13

Guía 13. Plan de vigilancia poscomercialización

Reglamento Europeo de Inteligencia Artificial

Empresas desarrollando cumplimiento de requisitos

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 aplicables, destacan las siguientes normas. Por un lado, **ISO/IEC WD 25059:2021 "Software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Quality model for AI systems"** para la definición de los indicadores a vigilar en nuestros sistemas inteligentes. Por otro lado, **ISO/IEC 25000:2021 "Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE)"** para diseñar el sistema de vigilancia y definir cómo se realiza dicha labor.

Fecha de revisión: 10 de diciembre de 2025

Contenido general

1. Preámbulo 4
2. Introducción 6
3. Reglamento de Inteligencia Artificial 9
4. ¿Qué elementos implantar y cómo debo hacerlo para desarrollar un adecuado sistema de vigilancia poscomercialización? 12
5. Otros elementos a considerar 23
6. Documentación técnica 25
7. Anexos 26
8. Referencias 33

Índice detallado

1. Preámbulo 4
1.1 Objetivo de este documento 4
1.2 ¿Cómo leer esta guía? 4
1.3 ¿A quién está dirigido? 5
1.4 Casos de uso y ejemplos dispuestos a lo largo de la guía 5
2. Introducción 6
2.1 ¿Qué entendemos por 'sistema de vigilancia poscomercialización' y por qué es necesario? 6
2.2 ¿Por qué se necesita un sistema de vigilancia poscomercialización? 8
2.3 Entidades sujetas a legislación sectorial 8
3. Reglamento de Inteligencia Artificial 9
3.1 Análisis previo y relación de los artículos 9
3.2 Contenido de los artículos en el Reglamento de IA 9
3.3 Correspondencia del articulado con los apartados de la guía 11
4. ¿Qué elementos implantar y cómo debo hacerlo para desarrollar un adecuado sistema de vigilancia poscomercialización? 12
4.1 Vigilancia continua 14
4.2 Vigilancia periódica 16
4.3 Informes de incidentes 18
4.4 Comunicación transparente 19
4.5 Capacitación 20
4.6 Flexibilidad 21
5. Otros elementos a considerar 23
5.1 Conexiones con otras guías 23
6. Documentación técnica 25
7. Anexos 26
7.1 Anexo A - Indicadores de vigilancia 26
7.1.1 Anexo A.I - Indicadores sobre el sistema inteligente 27
7.1.2 Anexo A.II - Indicadores sobre la infraestructura 29
7.1.3 Anexo A.III - Indicadores sobre las acciones de los usuarios 30
7.1.4 Anexo A.IV - Indicadores de seguridad 31
8. Referencias 33

1. Preámbulo

1.1 Objetivo de este documento

El Reglamento Europeo de Inteligencia Artificial (*AI Act*) indica la necesidad de realizar un **plan de vigilancia poscomercialización** para los sistemas de inteligencia artificial de alto riesgo. El objetivo de esta guía es documentar los procesos que deben llevarse a cabo dentro del plan de vigilancia poscomercialización y establecer recomendaciones prácticas para su correcta implementación.

Un plan de vigilancia poscomercialización es un **conjunto de actividades** conducidas por los **proveedores/responsable del despliegue**, para recolectar y evaluar experiencia obtenida de sistemas de inteligencia artificial, considerados de alto riesgo, que han sido **puestos en el mercado**, y así identificar la necesidad de tomar cualquier acción. Se trata de una herramienta importante para asegurar que los sistemas de IA siguen siendo seguros y funcionan correctamente. Además, de esta manera, se contempla el **desarrollo de acciones** en el caso de que el riesgo continuado del sistema de alto riesgo comience a pesar más que el beneficio. La evaluación que se lleva a cabo con esta vigilancia poscomercialización puede contribuir, asimismo, a una continua mejora del sistema en cuestión.

El objetivo de esta guía es presentar los procesos que se deben llevar a cabo en este plan y establecer recomendaciones para conseguirlo.

1.2 ¿Cómo leer esta guía?

Como se mencionaba anteriormente, el **presente documento proporciona medidas** de implementación para entidades proveedoras y usuarias de los sistemas de IA **que faciliten el cumplimiento** de las obligaciones expresadas en el artículo 72 del Reglamento, dedicado a la vigilancia poscomercialización.

Para ello el documento **recorre en orden** todos los apartados de dicho artículo, dando respuesta a las preguntas fundamentales necesarias para **facilitar el cumplimiento** de las obligaciones expresadas en dichos apartados.

Además, debemos tener las siguientes cuestiones en cuenta para una lectura eficiente de esta guía:

1. **Conexión con otras guías**: en el caso de que no se tenga conocimiento o contexto sobre el resto de las guías, se recomienda comenzar por la lectura de la sección 3.1 de la presente guía para comprender la relación con el resto de las guías y comprender los pasos previos a realizar.
2. **Desarrollo del sistema de vigilancia poscomercialización**: sección 4 de la guía donde se explica cómo desarrollar e implementar el sistema de vigilancia además de las medidas que deben de estar contempladas en el plan de vigilancia. Antes de

su lectura, se recomienda revisar el Anexo A para profundizar en el concepto de indicador y los listados de indicadores mínimos.

3. **Documentación técnica**: por último, se recomienda una lectura para obtener una intuición de cuáles son los objetivos por cubrir.

1.3 ¿A quién está dirigido?

Los requisitos descritos en el artículo 72 "*Vigilancia poscomercialización por los proveedores y plan de vigilancia poscomercialización para sistemas IA de alto riesgo*" están enfocados en las medidas que debe de tomar **el proveedor de servicio** una vez el sistema inteligente se encuentra en producción. Por lo tanto, es responsabilidad del proveedor evaluar que durante todo el ciclo de vida se cumplen los requisitos expuestos en dicho artículo.

