de IA se ha entrenado con perfiles muy cualificados, el resultado (desde esa óptica) puede ser positivo pues ofrece a los candidatos más capacitados. Ahora bien, es posible que dicho "sesgo positivo de capacitación" al beber de datos históricos de una profesión basculada históricamente hacia un sexo o una clase social pueda producir un sesgo que sea negativo y discriminatorio.

Al examinar la posible incidencia negativa sobre derechos fundamentales, pueden identificarse ejemplos en distintos ámbitos. No obstante, en la práctica, los casos que han adquirido mayor relevancia pública suelen vincularse a la igualdad ante la ley y, por tanto, la discriminación y la generación de posibles injusticias. Con todo, que no todos los sesgos implican un trato discriminatorio, ni desemboca de forma automática en una vulneración de derechos[3]:

> "El sesgo se refiere a una diferencia sistemática en el tratamiento de ciertas personas o grupos, sin implicar necesariamente si esta diferencia es "correcta" o "incorrecta". Por el contrario, la discriminación y la equidad introducen un juicio de valor sobre los resultados de un tratamiento sesgado. Un sistema de IA sesgado puede producir resultados que pueden considerarse "discriminatorios" o "injustos", según el contexto y los valores aplicados".

Otra posible consecuencia de los sesgos es la exclusión. A diferencia de la discriminación, que supone una situación de desventaja[4] comparativa, la exclusión constituye una forma de desigualdad que se materializa al impedir a la persona o grupo acceder a determinados servicios o recursos[5].

En este sentido, no todo error puede considerarse sesgo, ni todo sesgo es necesariamente negativo, del mismo modo que no todos los sesgos negativos resultan discriminatorios, ni toda discriminación deriva en una situación de exclusión.

Por último y como un fin, pero también como un mecanismo para evitar las consecuencias negativas (injusticias) que puedan generar los sesgos, hay que hacer referencia a otro concepto diferente y relacionado que es la equidad. En el contexto de la IA, la injusticia puede entenderse como el "trato diferencial injustificado que beneficia preferentemente a ciertos grupos sobre otros"[6] y "la equidad, por lo tanto, es la ausencia de tal trato diferencial injustificado o prejuicio hacia cualquier individuo o grupo"[7].

La equidad no significa que deba de tratarse de forma distinta a diferentes personas o grupos, pero que sí que es posible que deba de hacerse ese trato diferente para precisamente conseguir corregir desequilibrios o una representación incorrecta que suponen una injusticia.

Una vez realizadas estas precisiones conceptuales básicas, podemos abordar una aproximación general a la clasificación de los sesgos sobre la base de los tres tipos que, según el NIST[8], afectan a los sistemas basados en Inteligencia Artificial, complementando dicha clasificación con aportaciones propias:

1. Los sesgos computacionales, entendidos como aquellos que pueden identificarse y cuantificarse a partir de un modelo de Inteligencia Artificial ya entrenado, son la punta del iceberg. Aunque se pueden tomar medidas para subsanarlos, no constituyen todas las fuentes de posibles sesgos que intervienen y surgen a partir de errores que resultan cuando la muestra no es representativa de la población. Estos sesgos surgen de factores sistemáticos y no aleatorios, y pueden producirse sin que exista prejuicio, parcialidad o intención discriminatoria alguna.

En los sistemas de IA, estos sesgos están presentes en los conjuntos de datos y procesos algorítmicos utilizados en el desarrollo de aplicaciones de IA y, a menudo, surgen cuando los algoritmos se entrenan con un tipo de datos y no puede extrapolar más allá de esos datos.

Aquí podemos encontrar varios tipos a su vez como el sesgo de selección de datos, que se produce

[3] Como se cita en esta guía práctica de Rhite: Suzanne Snoek e Isabel Barberá, *From Inception to Retirement: Addressing Bias Throughout the Lifecycle of AI Systems: A Practical Guide* (Rhite, 5 de septiembre de 2024). Vid. https://rhite.tech/files/From-Inception-to-Retirement-Addressing-Bias-Throughout-the-Lifecycle-of-AI-Systems.pdf

[4] Por ejemplo, considerar que alguien con una discapacidad no relacionada con el trabajo pueda ser peor candidato que otro que no tiene dicha discapacidad.

[5] Por ejemplo, pensemos en un sistema de IA que no contempla opciones de vehículos adaptados para personas con determinadas discapacidades y los excluye.

[6] ISO/IEC 22989:2022.

[7] Snoek y Barberá, *From Inception to Retirement*.

[8] El National Institute of Standards and Technology (NIST) - agencia del gobierno de los EE. UU. encargada de promover la innovación y la competencia industrial - publicó en marzo de 2022 la NIST Special Publication 1270 con el título "Towards a Standard for Identifying and Managing Bias in Artificial Intelligence" cuyo objetivo es ofrecer una guía para futuro estándares en identificación y gestión de sesgos. Reva Schwartz et al., *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence*, NIST Special Publication 1270 (Gaithersburg, MD: National Institute of Standards and Technology, 2022), https://doi.org/10.6028/NIST.SP.1270

cuando los datos utilizados para entrenar un modelo no representan de manera equitativa la realidad[9]; o el sesgo algorítmico que se produce cuando los algoritmos favorecen ciertos resultados o a ciertos grupos[10].

2. Por su parte, el sesgo humano es el que tenemos cada uno de nosotros de manera implícita y que afecta a cómo percibimos e interpretamos la información que recibimos.

Los prejuicios humanos reflejan errores sistemáticos en el pensamiento humano basados en un número limitado de principios heurísticos y predicción de valores hasta operaciones de juicio más simples. Se conocen también como sesgos cognitivos al estar relacionados con la forma en que los seres humanos procesamos información y en cómo tomamos decisiones a partir de ella.

Dentro del sesgo humano podemos encontrar diversas categorías: como los sesgos de confirmación, que se producen cuando el sistema de IA se basa en creencias o asunciones preexistentes en los datos al existir una tendencia a buscar, interpretar y recordar aquellos datos que refuerzan dichas creencias u opiniones preexistentes[11]; los sesgos de anclaje, que se producen cuando hay una dependencia excesiva de la primera información o ancla[12]; el efecto halo, que valora a una persona o cosa en función de una característica sobresaliente[13]; y el sesgo de negatividad, que se produce cuando se da un peso mayor a la información negativa que a la positiva[14].

3. Y, por último, hay que considerar el conocido como sesgo sistémico, que es el que se encuentra integrado en la sociedad y en las instituciones por razones históricas. Este no tiene por qué ser el resultado de ningún prejuicio o prejuicio consciente, sino más bien de que la mayoría siga reglas o normas existentes (el racismo y el sexismo son los ejemplos más comunes).

Las distintas clases de discriminación y exclusión constituyen algunas de las posibles consecuencias negativas derivadas de los sesgos y, por tanto, riesgos asociados a su manifestación en sistemas de IA[15].

La tipología de posibles sesgos y riesgos son numerosos y diversos, aunque de ciertos sesgos no se habla porque no producen discriminación a pesar de poder tener efectos en decisiones que supongan consecuencias negativas, como por ejemplo para la salud.