Las obligaciones del **responsable del despliegue se centran en la notificación de incidentes y comportamientos anómalos al proveedor**. Concretamente, cuando se detecte una modificación anómala del comportamiento del sistema respecto a sus instrucciones de uso, se deberá notificar al proveedor. Además, en el caso de que el usuario no consiga contactar con el proveedor, será éste el responsable de aplicar los cambios necesarios y suspender el uso del sistema tal y como especifica el Artículo 26-Apartado 5 del Reglamento.

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

Para contextualizar, donde aplica, las medidas expuestas que permiten cumplir los requisitos del reglamento, se utilizarán ejemplos sobre dos casos de uso:

- Promoción de empleados
- Gestión de enfermedades crónicas - Bomba de insulina inteligente

Dichos casos de uso se desarrollan en detalle en la Guía práctica y ejemplos para entender el Reglamento de IA.

Los ejemplos sobre dichos casos de uso son expuestos a alto nivel, sin entrar en detalles ni ser exhaustivos, para intentar abarcar las mayores casuísticas posibles. Además, no responden a experiencias reales (pero sí con la intención de ser realistas desde un punto de vista didáctico), teniendo como objetivo únicamente aclarar un poco más las medidas, no pudiendo por tanto ser tomados como especificaciones en una implantación real.

2. Introducción

2.1 ¿Qué entendemos por 'sistema de vigilancia poscomercialización' y por qué es necesario?

Un sistema de vigilancia poscomercialización de sistemas de IA de alto riesgo se concibe como un **conjunto de procesos** y **herramientas** orientados a recabar datos de un sistema para transformarlos en una serie de indicadores sobre su actividad con el objetivo de supervisar los sistemas de inteligencia artificial (IA) **después de su lanzamiento al mercado** (Véase punto 2.5 "Diferencias entre dato e indicador" para comprender la diferencia entre datos e indicadores). El objetivo es que el proveedor pueda evaluar si los sistemas de IA cumplen los requisitos establecidos en el capítulo III, sección 2 (Sección referida a los 'Requisitos de los sistemas de alto riesgo'), a lo largo de todo el ciclo de vida del sistema inteligente.

El sistema de vigilancia poscomercialización funciona mediante los siguientes subsistemas:

- **Sistemas de captación de indicadores.** Diferentes procesos que recopilan datos del rendimiento del sistema inteligente, su infraestructura, las interacciones de usuarios y diferentes datos sobre seguridad (Véase Anexo A de la presente guía: "Indicadores de vigilancia" para ver el listado mínimo de indicadores).
- **Sistemas de registro de dichos indicadores**. Servicios de almacenamiento de dichos registros conforme a las medidas descritas en la sección 5 de la Guía de Registros: "¿Qué elementos debo implantar y cómo debo hacerlo para desarrollar un adecuado sistema de gestión de registros?".
- **Sistema de alertas automatizadas**. Procesos de vigilancia de cambios en los indicadores en base a sus escalas preestablecidas para alertar sobre posibles escenarios de riesgo (Véase Anexo A de la presente guía: "Indicadores de vigilancia" para ver el listado de umbrales mínimos y máximos).
- **Diferentes interfaces de análisis para los encargados de vigilancia**. Punto de acceso y análisis por parte de los vigilantes del sistema para poder analizar los indicadores extraídos. Puede ser en forma de aplicación web, listado de datos en bruto o a través de cualquier otra herramienta que permita realizar búsquedas y operaciones sobre grupos de registros.

Ejemplo – Sistema de evaluación para la promoción de personal

Pongamos el caso de una empresa que ha implementado un sistema de IA para evaluar la promoción del personal, que se basa en una variedad de parámetros y datos recopilados de los empleados. El sistema ha sido validado y aprobado por los reguladores pertinentes y ha sido lanzado al mercado.

Teniendo en cuenta que el sistema de IA establecido es considerado de alto riesgo, la empresa ha implementado un sistema de vigilancia poscomercialización para recopilar y analizar datos continuamente sobre la eficacia del sistema en el mundo real. Dicho sistema de vigilancia recopila los datos de promoción de los empleados, así como otros datos relevantes, como la tasa de rotación y la satisfacción de los empleados, y los analiza para detectar cualquier patrón o tendencia inesperados.

Imaginemos que, después de unos meses de vigilancia poscomercialización, el indicador de uno de los inputs (grupo étnico de los empleados) sufre una variación significativa con respecto a la media de dicho input en el registro histórico. En este caso, el sistema de vigilancia poscomercialización envía una alerta a través de emails desde el sistema de monitorización al equipo responsable del sistema de IA, que investiga el problema en la interfaz de análisis.

El equipo descubre que dicho sistema está utilizando un conjunto de datos históricos que produce que los empleados de cierto origen étnico o género tengan menos probabilidades de ser promovidos en comparación con otros, a pesar de tener calificaciones similares. El incidente se comunica a través del informe de incidentes a los responsables designados en el plan de vigilancia. Posteriormente, el equipo de la empresa trabaja rápidamente para desarrollar una solución y lanza una actualización del sistema de IA que elimina el sesgo. Una vez hecho esto, el sistema de vigilancia poscomercialización continúa recopilando y analizando datos para asegurarse de que el problema se haya resuelto y que el sistema de IA esté funcionando correctamente. Finalmente, los responsables del sistema documentarán lo sucedido en los medios establecidos en el plan de vigilancia adjuntando el informe del incidente, la solución aplicada y los resultados obtenidos.

A modo de resumen introductorio visual, se facilita una infografía que trata de dar una visión general del diseño poscomercialización de los sistemas de IA de alto riesgo:

2.2 ¿Por qué se necesita un sistema de vigilancia poscomercialización?

La vigilancia poscomercialización de sistemas de IA de alto riesgo incluye una variedad de tareas, teniendo en cuenta la finalidad prevista del sistema. Tiene además, en cuenta, la recolección de datos sobre el rendimiento y la seguridad del sistema, la evaluación de las posibles causas de problemas detectados, la implementación de soluciones para corregir problemas, y la comunicación de los resultados y recomendaciones a las partes interesadas.

Puede, además, incluir la implementación de controles de seguridad adicionales para proteger el sistema de ataques maliciosos, y la realización de evaluaciones periódicas para garantizar que el sistema continúa cumpliendo con los distintos requisitos que establece el AI Act para estos sistemas en cuanto a 'Datos y gobernanza de datos', 'Documentación Técnica', 'Registros', 'Supervisión humana', 'Transparencia y comunicación de información a los usuarios' y 'Precisión, solidez y ciberseguridad'. Estos procesos son importantes para garantizar la seguridad y la confiabilidad de los sistemas de IA de alto riesgo una vez implementados.

2.3 Entidades sujetas a legislación sectorial

Los proveedores sujetos por los actos legislativos especificados en la Sección A del Anexo I, que hayan implementado un sistema y un plan de vigilancia poscomercialización conforme a estas disposiciones, tienen la facultad de utilizar dichos mecanismos acorde con la normativa establecida.

Esta posibilidad se extiende igualmente a aquellos proveedores que desarrollan sistemas de inteligencia artificial (IA) considerados de alto riesgo, según se define en el punto 5 del Anexo III. Es decir, sistemas introducidos en el mercado o puestos en servicio por entidades financieras sujetas a requisitos en materia de gobernanza interna, mecanismos o procesos establecidos con arreglo a la legislación de la Unión en materia de servicios financieros.

3. Reglamento de Inteligencia Artificial

3.1 Análisis previo y relación de los artículos

En esta sección se simplifica y estructura el contenido del artículo con el objetivo de facilitar su comprensión. Asimismo, se presenta la estructura seguida en esta guía para dar cobertura al contenido de dicho artículo.

Qué entendemos del artículo

El artículo aborda los factores exigidos para la **implementación y documentación** de sistemas de vigilancia poscomercialización para sistemas de IA de alto riesgo. Con ello, pretende garantizar que dichos sistemas continúen **cumpliendo con los requisitos necesarios** establecidos en el AI Act una vez puestos en el mercado, y durante **todo el ciclo de vida** del sistema.

Qué medidas clave entendemos que establece el artículo

- Establecimiento de sistemas de vigilancia proporcionales a los riesgos.
- Recopilación, documentación y análisis de datos relevantes sobre el rendimiento de los sistemas de IA.
- Creación de un plan de vigilancia poscomercialización como parte de la documentación técnica.
- Aceptación de documentación de vigilancia existente si esta cumple con ciertos actos legislativos y regulaciones financieras.

3.2 Contenido de los artículos en el Reglamento de IA

AI Act

Art.72 Vigilancia poscomercialización por parte de los proveedores y plan de vigilancia poscomercialización para sistemas de IA de alto riesgo

1. Los proveedores establecerán y documentarán un sistema de vigilancia poscomercialización de forma proporcionada a la naturaleza de las tecnologías de IA y a los riesgos de los sistemas de IA de alto riesgo.

2. El sistema de vigilancia poscomercialización recopilará, documentará y analizará de manera activa y sistemática los datos pertinentes que pueden facilitar los responsables del despliegue o que pueden recopilarse a través de otras fuentes sobre el funcionamiento de los sistemas de IA de alto riesgo durante toda su vida útil, y que permiten al proveedor evaluar el cumplimiento permanente de los requisitos establecidos en el capítulo III, sección 2, por parte de los sistemas de IA. Cuando proceda, la vigilancia poscomercialización incluirá un análisis de la interacción con otros sistemas de IA. Esta obligación no comprenderá los datos operativos sensibles de los responsables del despliegue que sean autoridades garantes del cumplimiento del Derecho.

3. El sistema de vigilancia poscomercialización se basará en un plan de vigilancia poscomercialización. El plan de vigilancia poscomercialización formará parte de la documentación técnica a que se refiere el anexo IV. La Comisión adoptará un acto de ejecución en el que se establecerán disposiciones detalladas que constituyan un modelo para el plan de vigilancia poscomercialización y la lista de elementos que deberán incluirse en él a más tardar el 2 de febrero de 2026. Dicho acto de ejecución se adoptará de conformidad con el procedimiento de examen a que se refiere el artículo 98, apartado 2.

4. En el caso de los sistemas de IA de alto riesgo regulados por los actos legislativos de armonización de la Unión enumerados en el anexo I, sección A, cuando ya se haya establecido un sistema y un plan de vigilancia poscomercialización con arreglo a dichos actos, con el fin de garantizar la coherencia, evitar duplicidades y reducir al mínimo las cargas adicionales, los proveedores podrán optar por integrar, según proceda, los elementos necesarios descritos en los apartados 1, 2 y 3, utilizando el modelo a que se refiere el apartado 3, en los sistemas y planes que ya existan en virtud de dicha legislación, siempre que alcance un nivel de protección equivalente.

El párrafo primero del presente apartado también se aplicará a los sistemas de IA de alto riesgo a que se refiere el anexo III, punto 5, introducidos en el mercado o puestos en servicio por entidades financieras sujetas a requisitos relativos a su gobernanza, sus sistemas o sus procesos internos en virtud del Derecho de la Unión en materia de servicios financieros.