Dado que vamos a poner especial foco en la posible afectación a los derechos fundamentales conviene recordar, aunque sea obvio, algo que ya hemos indicado en otra obra[16]: "*los derechos humanos no son "algo nuevo" y con los siglos se ha producido una evolución en cuanto a su número, los sujetos beneficiarios, así como en cuanto a su alcance territorial; aunque su aplicación real diste de ser verdaderamente global. Pero, incluso en donde existe una "cultura" y sistemas tuitivos de los derechos humanos (como es el caso de Europa), la evolución industrial primero y la tecnológica después aportaron oportunidades y riesgos para los mismos que tuvieron que ser gestionadas*".

[9] Por ejemplo, si un algoritmo de contratación se entrena principalmente con currículos de hombres, podría sesgar las decisiones hacia candidatos masculinos.

[10] Por ejemplo, un sistema de crédito que penaliza automáticamente a personas de bajos ingresos debido a su historial financiero podría tener un sesgo algorítmico.

[11] Por ejemplo, si un algoritmo de recomendación de películas solo sugiere géneros específicos a un usuario, podría reforzar sus preferencias anteriores.

[12] Por ejemplo, si un sistema de precios en línea muestra un precio inicial alto, los usuarios pueden percibir como "oferta" o precio barato cualquier precio inferior a ese anclaje inicial.

[13] Por ejemplo, asumir que un candidato con una universidad prestigiosa en su currículo es automáticamente más competente.

[14] Por ejemplo, imaginemos un sistema de detección de fraudes podría ser más propenso a identificar falsos positivos debido a este sesgo derivado de que está entrenado con información "negativa".

[15] Eduard Chaveli Donet, "Sesgos en la IA (I): La necesaria distinción entre sesgos y conceptos afines," *Telefónica Tech Blog*, 11 de febrero de 2025, https://telefonicatech.com/blog/sesgos-en-ia-parte-i-distincion-entre-sesgos-y-conceptos-afines

[16] Lorenzo Cotino Hueso y Pere Simón Castellano, dirs., *Tratado sobre el reglamento de inteligencia artificial de la Unión Europea* (Madrid: Aranzadi, 2024).

2. La introducción de los sesgos en las diferentes etapas del ciclo de vida de los sistemas de IA

Antes de adentrarnos en las medidas para la gestión de los riesgos que los sesgos pueden generar para los derechos fundamentales, entendemos apropiado realizar una aproximación previa a la forma en que dichos sesgos se identifican y se incorporan en los sistemas de IA. El NIST considera que las organizaciones que diseñan y desarrollan la tecnología de IA utilizan el ciclo de vida de la IA para realizar un seguimiento de sus procesos y asegurar que la tecnología sea funcional, pero no necesariamente identifican posibles riesgos y daños y los gestionan.

Por otro lado, los enfoques actuales sobre los sesgos tienden a clasificarlos en función de su tipología (es decir, estadístico, cognitivo) o caso de uso y sector industrial (es decir, contratación, atención sanitaria, etc.). No obstante, estas aproximaciones pueden limitar la perspectiva necesaria para gestionar eficazmente el sesgo, entendido como un fenómeno intrínsecamente contextual. Por ello, el documento del NIST propone un enfoque para gestionar y reducir los efectos de los sesgos perjudiciales en todos los contextos teniendo en cuenta los "lugares o momentos" clave dentro de las etapas del ciclo de vida de un sistema de IA.

Identificar las fuentes de sesgo es el primer paso en cualquier estrategia de mitigación. Dichas fuentes pueden aparecer en múltiples fases del sistema de IA, lo que exige comprender con claridad cuáles son las etapas del ciclo de vida de los sistemas de IA y, a su vez, ponerlo en relación con el rol y tareas concretas que tienen los diferentes operadores. No es lo mismo cuando hablamos de un proveedor[17] que de un responsable del despliegue"[18] , por citar dos ejemplos muy claros. Por ello a continuación, realizaremos una visión general que hay que aterrizar en cada caso.

El desarrollo de un sistema de IA tiene unas fases básicas y, aunque hay diferentes aproximaciones o formas de englobarlas[19], consideramos que esta es la más clara de entender y la que sirve mejor para aterrizar los sesgos en cada una de ellas:

FASE 1. FASE DE INICIO: PRE-DISEÑO O DEFINICIÓN DEL ALCANCE

Los sistemas de IA comienzan en la fase de pre-diseño en la que:

a. Se identifica el problema que se quiere resolver con el sistema de IA (es decir: los objetivos). Si los objetivos del sistema están influenciados por prejuicios, el sistema reflejará esos sesgos. Por ejemplo, si se decide que un sistema de contratación debe priorizar candidatos de ciertas universidades que podemos calificar como prestigiosas, puede excluir a candidatos igualmente cualificados de otras universidades.

Por tanto, nos encontramos aquí con sesgos institucionales o sistémicos, por lo que hay que revisar esos posibles perjuicios introduciendo diversas medidas que van desde la revisión de los datos utilizados hasta la evaluación de los posibles impactos. Para ello hay que involucrar a expertos en áreas como, por ejemplo, la ética y los derechos fundamentales.

b. Se recopilan los requisitos funcionales (qué debe hacer el sistema) y no funcionales del sistema (cómo debe comportarse el sistema). Aquí podemos incluir la recopilación de datos en cuanto al análisis de requisitos (determinar qué datos son necesarios para resolver el problema) y la recopilación preliminar (obtener datos iniciales para entender mejor el ámbito y los desafíos a los que se enfrenta).

Si los datos recopilados no son representativos de la población objetivo, el modelo aprenderá patrones incorrectos. Por ejemplo, si se recopilan datos de salud solo de una región específica, el sistema puede no funcionar bien en otras regiones con diferentes características demográficas.

[17] En los términos del RIA "*aquellas personas que desarrollen un sistema de IA o modelo de IA de uso general para introducirlo en el mercado de la Unión Europea bajo su propio nombre o marca comercial*". Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial (Reglamento de Inteligencia Artificial), art. 3.

[18] Según el RIA "*toda persona que utiliza un sistema de IA bajo su propia autoridad en el ámbito profesional*". Así, toda organización que incorpore un sistema de IA en sus procesos actuaría como responsable del despliegue de dicho sistema. Reglamento (UE) 2024/1689, art. 3.

[19] Sobre las fases de un sistema de IA pueden consultarse, por ejemplo: Keyrus, "El ciclo de vida de la inteligencia artificial," *Keyrus Insights*, https://keyrus.com/sp/es/insights/el-ciclo-de-vida-de-la-inteligencia-artificial-alcance-diseno-de-modelos-y/; Enzyme Advising Group, "Las tres fases para poner en marcha tu modelo de IA," *Enzyme Blog*, https://enzyme.biz/blog/el-ciclo-de-vida-de-ia-pasos-para-poner-en-marcha-tu-modelo; o una aproximación más detallada en la figura 3 de *ISO/IEC 22989:2023, Artificial Intelligence — Artificial Intelligence Concepts and Terminology*.

Se trata de trampas de abstracción derivadas de traducir del "mundo real al sistema de IA" y "sus entradas y salidas se simplifican en exceso o se ignoran".