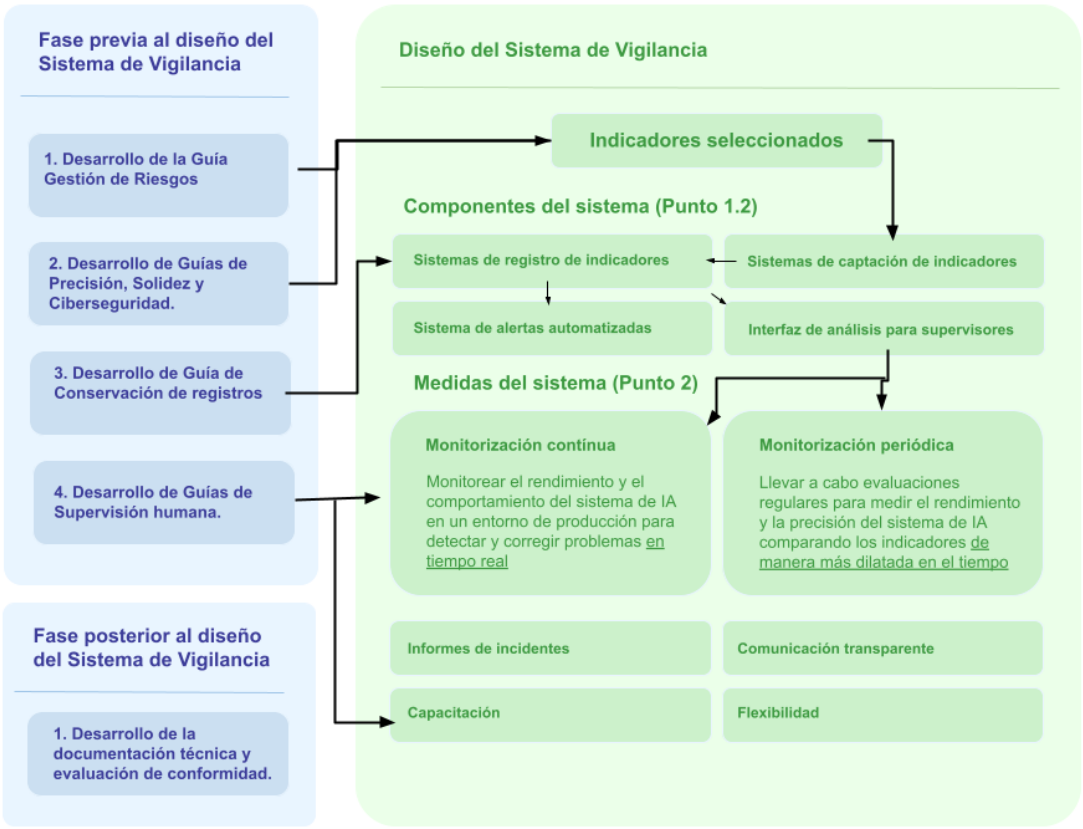
3.3 Correspondencia del articulado con los apartados de la guía

En esta tabla se detalla la correspondencia de las secciones de esta guía que abordan los elementos de dicho artículo:

Artículo Reglamento	Requerimiento Reglamento	Sección guía
72.1	Proporcionalidad del sistema de vigilancia poscomercialización en base a la naturaleza de las tecnologías de IA aplicadas.	Apartado 4
72.2	Datos facilitados por los responsables del despliegue o recopilados a través del sistema de vigilancia sobre el funcionamiento de los sistemas de IA de alto riesgo durante toda su vida útil.	Apartado 4.1 y Apartado 4.2
72.3	El Plan de vigilancia poscomercialización formará parte de la documentación técnica al que se refiere el anexo IV.	Apartado 6
72.4	Documentación específica de sistemas de IA de alto riesgo regulados por los actos legislativos de armonización de la Unión enumerados en el anexo I, sección A.	Apartado 6

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

En base a lo comentado en el punto 1 acerca de qué es un sistema de vigilancia poscomercialización y qué componentes debe de tener, podemos seguir el siguiente diagrama de flujo en su diseño:



¿Cómo implementar el sistema de vigilancia poscomercialización?

Para poder llevar a cabo las principales actividades de vigilancia, el proveedor deberá diseñar, desplegar y validar los componentes que forman parte del sistema. En concreto:

1. **Selección de indicadores**: se deberá establecer cuáles son los indicadores más relevantes para la correcta vigilancia del sistema (véase Anexo A de la presente guía: "Indicadores de vigilancia").
   La selección de estos indicadores deberá basarse en el sistema de gestión de riesgos previamente desarrollado.
   Una vez seleccionados, el sistema inteligente deberá generar registros con la información asociada a dichos indicadores, basándose en las recomendaciones técnicas de la Guía de Registros: "¿Qué elementos debo implantar y cómo debo hacerlo para desarrollar un adecuado sistema de gestión de registros?".
   Tras concretar el diseño de los registros, el sistema inteligente deberá depositarlos en un archivo de log temporal o similar para que el sistema de captación comience su procesamiento.
   También puede darse el caso de que el propio sistema inteligente envíe directamente los registros al sistema de registro de indicadores, sin necesidad de utilizar un archivo temporal. Ambas opciones son válidas y dependerán de la arquitectura del sistema inteligente.
2. **Desarrollo de sistemas de captación y envío de indicadores:** Los sistemas de captación recolectarán los datos de los archivos logs donde sean depositados por parte del sistema inteligente y serán enviados a los sistemas de registro de indicadores. Estos sistemas deberán prestar especial atención a la seguridad en el envío de registros.
3. **Sistema de registro de indicadores**: Se deberá de implementar un sistema de gestión de registros donde los sistemas de captación de indicadores enviarán los datos generados por el sistema. Este sistema podrá ser cualquier tipo de base de datos estructurada o no estructurada de almacenamiento de registros. Además, se tendrán que establecer las políticas de control de acceso, retención y eliminación de dichos registros tal como indica la Guía de Registros.
4. **Desarrollo conjunto del sistema de alertas y la interfaz de análisis para supervisores**: A partir de los registros se contará con un sistema de vigilancia de indicadores para la alerta temprana frente a valores anómalos. Se podrá desarrollar una solución propia o emplear alguna de las soluciones de visualización y vigilancia de registros. Se deberá de contar tanto con:
   a. Un sistema de alerta asignado a los supervisores del sistema. Se recomienda que el sistema envíe las alertas por diferentes medios de comunicación y notifique también periódicamente sobre el comportamiento adecuado del sistema.
   b. Una interfaz para la supervisión en tiempo real de los indicadores del sistema. Dicha interfaz deberá mostrar de forma evidente cuándo uno de los indicadores del sistema se encuentra fuera de la escala esperada. También deberá permitir la exploración manual de los registros para realizar una inspección más profunda sobre los datos recabados.

Todos los componentes del sistema deben de tener tests de funcionalidad a nivel de unidad y de integración con el resto para asegurar el correcto funcionamiento del sistema de vigilancia.

Antes de llevar a cabo el diseño y la implementación de dicho sistema de vigilancia, se debe de establecer un plan de vigilancia poscomercialización en el que se incluyan las siguientes labores y características: vigilancia continua, vigilancia periódica, informe de incidentes, comunicación transparente, capacitación, flexibilidad y evaluación independiente (si es posible). A continuación, se describen los elementos enumerados:

- **Vigilancia continua:** monitorear el rendimiento y el comportamiento del sistema de IA en un entorno de producción para detectar y corregir problemas a partir de los indicadores seleccionados en la evaluación de riesgos.
- **Vigilancia periódica:** llevar a cabo evaluaciones regulares para medir el rendimiento y la precisión del sistema de IA y detectar cualquier problema.
- **Informes de incidentes:** establecer un sistema para recopilar y analizar informes de incidentes relacionados con el sistema de IA, incluidos errores, problemas relativos a la privacidad y la seguridad.
- **Comunicación transparente:** proporcionar, a través de los informes, información clara y transparente sobre el rendimiento y la seguridad del sistema de IA a los usuarios, reguladores y otros interesados.
- **Capacitación:** capacitar a los usuarios y operadores del sistema de IA para que puedan detectar y manejar problemas asociados al funcionamiento anómalo del sistema.
- **Flexibilidad:** tener un plan flexible y escalable para adaptarse a los cambios en el rendimiento y la seguridad del sistema de IA y para cumplir con las regulaciones actuales y futuras.

En el caso de que fuera posible, sería valorable una **evaluación independiente**, que pudiera proporcionar una visión objetiva sobre su rendimiento y seguridad. Por ejemplo, la consecución de un sello o certificación que dé garantía de ello por parte de una entidad acreditada.

Las siguientes secciones desarrollan cada uno de estos elementos en mayor profundidad. Posteriormente, en la sección 4 de "Documentación técnica" se indicará la manera en la que documentar tanto el sistema de vigilancia como el plan de vigilancia propuesto.

4.1 Vigilancia continua

La vigilancia continua de los sistemas de IA de alto riesgo es necesaria para garantizar que el sistema continúe funcionando de manera segura y eficaz una vez en el mercado, y para evitar problemas de rendimiento, seguridad y responsabilidad legal. Esta vigilancia resulta especialmente relevante ante **cambios abruptos en el comportamiento del sistema**.

Hay que tener en cuenta que los sistemas de IA operan en entornos cambiantes, y pueden verse afectados por cambios en los datos de entrada, las condiciones del entorno, y las

regulaciones. La vigilancia continua de los indicadores seleccionados tras la evaluación de riesgos permite detectar y corregir problemas relacionados con estos cambios.

Además, los sistemas de IA pueden experimentar problemas de rendimiento debido a una variedad de factores, como el envejecimiento de los datos de entrenamiento, el uso de datos de entrenamiento insuficientes o inexactos, o la falta de capacitación adecuada. La vigilancia continua permite detectar y corregir problemas de rendimiento antes de que afecten significativamente el sistema.

No se debe olvidar que estos sistemas pueden ser objeto de ataques maliciosos, como el aprendizaje adversarial, el phishing, y el robo de datos. La **vigilancia continua permite detectar y corregir problemas de seguridad antes de que causen daños significativos**.

Medidas para llevarlo a cabo

Entre las técnicas existentes para monitorizar el rendimiento y el comportamiento de un sistema de IA en un entorno de producción se destacan las siguientes:

- **Vigilar los indicadores del sistema inteligente:** consiste en la recopilación y análisis de datos sobre el rendimiento y el comportamiento del sistema de IA para detectar patrones y tendencias. Se pueden utilizar herramientas de análisis de datos para detectar problemas, como errores y desviaciones.
- **Vigilar los indicadores sobre las acciones de los usuarios:** recopilación de información sobre cómo los usuarios interactúan con el sistema de IA, incluidos los comandos, las consultas y las respuestas. Esto puede ayudar a detectar problemas de usabilidad y a proporcionar retroalimentación para mejorar el sistema.
- **Vigilar los indicadores de la infraestructura:** monitorizar el estado de los componentes del sistema de IA, como la CPU, la memoria y el almacenamiento, para detectar problemas de rendimiento y capacidad.
- **Vigilar los indicadores de seguridad:** El objetivo es detectar y prevenir brechas en la seguridad, incluidos ataques de phishing, fuerza bruta y malwares.
- **Supervisión de cambios en los indicadores mediante alertas:** configuración de alertas para detectar eventos anómalos, como errores, violaciones de seguridad y problemas de privacidad. Esto puede incluir la configuración de reglas para detectar patrones específicos o la configuración de umbrales para detectar desviaciones.

Ejemplo – Bomba de insulina inteligente

Usando el ejemplo de la bomba de insulina inteligente, la monitorización continua se centraría en la supervisión de los indicadores seleccionados para **evitar cambios de rendimiento abruptos.**

Supongamos que el indicador de número de predicciones, que de forma continuada tenía valores cercanos a 45 predicciones por minuto, ha pasado a ser 3640 predicciones por minuto.

Primero, los registros con los indicadores serán transferidos por el sistema de captación y envío de indicadores al sistema de gestión de registros a través de un protocolo de comunicación seguro. Posteriormente, el sistema de alertas analizará los datos obtenidos y detectará la anomalía respecto a la escala de normalidad del indicador en base a las reglas preestablecidas en su diseño. Por último, el sistema de alertas lanzará las notificaciones a los supervisores a través de los medios definidos en su diseño: emails, SMS o cualquier otro sistema de comunicación de la organización.

Tras la notificación el equipo responsable de supervisión evaluará a través de la interfaz de supervisión los indicadores captados y procederá a la revisión del sistema.