Las principales trampas de abstracción incluyen la de formalismo, la del efecto dominó y la del solucionismo.

c. Se analiza la viabilidad tanto técnica como económica y operativa del proyecto.

d. Se seleccionan las tecnologías para desarrollar el sistema de IA.

f. Y se planifica el proyecto

FASE 2. FASE DE DISEÑO Y DESARROLLO

En esta etapa del ciclo de vida de la IA se llevan a cabo decisiones de calado como si se van a realizar determinados desarrollos o se van a adquirir, si se utilizarán soluciones de código abierto o propietario etc.

Los principales hitos son los siguientes:

1. Se analizan los requisitos y datos disponibles, lo que supone incluir la recopilación, limpieza y preparación de los datos concretos que son necesarios para entrenar el modelo de IA.

2. Se diseña y desarrolla el modelo de IA, lo que incluye la selección de algoritmos y técnicas de aprendizaje.

3. Se lleva a cabo el entrenamiento del modelo con los datos y se ajustan los parámetros para optimizar su rendimiento.

4. Se prueba y valida el modelo para asegurar que cumple con los requisitos y funciona correctamente.

5. Se realizan iteraciones para mejorar el modelo basado en los resultados de las pruebas.

Dada la importancia del diseño en el resultado, los sesgos de validez del constructo (de la construcción teórica para entender el problema) son especialmente importantes en esta fase. Este se produce cuando una variable no mide con previsión el constructo que se quiere representar para la construcción del sistema de IA, cuando hablamos de problemas complejos. Por ejemplo: imaginemos que confundimos el estatus socioeconómico con un elemento único como los ingresos, cuando realmente hay otros que pueden afectarlo como la educación, la riqueza, ocupación o prestigio, por ejemplo. Por tanto, hay que tener en cuenta diversas medidas de dichos elementos complejos y considerar diversas "formas" de interpretarlos como, por ejemplo, las derivadas de diferentes visiones culturales.

También es posible que los algoritmos hayan sido diseñados de manera que favorezcan ciertos resultados[20].

Por otro lado, aquí también se procede a la comprensión y preparación de los datos, siendo el sesgo de representación el más común y el más tratado en esta fase. Si el modelo se entrena con datos sesgados, aprenderá esos sesgos[21]. Para ello hay que garantizar la representación correcta, pudiendo utilizar, por ejemplo, técnicas de muestreo.

Otros sesgos importantes en esta fase son el sesgo de medición, el sesgo histórico, el sesgo de selección (de los que hay muchos ejemplos[22]) y el sesgo de etiquetado.

[20] Por ejemplo, un algoritmo de recomendación de empleo que prioriza ciertos términos en los currículums (como "agresivo", pensemos en "comercial agresivo" o "ejecutivo agresivo") puede sesgarse hacia candidatos masculinos, ya que estos términos son más comúnmente utilizados por hombres o por ejemplo términos como persona empática y colaborativa, más asociados a mujeres.

[21] Por ejemplo: 1.Si la población objetivo-definida no representa correctamente la población de uso posterior. Por ejemplo, un sistema de reconocimiento facial entrenado con imágenes mayoritariamente de personas de piel clara puede tener dificultades para reconocer a personas de piel más oscura. 2.O si hay grupos subrepresentados en dicha población puede no funcionar bien parta dicho subgrupo.

[22] Algunos ejemplos de sesgos de selección son el de muestreo, el de autoselección, el sesgo de cobertura y el sesgo de observación.

Durante la fase de desarrollo se construyen los modelos seleccionados y se entrenan los datos.

Los principales sesgos de esta fase son los sesgos de algoritmo. La principal característica de este sesgo es que no está en los datos, sino en el propio algoritmo. Un ejemplo de ello es que un algoritmo de selección de personas utilice datos de entrenamiento equilibrados asignando más peso a algún criterio que no tenga que ver con el posible rendimiento.

Hay diversos tipos de sesgos de algoritmo, pero en esta fase algunos que inciden en el desarrollo son los siguientes: el sesgo de agregación, el sesgo de variable omitida y el sesgo de aprendizaje.

Como dice el NIST, al final de la fase de diseño y antes del despliegue es necesaria una evaluación exhaustiva de la mitigación de sesgo para garantizar que el sistema se mantenga dentro de los límites preespecificados, lo que debe incluir:

- Las fuentes de sesgo identificadas
- Las técnicas de mitigación implementadas
- Evaluaciones de desempeño relacionadas antes de que el modelo pueda lanzarse para su implementación.

Como medidas para solucionar dichos riesgos, el NIST (en su informe citado) menciona por ejemplo el "desafío cultural efectivo" (*cultural effective challenge*), una práctica que busca crear un entorno en el que los desarrolladores de tecnología puedan desafiar y cuestionar activamente los pasos en el modelado y la ingeniería para ayudar a erradicar los sesgos estadísticos y los sesgos inherentes a la toma de decisiones humanas. Aunque lo hemos incardinado en esta fase debería ser iterativo. Como ya hemos sostenido anteriormente[23], pensamos que la existencia de una "parada formal", quizá incluso su constancia en un informe como sucede por ejemplo en las evaluaciones de impacto en protección de datos conforme al RGPD o las EIDF del RIA, sería una buena forma de "obligar a realizar dicha parada" y a que tuviese el calado oportuno. Si derivado de ello se constatase el sesgo en los algoritmos y el impacto que ello podría tener podría/debería incluso evitarse avanzar en las siguientes fases.

FASE 3. FASE DE PRUEBA Y EVALUACIÓN (VERIFICACIÓN Y VALIDACIÓN ANTES DEL DESPLIEGUE)

En esta fase:

1. Se monitoriza el rendimiento del modelo previo al despliegue. Si las métricas de evaluación no consideran la equidad, el modelo puede parecer preciso, pero ser injusto y una vez desplegado posteriormente, el sistema puede perpetuar y amplificar los sesgos existentes. Por ejemplo, un sistema de recomendación de préstamos, que ha sido entrenado con datos históricos, puede continuar discriminando a ciertos grupos si esos datos reflejan prácticas discriminatorias pasadas.

2. Se actualiza el modelo, se ajusta con datos de entrenamiento y validación, y se procede a las mejoras necesarias.

En esta fase se puede producir un sesgo denominado sesgo de evaluación, que es cuando los procedimientos o las métricas a utilizar para evaluar el modelo no están alineadas con el modelo a implementar.

Por tanto, hay que adoptar medidas como, por ejemplo, evaluar las métricas, los ajustes de datos e idealmente esta evaluación debería realizarse junto con otras partes interesadas para garantizar que todos los problemas previamente identificados se resuelvan a satisfacción de todos.

FASE 4. FASE DE DESPLIEGUE O IMPLEMENTACIÓN

Es en esta fase en la que los responsables del despliegue ya trabajan con esta tecnología, pues pasan del desarrollo a su implantación en el entorno de producción. Hay que tener en cuenta

[23] Eduard Chaveli Donet, "Sesgos en la IA (V): Introducción de los riesgos en el ciclo de vida de los sistemas de IA (parte 1)," *Telefónica Tech Blog*, 22 de abril de 2025, https://telefonicatech.com/blog/sesgos-ia-v-introduccion-riesgos-en-ciclo-de-vida-sistemas-de-ia-parte-1

especialmente el sesgo de implementación, que se produce si el sistema se implementa en un entorno que no refleja las condiciones del entrenamiento, lo que supone que puede comportarse de manera sesgada. Por ejemplo, un sistema de traducción automática entrenado principalmente con textos formales puede no funcionar bien con lenguaje coloquial.