4.2 Vigilancia periódica

La realización de evaluaciones regulares para medir el rendimiento y la precisión del sistema de inteligencia artificial es esencial para garantizar que este sigue siendo preciso, escalable y útil en un entorno real, una **revisión más dilatada en el tiempo,** dependiendo dicha periodicidad de las características del propio sistema, permitiendo detectar problemas temprano y tomar medidas para corregirlos.

Por un lado, es importante asegurar la precisión del sistema: a través de pruebas de precisión, se puede evaluar cómo el sistema está realizando tareas específicas, como el reconocimiento de objetos en imágenes o la traducción de idiomas, y detectar si hay algún problema que esté afectando su precisión.

Además, también es necesario identificar problemas con antelación: esto se puede conseguir realizando pruebas de seguimiento, se puede monitorear el rendimiento del sistema en el tiempo, detectando cualquier disminución en la precisión o el rendimiento, lo cual permite detectar con antelación posibles problemas y tomar medidas para corregirlos antes de que se vuelvan graves.

También permite asegurar la escalabilidad del sistema: a través de pruebas de rendimiento, se puede evaluar cómo el sistema está manejando grandes cantidades de datos y detectar cualquier problema que pueda afectar su capacidad para escalar.

Ejemplo – Sistema de evaluación para la promoción de personal

Empleando el ejemplo del sistema de promoción de personal, el sistema inteligente podría perder rendimiento al no reentrenar el sistema durante años y tomar decisiones de promoción que terminan en despidos o equipos poco cohesionados. La monitorización periódica de las métricas de rendimiento permitirá evaluar su utilidad y prevenir este tipo de escenarios.

Medidas para llevarlo a cabo

Existen varias técnicas para evaluar el rendimiento y la precisión de un sistema de IA, algunas de las cuales incluyen:

- **Pruebas de rendimiento:** se utilizan para medir el tiempo de respuesta del sistema y su capacidad para manejar grandes cantidades de datos. Son las conocidas pruebas de estrés que permiten llevar al límite el sistema inteligente en un entorno de prueba y valorar la respuesta ofrecida ante dicho escenario.
- **Pruebas de precisión:** se utilizan para medir la precisión del sistema al realizar tareas específicas, como el reconocimiento de objetos en imágenes o la traducción de idiomas. En concreto, se deberá seleccionar previamente las métricas de precisión asociadas al modelo de IA (sección de la Guía de Precisión: "Métricas de precisión asociadas al modelo de IA") y los niveles de precisión pertinentes para el sistema (sección de la Guía de Precisión: "Niveles de Precisión, Documentación y Monitorización de los modelos de IA.").

No obstante, en algunos casos, medir la precisión del algoritmo puede ser muy complicado debido al caso de uso en sí, como puede verse en el siguiente ejemplo.

Ejemplo – Sistema de evaluación para la promoción de personal

Siguiendo con el ejemplo propuesto del sistema inteligente de promoción, ¿Cómo podríamos evaluar si el sistema inteligente ha acertado o no en su predicción? ¿Cómo podemos valorar que ha acertado al no promocionar a un empleado en particular? En estos casos se recomienda mantener una muestra control que siga el procedimiento anterior a la implantación del sistema para poder comparar los resultados obtenidos con los que hubiera ofrecido el sistema. Para ello, se dejaría una muestra de los empleados para ser evaluada y promocionada de forma manual y tras dicho proceso se evaluaría las decisiones que hubiera tomado el modelo en base a las métricas anteriores frente a las decisiones manuales obtenidas por el equipo de RRHH.

Además de las pruebas anteriores hay que tener en cuenta las siguientes recomendaciones:

- **Histórico de resultados:** Tal y como indica la Guía de Precisión en su sección "Medidas técnicas Generales", se deberá de contar con un historial de los resultados obtenidos en las pruebas anteriores para descubrir tendencias de reducción de rendimiento de forma temprana. A modo de ejemplo, si los encargados del sistema de promoción realizarán el primer mes de despliegue en producción de un sistema inteligente diferentes pruebas obteniendo una precisión del 99.8%, 99.6% y 99.2% de forma sucesiva, si el criterio mínimo de precisión fuera del 98%, esta tendencia de pérdida de rendimiento podría pasar desapercibida. Sin embargo, al contar con un historial de pruebas se podrá observar la tendencia en los resultados.
- **Especial cuidado con la mejora en las métricas sin variación del sistema:** en algunos casos puede que las decisiones tomadas por los resultados del sistema inteligente produzcan una tendencia a la mejora de las métricas del sistema buscando de forma indirecta las entradas que se adaptan a la mejora del rendimiento del algoritmo.

Ejemplo – Sistema de evaluación para la promoción de personal

Siguiendo con nuestro ejemplo, si el sistema de selección de candidatos de promoción se reentrenara sobre datos sobre los que ha predicho, se estaría sesgando a sí mismo. Es decir, si el algoritmo da una mayor probabilidad a empleados con unas determinadas características, estos terminan siendo ascendidos y el sistema es de nuevo entrenado sobre los resultados obtenidos, irá progresivamente sesgando hacia las decisiones que tomó de forma inicial acrecentando su seguridad. Este escenario puede corregirse a través del entrenamiento sobre una muestra de control que ha sido promocionada a través de expertos en recursos humanos.

4.3 Informes de incidentes

A la hora de establecer un sistema de recopilación y análisis de informes de incidentes relacionados con el sistema de IA ofrecemos las siguientes recomendaciones:

Definir los objetivos del sistema: determinar los objetivos del sistema de recopilación y análisis de incidentes. Es decir, responder a la siguiente pregunta: ¿Qué información se espera recopilar y cómo se utilizará esta información?

Diseño del sistema: definición del proceso de recopilación y análisis de los informes de incidentes. Esto incluye establecer los procesos para reportar incidentes, la estructura de los informes y los mecanismos de análisis de datos.

Implementación del sistema: Incluye el desarrollo de una plataforma para recopilar y almacenar los informes, así como crear herramientas de análisis de datos.

Capacitación: Se debe trabajar en la formación de las personas involucradas en el sistema, incluyendo a los usuarios finales, los administradores del sistema y el equipo de análisis de datos.

Vigilar continuamente: Esta vigilancia incluye: recolectar retroalimentación de los usuarios, revisar los informes de incidentes y analizar las tendencias para identificar posibles problemas y oportunidades de mejora.

Ejemplo – Bomba de insulina inteligente

Tras la detección de un comportamiento anómalo en la bomba de insulina inteligente se ha generado un informe con la siguiente información:

- **Características y objetivos del sistema**: una descripción del sistema en cuestión para facilitar la evaluación de impacto por parte de personas sin conocimientos relacionados con el sistema. Esta información debe de estar previamente cumplimentada en el informe.
- **Comportamiento anómalo detectado:** en este caso el indicador anómalo ha sido el número de predicciones por minuto de la bomba. También se describe información del contexto como el número de sistemas afectados, inicio de la anomalía, impacto producido y demás factores relacionados que puedan ayudar a la comprensión de la situación.
- **Acciones desarrolladas hasta el momento**: se ha revisado la interfaz de monitorización y los registros de indicadores. El problema se ha producido en un solo dispositivo y se requiere de una revisión in situ.
- Otras secciones que puedan ofrecer más contexto de la anomalía detectada en el sistema inteligente:
    - Configuración y Ajustes del Sistema en el momento en el que se produjo el comportamiento anómalo.
    - Extracto del histórico de registros e indicadores del sistema inteligente.
    - Notas de los supervisores que puedan añadir contexto a la situación.

Una vez completado el informe se deberá de trasladar a través del sistema de envío predefinido.

4.4 Comunicación transparente

La comunicación de las características del sistema, el rendimiento de este y las consecuencias de su uso en producción deben adaptarse al receptor de dicha información para facilitar una comprensión correcta de todas las implicaciones de su empleo.

Medidas para llevarlo a cabo

Es importante proporcionar información clara y transparente sobre el rendimiento y la seguridad del sistema de IA a los responsables del despliegue, reguladores y otros interesados. Para ello se recomienda establecer indicadores de rendimiento y seguridad claros y cuantificables para medir el rendimiento y la seguridad del sistema de IA. Estos indicadores deben ser relevantes para los responsables del despliegue, reguladores y otros interesados.

Por otro lado, también es relevante la recopilación de datos sobre el rendimiento y la seguridad del sistema de IA, así como su posterior análisis, con el fin de obtener información sobre su desempeño.

Como medida destacada en este ámbito, resulta fundamental ofrecer información clara sobre el rendimiento y la seguridad del sistema de IA a los responsables del despliegue, los reguladores y otros interesados. Esto puede incluir informes periódicos, paneles de seguimiento o un sitio web dedicado.

En este sentido, es recomendable proporcionar detalles sobre el funcionamiento del sistema, incluyendo su algoritmo, los datos que utiliza y las decisiones que toma. Esto ayudará a los distintos interesados a comprender mejor el sistema y a evaluar su rendimiento y su seguridad.

Ejemplo – Sistema de evaluación para la promoción de personal

Un ejemplo de estas medidas, empleando el ejemplo del sistema de promoción de empleados, sería la adaptación en la forma de comunicar las métricas de precisión por parte de los supervisores del sistema a la hora de ser reportada a los decisores de este. En este caso, en lugar de mencionar que:

"El sistema inteligente cuenta con un 96% de precisión"

se podría digerir el mensaje para indicar que:

"El sistema inteligente cuenta con un 96% de precisión, es decir, de cada 100 empleados que el sistema indicó que deberían de ser promocionados el sistema acertó en 96 de ellos y cometió error en indicar una promoción para 4 de ellos. Sin embargo, esta métrica no considera los errores que se produjeron entre los empleados que no fueron promocionados y se deben de considerar otras métricas junto con la actual."

4.5 Capacitación

A la hora de capacitar a los supervisores se debe proporcionar una formación básica sobre cómo funciona el sistema de IA y cómo se utiliza. Esto incluye información sobre los algoritmos utilizados, los datos que se utilizan y cómo se toman decisiones. También es importante proporcionar ejemplos de funcionamiento anómalo, incluyendo ejemplos de errores, fallos y comportamiento inesperado.

Ejemplo – Sistema de evaluación para la promoción de personal

En el caso del ejemplo, el proveedor del sistema inteligente de selección de personal para promociones internas deberá contar con supervisores con los siguientes requisitos:

- Formación acerca del contexto en el que se está empleando el sistema y las consecuencias devenidas de las predicciones generadas para los empleados.
- Conocer los datos que está teniendo en consideración para llevar a cabo las predicciones (experiencia del empleado, proyectos realizados, disponibilidad de traslado entre otras) y cómo pueden verse afectadas por factores externos (latencia en la introducción de nuevas actualizaciones en los CV de los empleados por parte del equipo de recursos humanos).

  Comprensión acerca del algoritmo o algoritmos empleados en el sistema y qué sistema emplea para la generalización de conocimiento sobre los empleados. A modo de ejemplo, conocer que se trata de un algoritmo basado en reglas ayudará.
- Experiencia a través de ejemplos audiovisuales o en escenarios prácticos de comportamientos anómalos del sistema inteligente en el que no genere una predicción adecuada para los empleados o no permita realizar predicciones por una sobrecarga de la infraestructura y cómo deberían de actuar en tal caso.

Por la parte de los propios usuarios del sistema, los especialistas de recursos humanos, también se deberían de formar en la detección de este tipo de escenarios y las medidas adecuadas para reportar cualquier situación anómala al proveedor del servicio.

4.6 Flexibilidad

Mantener un plan flexible y escalable es fundamental para mejorar la vigilancia del sistema. Por un lado, se encuentra el hecho de que los sistemas de IA están en constante evolución y mejora, por lo que esta característica puede permitirle adoptar nuevas tecnologías y técnicas para mejorar el rendimiento. Igualmente, en cuestiones de seguridad, la flexibilidad permite adaptarse a nuevas amenazas y vulnerabilidades, así como garantizar la protección de los datos y la privacidad de los usuarios, y adaptarse a nuevos cambios en la regulación.