Las trampas de abstracción (de las que ya hemos hablado) también son propias de esta fase.

FASE 5. FASE DE OPERACIÓN Y MONITORIZACIÓN

Cuando los sistemas están en producción (operando) se deben de monitorizar y realizar los cambios necesarios tanto a nivel de hardware y software, como efectuar ajustes en el propio algoritmo, en los datos utilizados (añadiendo, limpiando etc.), etc.

En los sistemas que utilizan aprendizaje continuo[24] se producen más estos sesgos a diferencia de los sistemas de IA en los que no[25].

En esta fase se puede producir el conocido como "bucle de retroalimentación de refuerzo", lo que conlleva que al volver a entrenar el modelo con los nuevos datos y partiendo de un sesgo inicial no solucionado, se pueda propagar el sesgo, a lo que se puede añadir, por ejemplo, el sesgo de automatización que puede tener un efecto multiplicador. Para ello hay que establecer mecanismos de retroalimentación continua para identificar posibles sesgos y corregirlos en tiempo real.

FASE 6. VALIDACIÓN CONTINUA

La "validación continua" está interrelacionada con el monitoreo y a veces se superponen, pero en teoría son cosas diferentes:

1. En el monitoreo se supervisa el rendimiento del modelo de IA en producción en tiempo real revisando métricas para ver si hay alguna que detecta posibles anomalías. Se trata de asegurar que el modelo continúa funcionando bien y de detectar posibles problemas cuanto antes.

2. En la validación continua se trata de evaluar regularmente el modelo con nuevos datos para ver si continúa siendo preciso.

3. La reevaluación es una revisión completa y más profunda en la que se analiza periódicamente el modelo completo y su rendimiento general para ver si se requieren cambios más importantes.

En consecuencia, la validación continua se puede realizar en sistemas de IA donde no aplica el aprendizaje continuo para, por ejemplo "detectar desviaciones de datos, de conceptos o para detectar cualquier mal funcionamiento técnico" (ISO/IEC 5338), pero es especialmente relevante con nuevos datos, por lo que es fundamental en los supuestos de aprendizaje continuo donde el reentrenamiento existe aunque no sea explícito. En los sistemas con aprendizaje continuo, los modelos integran nuevos datos de forma continua sin un reentrenamiento explícito, por lo que es fundamental: comprobar la coherencia de los datos en producción con los iniciales del entrenamiento y tener que actualizar los propios datos de prueba.

Por tanto, los principales sesgos en esta fase son los sesgos de los datos, de entre los que cabe destacar los de representación, selección, medición, el de etiquetado y el de *proxies*, por lo que habrá que poner especial foco en las medidas para gestionarlos en esta fase.

FASE 7. RE-EVALUACIÓN

A diferencia del monitoreo y validación continua que se refieren a ajustes constantes cada uno con la finalidad que hemos visto, la de reevaluación es más profunda.

Aparte de los sesgos de evaluación y las trampas de abstracción que ya conocemos, y que en

[24] Pensemos por ejemplo en los asistentes virtuales como Alexa de Amazon o el asistente de Google, por citar varios que aprenden y se actualizan con las interacciones de los usuarios.

[25] Por ejemplo, los "sistemas de calculadoras matemáticas" para operaciones complejas que se basan en reglas predefinidas u otros que se basan en reglas predefinidas y no aprenden de forma continua.

estas fases pueden servir para refinar el sistema con decisiones, hay varios propios de esta fase:

1. Falacia del costo volcado a la suma: es la tendencia de las personas u organizaciones a seguir invirtiendo en un proyecto o esfuerzo en el que ya han invertido recursos significativos (dinero, tiempo, esfuerzo, reputación etc.), incluso cuando los costos actuales superan los beneficios potenciales y de hecho cuanto mayor es el costo volcado mayor es el sesgo.

2. Y aunque tienen relación hay otro sesgo diferente que es el sesgo de *estatus quo*, que se refiere a la preferencia por mantener el estado actual de las cosas.

En ambos casos es fundamental que las partes interesadas lo conozcan, reconozcan y tomen decisiones al respeto.

FASE DE RETIRO

Incluso si se toma la decisión de retirar el sistema, lo que se puede deber a diferentes motivos (no sirve a los propósitos, se ha buscado otra solución, se entiende que no es justo etc.), ello puede producir un sesgo conocido como sesgo histórico (o sesgo de legado) dado que el sistema se ha entrenado con datos históricos sesgados que se replican. Un ejemplo son los algoritmos de recomendación de noticias que se puedan basar en aquellas más relevantes, aunque quizá no sean más verídicas o contrastadas. Obviamente al que era usuario de ese sistema ya no le afectará, pero que les afectará a otros usuarios del sistema de IA que lo adquieran o usen.

3. Gestión de los riesgos derivados de los sesgos. Especial referencia a la gobernanza de los datos y gobernanza de la IA

3.1. Introducción

Antes de adentrarnos en las medidas previstas en el RIA para gestionar los riesgos que los sesgos pueden generar para los derechos fundamentales, resulta necesario formular algunas consideraciones generales previas sobre los mismos.

En primer lugar, que cuando hablamos de riesgos es necesario distinguir tres categorías diferenciadas: los riesgos del propio sistema de IA, los riesgos del sistema de gestión de IA (especialmente relevante si hablamos de un sistema certificado bajo un estándar, en concreto la BS ISO/IEC 42001:2023 (Information technology — Artificial intelligence — Management system, ahora ya norma española mediante la UNE-ISO/IEC: 2025), en adelante ISO 42001; y, por último, los riesgos para los derechos fundamentales.

Mientras que en los dos primeros casos el análisis se orienta principalmente al "dolor" organizativo (aunque obviamente pueda irradiar a terceros), en el caso de las evaluaciones de impacto en derechos fundamentales, el enfoque se desplaza a la posible afectación a terceros (individuos o colectivos).

En segundo lugar, el enfoque respecto de los riesgos del RIA, según se desprende de su Considerando 14 – consiste en "*introducir un conjunto proporcionado y eficaz de normas vinculantes para los sistemas de IA" adaptando "el tipo y el contenido de tales normas a la intensidad y el alcance de los riesgos que pueden generar los sistemas de IA*". Esto lo hace el RIA mediante la siguiente fórmula que algunos autores han criticado[26]: "*prohibir determinadas prácticas inaceptables de inteligencia artificial, establecer requisitos para los sistemas de IA de alto riesgo y obligaciones para los operadores correspondientes, y establecer obligaciones de transparencia para determinados sistemas de IA*".

Este enfoque basado en riesgos implica que el propio RIA incorpora ya una medida normativa de gestión del riesgo: la prohibición de ciertos sistemas y la diferenciación entre niveles de riesgo.

[26] Mantelero indicaba que *"Aunque esto es eficaz en términos de impacto político y aceptabilidad, es una forma débil de prevención de riesgos. La Propuesta hace una distinción bastante rígida entre el riesgo de alto nivel y el resto, no proporcionando ninguna metodología para evaluar el primero, y eximiendo en gran medida al segundo de cualquier mitigación (con la limitada excepción de las obligaciones de transparencia en ciertos casos)"*. Alessandro Mantelero, *Artificial Intelligence and Data Protection: Challenges and Possible Remedies* (Estrasburgo: Consejo de Europa, 2019), 173.

Asimismo, existen ejemplos históricos de discriminación vinculada a sesgos (tanto reales como otros recreados en obras de ficción[27]) que ya no deberían de producirse en el marco del RIA, pues los prohíbe "de plano".

La tercera reflexión necesaria es que el riesgo cero, en términos generales, no existe. Tampoco en el ámbito de la IA. Por ello, debemos de abordar el tema del nivel de riesgo aceptable. Como hemos mantenido ya anteriormente[28] el artículo 9.4. del RIA señala que las medidas de gestión de riesgos "*serán tales que el riesgo residual pertinente asociado a cada peligro, así como el riesgo residual global de los sistemas de IA de alto riesgo, se consideren aceptables"*. Por tanto *"asumiendo lege data la utilización de la metodología de gestión de riesgos que ha establecido el RIA y asumiendo también que no pueden tolerarse impactos altos en derechos fundamentales, el riesgo aceptable será por definición bajo*". En consecuencia, deberán de adoptarse salvaguardas y controles que permitan reducir el riesgo inicial hasta situarlo por debajo del umbral de tolerancia definido con carácter previo.

La protección de los derechos fundamentales constituye, junto con la salud y la seguridad, uno de los objetivos esenciales del RIA. En consecuencia, esta protección debe de integrarse en todas las fases del ciclo de vida de los sistemas de IA, como parte inherente del enfoque basado en el riesgo. de manera distinta según el tipo de sistemas, y el operador involucrado, lo que podemos ver de forma general en esta tabla (referida a los sistemas de alto riesgo):

Obligaciones sistemas de alto riesgo	Proveedor	Responsable despliegue	Representante autorizado	Importador	Distribuidor
Art 8 Cumplimiento de los requisitos sistemas Alto riesgo (Art 8 a 15)	X				
Art. 9 gestión de riesgos	X				
Art. 10 Gobernanza de datos	X				
Art. 16 Obligaciones de los proveedores de sistemas de IA de alto riesgo	X				
Art. 17 Sistema de gestión de la calidad	X				
Art. 18 Conservación de la documentación	X				
Art 19. Archivos de registro generados automáticamente	X				
Art 20. Medidas correctoras y obligación de información	X				
Art .21 Cooperación con las autoridades competentes	X				
Art. 22 Representantes autorizados de los proveedores	X		X		
Art. 23 Obligaciones de los importadores				X	X
Art. 24 Obligaciones de los distribuidores					X
Art. 25 Responsabilidades cadena de valor de la IA		X	X	X	X
Art. 26 Obligaciones responsables del despliegue		X			
Art. 27 EIDF		X			
Art. 43 Evaluación de la conformidad	X				

[27] Por ejemplo, sistemas que permitan evaluar o predecir la probabilidad de que una persona física cometa una infracción penal; o los más recientes intentos de poner de moda en algunos entornos los Sistemas de puntuación social.

[28] Eduard Chaveli Donet, "La evaluación de impacto de derechos fundamentales por quienes despliegan sistemas de inteligencia artificial en el Reglamento", en *Tratado sobre el Reglamento de Inteligencia Artificial de la Unión Europea*, dirs. Lorenzo Cotino Hueso y Pere Simón Castellano (Las Rozas, Madrid: Editorial Aranzadi, 2024), 495–533.

Obligaciones sistemas de alto riesgo	Proveedor	Responsable despliegue	Representante autorizado	Importador	Distribuidor
Art. 47 Declaración UE de conformidad	X				
Art. 48 Marcado CE	X				
Art. 49 Registro	X	X	X		
Art. 52 transparencia hacia los usuarios		X			
Art. 72 Vigilancia poscomercialización (Alto riesgo)	X	X			
Art. 73 Notificación de incidentes graves	X	X			
Art. 86 Derecho a explicación de decisiones tomadas individualmente		X			

Fuente. Curso RIA de la AEC. Equipo Govertis, part of Telefónica Tech.

Tras la lectura sistemática del RIA, se aprecia que existen obligaciones estrechamente vinculadas con la gestión de los riesgos para los derechos fundamentales: los análisis de riesgos como parte de los sistemas de gestión y, en su caso, de los procesos de evaluación de conformidad; las exigencias adicionales impuestas a los modelos de propósito general con riesgo sistémico (como disponer de documentación técnica, políticas de gobernanza y supervisión), en las que aunque no exista una referencia explícita es normal tenerlos en cuenta; por supuesto en las EIDF, cuando son requeridas y dónde sí que se pone absoluto foco en ellos; y también en el caso de la supervisión humana puesto que el elemento humano, que está en la base de los sesgos, también puede desempeñar un papel decisivo mediante vigilancia humana, la participación de las partes interesadas o la intervención de los profesionales cualificados. Finalmente, la calidad de los datos y - por tanto - su gobernanza, constituye la herramienta fundamental a la que dedicamos el apartado siguiente.

3.2. La gobernanza de datos como herramienta para gestionar los riesgos derivados de los sesgos

La importancia y el valor de los datos —tanto personales como no personales— para la economía y la sociedad no constituye un fenómeno reciente. Debe recordarse que el propio Reglamento General de Protección de Datos (RGPD) no se limita a la protección de las personas físicas en lo que respecta al tratamiento de sus datos personales, sino que también promueve la libre circulación de estos datos dentro de la Unión Europea. Tal y como enfatiza el considerando 6 del RGPD, "*la tecnología ha transformado tanto la economía como la vida social, y ha de facilitar aún más la libre circulación de datos personales dentro de la Unión y la transferencia a terceros países y organizaciones internacionales", aunque "garantizando al mismo tiempo un elevado nivel de protección de los datos personales*".

Este equilibrio entre protección y circulación ha guiado múltiples iniciativas europeas en la última década para construir un verdadero mercado único de datos, entre ellas la Estrategia Europea de Datos que en 2020 marcó el rumbo hacia una economía basada en datos interoperables, accesibles y reutilizables, a la que se suman otros instrumentos y normas clave aprobados con posterioridad[29]

[29] La Brújula Digital 2030 y el Decenio Digital Europeo, que fijan metas concretas para la transformación digital de Europa en ámbitos como conectividad, competencias digitales, digitalización de servicios públicos y empresas. El Reglamento (UE) 2022/868 de Gobernanza de Datos (Data Governance Act), que establece mecanismos para compartir datos entre sectores públicos y privados de forma segura, voluntaria y bajo condiciones claras. El Reglamento (UE) 2023/2854 sobre normas armonizadas para un acceso justo a los datos y su utilización (Reglamento de Datos), que garantiza que los usuarios de productos conectados puedan acceder a los datos que generan, fomenta la equidad contractual y refuerza la seguridad jurídica. El Reglamento (UE) 2018/1807 sobre la libre circulación de datos no personales, que complementa al RGPD al eliminar obstáculos al movimiento de datos industriales, de IoT o empresariales dentro del mercado único. El Reglamento (UE) 2025/327 sobre el Espacio Europeo de Datos de Salud (EEDS), que crea un marco para el acceso transfronterizo a datos sanitarios electrónicos y su reutilización ética para investigación, innovación y formulación de políticas, con especial atención a la protección de datos sensibles; Reglamento (UE) 2024/2847 del Parlamento Europeo y del Consejo, de 23 de octubre de 2024, relativo a los requisitos horizontales de ciberseguridad para los productos con elementos digitales y por el que se modifica el Reglamento (UE) nº 168/2013 y el Reglamento (UE) 2019/1020 y la Directiva (UE) 2020/1828 (Reglamento de Ciberresiliencia, CRA) que pretende garantizar que los productos que generan, procesan o transmiten datos lo hagan de forma segura lo que refuerza la confianza en la infraestructura digital, lo cual es esencial para que los datos puedan circular, compartirse y reutilizarse sin poner en riesgo a los usuarios ni a las empresas.

así como otros que están fase de elaboración actualmente.

Este ecosistema normativo refleja una visión europea que reconoce el valor estratégico de los datos, sin renunciar a la protección de los derechos fundamentales. La gobernanza de los datos -y por extensión, de las tecnologías que los procesan, como la inteligencia artificial- se configura como un elemento clave para garantizar una transformación digital competitiva, pero a la vez respetuosa con los derechos fundamentales.

Para garantizar ese mercado único de datos (gobernanza a nivel "macro") es imprescindible, como sostiene Loza[30], que las organizaciones cuenten con una sólida gobernanza de datos a nivel interno (nivel "micro"), la cual además permitirá avanzar hacia la gobernanza de la inteligencia artificial".

No vamos a entrar en analizar la aproximación detallada al concepto de gobernanza que tiene múltiples aproximaciones - algunas procedentes del ámbito- como el de las TI, pero sí dar unas pinceladas sobre lo que se entiende por gobernanza de los datos y aterrizarla a los sistemas de IA.

Sostiene también esta autora que "*no existe una definición unívoca o normativa para el concepto de Gobernanza de datos*" y que este concepto ha ido evolucionando de referirse al contexto interno de las organizaciones sólo en lo relativo al control y gestión de sus datos a evolucionar hacia un concepto más amplio y elaborado como el que propone el Instituto de Gobernanza de Datos (*Data Governance Institute*), que lo define[31] como "*el ejercicio de la toma de decisiones y la autoridad en asuntos relacionados con los datos", y de forma más amplia, como "un sistema de derechos de decisión y responsabilidades para los procesos relacionados con la información, ejecutados según modelos acordados que describen quién puede tomar qué acciones con qué información, y cuándo, bajo qué circunstancias, utilizando qué métodos*".

Por su parte la Agencia Española de Protección de Datos (AEPD) afirmó ya en 2020 que "*la gobernanza de datos, o data governance, es la estrategia para la correcta administración y gestión de la política de datos en la organización*".

Y que es "*una parte importante de esa política ha de ser la política de protección de datos establecida en el artículo 24 del RGPD y la necesidad de adoptar dichas políticas expresada en el considerando 78*", y concreta los objetivos de la gobernanza de datos, los específicos a añadir cuando se traten datos personales, así como determinados factores de éxito de esta[32].

A partir de las aportaciones doctrinales, pueden sintetizarse las siguientes características fundamentales de la gobernanza de datos[33]:

1. Poner en valor los datos como un activo de la organización que debe gestionarse.

2. Establecer responsabilidades, tanto a nivel estratégico, táctico como operativo.

3. Definir pautas y normas para velar por la calidad de los datos y su uso adecuado.

4. Existencia de un liderazgo estratégico de la dirección que no dependa de un exclusivo departamento o área de la compañía, de forma que, como sistema transversal, sea coherente con los objetivos y cultura de la organización y, por supuesto, con la normativa vigente.

5. Y añadiría que la propia gobernanza de datos exige que se integre con políticas de transformación digital, seguridad y cumplimiento normativo para que dicho flujo de datos sea seguro.

Vamos a centrarnos ahora en la Gobernanza de los datos en sistemas de IA en el marco del RIA

[30] María Loza Corera, "Datos y gobernanza de datos y conexiones con principios protección de datos en el artículo 10 del Reglamento", en *Tratado sobre el Reglamento de Inteligencia Artificial de la Unión Europea*, dirs. Lorenzo Cotino Hueso y Pere Simón Castellano (Las Rozas, Madrid: Editorial Aranzadi, 2024), 567–591

[31] Data Governance Institute, "Definitions of Data Governance," *datagovernance.com*, consultado el 8 de marzo de 2026, https://datagovernance.com/the-data-governance-basics/definitions-of-data-governance/

[32] Agencia Española de Protección de Datos, "Gobernanza y política de protección de datos," *Blog de la AEPD*, 2 de septiembre de 2020, https://www.aepd.es/prensa-y-comunicacion/blog/gobernanza-y-politica-de-proteccion-de-datos

[33] Las tres primeras definiciones proceden de Manuel Salvador Serna, "Inteligencia artificial y gobernanza de datos en la Administración Pública: sentando las bases para su integración a nivel corporativo", en *Repensando la administración pública: administración digital e innovación pública* (Madrid: INAP, 2021), 126–148, a las que hemos añadido el matiz de que entendemos que dichas responsabilidades deben definirse en los diferentes niveles organizativos; la cuarta definición procede de María Loza Corera, Loza Corera, "Datos y gobernanza de datos…".

sin desconocer que existen otros en el ámbito del conocido como *soft law*[34]. En conjunto, estas iniciativas configuran un ecosistema internacional convergente hacia una gobernanza responsable de los datos y de la IA.

Como se ha indicado, el RIA (en su aproximación a riesgos) clasifica los sistemas de IA en diferentes niveles y para cada uno dispone una serie de obligaciones. En concreto, la Gobernanza de datos se regula en el artículo 10, enmarcado en el capítulo II del Título III dedicado a los Sistemas de IA alto riesgo. Sus obligaciones se aplican con independencia de si se tratan datos personales o no, lo que refleja evidencia su carácter transversal. En un sistema basado en datos, no disponer de datos adecuados (de entrenamiento, de validación y prueba etc.) puede generar sesgos que a su vez pueden amplificarse por mecanismos de retroalimentación del propio sistema. Obviamente, muchos sistemas de IA tratan datos de carácter personal, por lo que el RGPD directamente aplicable, aunque el articulado del RIA no dispone expresamente el cumplimiento de ningún principio de protección datos, cosa que sí que hacen los Considerandos. Por ejemplo, el Considerando 67, que hace referencia a la transparencia sobre el fin original de la recopilación de datos; o el Considerando 69, que se refiere a la necesidad de garantizar el derecho a la protección de los datos personales a lo largo de todo el ciclo de vida del sistema de IA; mencionando de forma expresa los principios de minimización de datos y de protección de datos desde el diseño y por defecto, cuando el sistema trate datos personales.

El Considerando 67 también está relacionado con gobernanza de los datos cuando al disponer que "*los conjuntos de datos para el entrenamiento, la validación y la prueba, incluidas las etiquetas, deben ser pertinentes, lo suficientemente representativos y, en la mayor medida posible, estar libres de errores y ser completos en vista de la finalidad prevista del sistema",* añadiendo de forma complementaria el Considerando 69 que *"las medidas adoptadas por los proveedores para garantizar el cumplimiento de estos principios podrán incluir no solo la anonimización y el cifrado, sino también el uso de una tecnología que permita llevar los algoritmos a los datos y el entrenamiento de los sistemas de IA sin que sea necesaria la transmisión entre las partes ni la copia de los datos brutos o estructurados, sin perjuicio de los requisitos en materia de gobernanza de datos establecidos en el presente Reglamento*".

Por otra parte, cuando el artículo 10 exige que los datos estén exentos de errores y sean completos con vistas a la finalidad prevista, entendemos que se está haciendo referencia directa al principio de exactitud, en consonancia con la que afirma la AEPD[35].

Como vemos, la conexión de ambas normas en materia de gobernanza se puede dar en diferentes puntos, pero merece especial referencia el considerando 70 del RIA, que desarrolla el artículo 10.5, al disponer la posibilidad excepcional bajo ciertos requisitos adicionales a las garantías del RGPD, de que los proveedores de dichos sistemas puedan tratar categorías especiales de datos personales siempre que ofrezcan las garantías adecuadas en relación con los derechos y las libertades fundamentales de las personas físicas siempre que además de las disposiciones establecidas en el RGPD, la Directiva (UE) 2016/680 y el Reglamento (UE) 2018/1725, cumplan todas las condiciones: (i) que no pueda realizarse mediante datos sintéticos, anónimos u otros datos; (ii) que las categorías especiales de datos personales tratados se sujeten a limitaciones técnicas en cuanto a la reutilización y a las medidas más avanzadas de seguridad (iii) se sujeten a medidas que garanticen la seguridad y protección de los datos personales tratados (iv) no se transmitirán, transferirán ni serán accesibles de otro modo a terceros; (v) se suprimirán una vez que se haya corregido el sesgo o los datos personales hayan llegado al final de su período de conservación.

Por último, nos parece especialmente relevante la contundencia con la que el Considerando 63

[34] Podemos citar, entre otras: Las OECD AI Principles (2019), adoptadas por más de 40 países, que promueven una IA centrada en el ser humano, responsable y transparente. La OCDE ha desarrollado además el OECD Framework for the Classification of AI Systems, que incluye criterios sobre gobernanza de datos y riesgos. En el ámbito europeo, además del RIA, se han publicado documentos como las Ethics Guidelines for Trustworthy AI del Grupo de Expertos de Alto Nivel en IA de la Comisión Europea (2019), que identifican la gobernanza de datos como uno de los requisitos clave para una IA fiable. La UNESCO Recommendation on the Ethics of Artificial Intelligence (2021), que establece principios globales sobre gobernanza, protección de datos, equidad y supervisión humana, con especial atención a los contextos culturales y sociales. El Artificial Intelligence Risk Management Framework (AI RMF) del NIST (EE.UU.), publicado en 2023, que ofrece un enfoque voluntario para identificar, evaluar y gestionar riesgos asociados al uso de IA, incluyendo la gobernanza de datos como uno de sus pilares fundamentales. Las Governance Guidelines for the Implementation of AI Principles del Gobierno de Japón, que proporcionan recomendaciones prácticas para aplicar principios éticos en el desarrollo y uso de sistemas de IA, con énfasis en la trazabilidad, la transparencia y la calidad de los datos. Así como otras iniciativas como la Global Partnership on AI (GPAI), que impulsa investigaciones y recomendaciones sobre gobernanza responsable de la IA, incluyendo la gestión de datos, la equidad algorítmica y la interoperabilidad.

[35] Agencia Española de Protección de Datos (AEPD), "Inteligencia artificial: principio de exactitud en los tratamientos," *Blog de la AEPD*, 31 de mayo de 2023, https://www.aepd.es/prensa-y-comunicacion/blog/inteligencia-artificial-principio-de-exactitud-en-los-tratamientos

indica algo que pudiera parecer obvio, pero que conviene "subrayar" mediante dicho considerando: "*No debe entenderse que el presente Reglamento constituye un fundamento jurídico para el tratamiento de datos personales, incluidas las categorías especiales de datos personales, en su caso, salvo que el presente Reglamento disponga específicamente otra cosa*", que es el supuesto analizado en el párrafo anterior.

3.3. La gobernanza de la IA como herramienta para gestionar los riesgos derivados de los sesgos

La gobernanza de los datos constituye una dimensión esencial de un marco más amplio: la gobernanza de la IA. Mientras que la gobernanza de datos en el RIA se centra en cómo se gestionan los datos utilizados por los sistemas de IA, especialmente los de alto riesgo, la gobernanza de la IA abarca todo el ciclo de vida (desde el diseño hasta el despliegue y supervisión) y diversas tareas que deben de estar "correctamente ordenadas".

El RIA no ofrece una definición de gobernanza de la IA ni tampoco encontramos una definición en la ISO *Information technology-Artificial intelligence -Artificial intelligence concepts and terminology ISO/IEC 22989*, pero puede entenderse como el conjunto de políticas, procesos, roles y controles que una organización establece para garantizar que sus sistemas de IA se diseñan, desarrollan, despliegan y supervisan de forma responsable.

Hablar de Gobernanza de la IA supone,por tanto,definir un Sistema de Gestión de Inteligencia Artificial (SGIA) e implementarlo, siguiendo el enfoque PDCA (plan, do, check, act), permitiendo una gestión proactiva y demostrable del riesgo, incluyendo los sesgos.

La adopción de un estándar, como por ejemplo la citada ISO 42001, se perfila como la opción óptima por varias razones: su carácter de estándar, lo que supone que haya sido analizada y contrastada; por su validez "global", lo que permite tener en cuenta buenas prácticas aplicables a diferentes países y a contextos multinacionales; y también porque el RIA reconoce y fomenta el uso de herramientas válidas para demostrar cumplimiento, especialmente en sistemas de alto riesgo, que pueden ser adoptadas como base para normas armonizadas europeas si lo aprueba la comisión, lo que pudiera suceder en un futuro este caso, y tal y como ha sucedido en otros precedentes[36]. En esos supuestos el reglamento establece los requisitos legales generales y la norma ISO proporciona un modelo que aterriza para ayudar a cumplir con esos requisitos de forma complementaria. Otro motivo por el que apostar por esta norma ISO es porque no se limita a sistemas de alto riesgo y se puede aplicar a cualquier tipo de operador (principalmente proveedor y responsable del despliegue), sea grande o pequeño y sea cual sea el sector de su actividad. Adicionalmente, al tratarse de una norma certificable permite además demostrar confianza ante posibles clientes y mercado, ciudadanos y sociedad.

Tomando como referencia la ISO/IEC 42001, el Sistema de Gestión de la Inteligencia Artificial (SGIA) puede estructurarse de manera coherente con la lógica de mejora continua propia de los sistemas de gestión conforme al tradicional ciclo PDCA (Plan–Do–Check–Act) y que cada actor tenga su propio SGIA, pero alineado y complementario. Por ejemplo, el proveedor se enfoca en el diseño ético y técnico, mientras que el usuario o responsable del despliegue se centra en el uso responsable y el impacto real en su caso concreto. Por poner un par de ejemplos centrados en la fase de plan: En relación con la obligación de disponer de una política de IA, el proveedor define principios éticos, transparencia, mitigación de sesgos desde el diseño y el responsable del despliegue adapta la política a su contexto de uso. Si hablamos de roles y responsabilidades, el proveedor asigna responsables de diseño ético, gestión de datos, validación de modelos y el responsable del despliegue asigna responsables de supervisión, uso seguro, atención a impactos en personas.

Asimismo, cada organización debe de adaptarlo a su tamaño, actividad etc. y en función de todo lo anterior aplicarán una serie de controles u otros (vid anexo 1 de la norma) que serán reflejados en la declaración de aplicabilidad, e incluso pueden tener que ampliarse. Si hablamos de sesgos vemos algunos ejemplos de controles específicos como el control A.6.3.1 (evaluación de impacto de IA), que promueve identificar impactos en derechos fundamentales desde el diseñ;, o el control A.6.4.1 (gestión de riesgos de IA), que obliga a establecer procesos para detectar y mitigar sesgos. Por último, el hecho de que disponga de una estructura armonizada de alto nivel con otras normas ISO (sin perjuicio de peculiaridades propias de la norma), que además es habitual y posible que

[36] Por ejemplo, en el ámbito sanitario tenemos la ISO 13485: 2016 (adaptada por la UNE-EN ISO 13485:2016) de "Gestión de calidad en productos sanitarios" en el Reglamento (UE) 2017/745 del Parlamento Europeo y del Consejo, de 5 de abril de 2017, sobre los productos sanitarios; que fue reconocida en la decisión de Ejecución (UE) 2021/1182 de la Comisión, de 16 de julio de 2021.

se hayan desplegado en las organizaciones que acometan un proyecto de ISO 42001, permite la integración coherente y eficiente de los diversos sistemas de gestión. Ejemplos claros y "cercanos" son desde la más general ISO 9001:2015 (sobre Sistemas de Gestión de Calidad) a otras más específicas del ámbito TI como por ejemplo la ISO/IEC 27001:2022 (Gestión de la Seguridad de la Información -SGSI- o la ISO/IEC 27701:2025, que es la versión actualizada del Sistema de Gestión de Información de Privacidad (PIMS).

4. Consideraciones finales

A modo de conclusiones realizamos las siguientes consideraciones finales:

1. Es fundamental conocer el concepto de sesgos, diferenciarlo de otros afines y comprender cómo puede introducirse en los sistemas de IA a través de las diferentes etapas de su ciclo de vida. Solo desde ese conocimiento es posible identificar sesgos y riesgos asociados, puesto que este es el primer paso para poder gestionarlos.

2. La legislación y particularmente el RIA proporciona un conjunto de "herramientas" (materializadas como obligaciones para los distintos operadores), que guardan una relación clara con la gestión de los riesgos para los derechos fundamentales. Entre ellas destacan, por especial relevancia, las Evaluaciones de Impacto en Derechos Fundamentales (EIDF), sin perjuicio de otras obligaciones que también contribuyen a mitigar dichos riesgos.

3. De entre dichas obligaciones, adquiere especial relevancia la gobernanza de los datos en el RIA, que aborda cómo se gestionan los datos utilizados por los sistemas de IA de alto riesgo. La calidad de los datos se encuentra estrechamente vinculada a la aparición o amplificación de sesgos, por lo que su adecuada gestión constituye un elemento nuclear para la fiabilidad del sistema.

4. Finalmente, todas estas obligaciones deben de integrarse en un sistema de gestión estructurado, de tal forma que permita planificar, operar, supervisar y mejorar de forma continua los procesos asociados. En este sentido, se ha puesto el acento en el valor de apoyarse en un estándar como la ISO 42001, por los beneficios y las ventajas que ofrece para la ordenación coherente y verificable de la gobernanza de la IA como se ha abordado a lo largo del artículo.

Bibliografía

Agencia Española de Protección de Datos. "Gobernanza y política de protección de datos." *Blog de la AEPD*. 2 de septiembre de 2020. https://www.aepd.es/prensa-y-comunicacion/blog/gobernanza-y-politica-de-proteccion-de-datos

"Inteligencia artificial: principio de exactitud en los tratamientos." *Blog de la AEPD*. 31 de mayo de 2023. https://www.aepd.es/prensa-y-comunicacion/blog/inteligencia-artificial-principio-de-exactitud-en-los-tratamientos

Chaveli Donet, Eduard. "La evaluación de impacto de derechos fundamentales por quienes despliegan sistemas de inteligencia artificial en el Reglamento". En *Tratado sobre el Reglamento de Inteligencia Artificial de la Unión Europea*, dirigido por Lorenzo Cotino Hueso y Pere Simón Castellano, 495–531. Las Rozas (Madrid): Editorial Aranzadi, 2024.

"Sesgos en la IA (I): La necesaria distinción entre sesgos y conceptos afines." *Telefónica Tech Blog*, 11 de febrero de 2025. https://telefonicatech.com/blog/sesgos-en-ia-parte-i-distincion-entre-sesgos-y-conceptos-afines

"Sesgos en la IA (V): Introducción de los riesgos en el ciclo de vida de los sistemas de IA (parte 1)." *Telefónica Tech Blog*. 22 de abril de 2025. https://telefonicatech.com/blog/sesgos-ia-v-introduccion-riesgos-en-ciclo-de-vida-sistemas-de-ia-parte-1

Cotino Hueso, Lorenzo, y Pere Simón Castellano, dirs. *Tratado sobre el reglamento de inteligencia artificial de la Unión Europea*. Madrid: Aranzadi, 2024.

Data Governance Institute. "Definitions of Data Governance." *datagovernance.com*. Consultado el 8 de marzo de 2026. https://datagovernance.com/the-data-governance-basics/definitions-of-data-governance/

ISO/IEC. ISO/IEC 22989:2023. Artificial Intelligence — Artificial Intelligence Concepts and Terminology. Geneva: International Organization for Standardization, 2023.

Loza Corera, María. "Datos y gobernanza de datos y conexiones con principios protección de datos en el artículo 10 del Reglamento". En *Tratado sobre el Reglamento de Inteligencia Artificial de la Unión Europea*, dirigido por Lorenzo Cotino Hueso y Pere Simón Castellano, 567–591. Las Rozas (Madrid): Editorial Aranzadi, 2024.

Mantelero, Alessandro. *Artificial Intelligence and Data Protection: Challenges and Possible Remedies*. Estrasburgo: Consejo de Europa, 2019.

Salvador Serna, Manuel. "Inteligencia artificial y gobernanza de datos en la Administración Pública: sentando las bases para su integración a nivel corporativo". En *Repensando la administración pública: administración digital e innovación pública*, 126–148. Madrid: Instituto Nacional de Administración Pública, 2021

Schwartz, Reva, Apostol Vassilev, Kristen Greene, Lori Perine, Andrew Burt, y Patrick Hall. *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence*. NIST Special Publication 1270. Gaithersburg, MD: National Institute of Standards and Technology, 2022. https://doi.org/10.6028/NIST.SP.1270

Snoek, Suzanne, e Isabel Barberá. *From Inception to Retirement: Addressing Bias Throughout the Lifecycle of AI Systems: A Practical Guide*. Rhite, 2024.