En definitiva, la flexibilidad es la adaptación del sistema inteligente a los cambios internos y externos que puedan afectar a su funcionamiento.

Medidas para llevarlo a cabo

a) **Identificar las regulaciones aplicables:** Incluye las regulaciones sectoriales y las regulaciones de protección de datos.

b) **Evaluar el rendimiento y la seguridad del sistema:** se recomienda la utilización de indicadores de rendimiento y seguridad claros y cuantificables.
c) **Identificar los riesgos:** Incluye los riesgos de rendimiento y los riesgos de seguridad.
d) **Establecer procedimientos para prevenir los riesgos:** Incluye procedimientos para comunicar incidentes, investigar incidentes e implementar soluciones.
e) **Vigilar el cumplimiento con la regulación existente:** Incluye el cumplimiento de las regulaciones sectoriales y las regulaciones de protección de datos.
f) **Establecer un plan de contingencia:** Esta cuestión es importante con el objetivo de prevenir incidentes o fallos críticos en el sistema de IA.
g) **Implementar un sistema de revisión continua:** Su objetivo es la evaluación del rendimiento y la seguridad del sistema de IA, identificar problemas y tomar medidas para mejorarlo.
h) **Formación específica sobre las regulaciones existentes** en IA, y protección de datos, además de '**mejores prácticas**' (Ética en IA).

Ejemplo – Sistema de evaluación para la promoción de personal

La empresa proveedora del sistema inteligente de promoción de empleados tiene un cambio en su dirección y en su política de incorporación y promoción de empleados, tras producirse una fusión con otra empresa cuyos datos no han sido incorporados al sistema. El proveedor deberá adaptar el plan de vigilancia y el sistema de vigilancia a los nuevos cambios:

- Evaluar si tras la incorporación existen nuevas regulaciones aplicables.
- Revisar que los indicadores sigan siendo los adecuados para una vigilancia adecuada en base a la nueva evaluación de riesgos.
- Llevar a cabo las actividades formativas necesarias en el caso de que se incorporen al equipo de supervisión miembros de la nueva empresa.
- Revisar los procedimientos de notificación si es necesario incluir nuevos decisores en la cadena de comunicación.
- Evaluar si la infraestructura sobre la que está alojada el sistema de vigilancia debe de ser actualizada con nuevas conexiones.
- Todas aquellas medidas destinadas a mantener activo el sistema y el plan de vigilancia.

5. Otros elementos a considerar

5.1 Conexiones con otras guías

Dado el carácter transversal de los sistemas de vigilancia poscomercialización resulta imprescindible considerar las medidas de vigilancia consideradas en el resto de las guías.

En concreto:

- **Guía de Gestión de riesgos:** Esta guía es la primera que debemos de considerar en nuestra implantación y por tanto la primera que debemos de abordar. Su relación con la Guía de vigilancia se deriva en tres vertientes:
    - Diseño: Tras realizar la evaluación de riesgos estaremos en disposición de definir qué información y datos deberemos de recabar del sistema para diseñar nuestros indicadores del sistema de vigilancia poscomercialización (Véase la sección correspondiente a: "¿Qué elementos debo implantar y cómo debo hacerlo para desarrollar un adecuado sistema de gestión de riesgos?" de la Guía de Gestión de Riesgos).
    - Supervisión: El sistema de vigilancia poscomercialización deberá de ofrecernos indicadores para saber si nuestro sistema se está acercando a una situación de riesgo.
    - Retroalimentación: La actividad del propio sistema de vigilancia poscomercialización ofrecerá *feedback* acerca de cuándo se deberá de actualizar la evaluación de riesgos desarrollada en base a cambios en los indicadores seleccionados. (Artículo 9.2.c)
- **Guía de Conservación de registros:** La relación se encuentra en la Guía de Conservación de registros, donde se indica que deberemos definir la información y los datos que necesitaremos recabar del sistema. Tras esta definición, pasaremos a diseñar los registros necesarios. Por lo tanto, la presente Guía de vigilancia nos facilita cuáles son los indicadores mínimos que debemos de recabar del sistema (Véase punto 5 de la presente guía: "Indicadores de vigilancia") y la Guía de Conservación de Registros nos indica cómo generar dichos registros a nivel técnico (Véase sección de la Guía de Conservación de Registros: "¿Qué elementos debo implantar y cómo debo hacerlo para desarrollar un adecuado sistema de gestión de registros?").
- **Guía de Supervisión humana:** La relación se encuentra dentro de las medidas aplicables de vigilancia humana. En concreto, las medidas de la Guía de Supervisión humana: "Medidas aplicables" deberán de ser completadas para, por ejemplo, contar con los usuarios encargados de la vigilancia del sistema.
- **Guía de Precisión:** Se hace referencia a la necesidad de evaluar la degradación del modelo a través de la vigilancia con *dashboards* y herramientas de visualización de la precisión. Para ello se deberá de seleccionar previamente las métricas de precisión asociadas al modelo de IA (sección de la Guía de Precisión: "Métricas de precisión asociadas al modelo de IA") y los niveles de precisión pertinentes para el

sistema (sección de la Guía de Precisión: "Niveles de Precisión, Documentación y Monitorización de los modelos de IA").

- **Guía de Solidez:** El desarrollo de esta guía debe de ser previo a la Vigilancia poscomercialización dado que en ella se especificarán:
    - Las métricas seleccionadas para la supervisión de solidez del sistema que deberán de ser vigiladas (sección de la Guía de Solidez: "Selección de métricas").
    - Los métodos de validación y verificación (sección de la Guía de Solidez: "Validación y verificación").
    - Los aspectos relacionados con la vigilancia de indicadores de solidez (sección de la Guía de Solidez: "Monitorización de la solidez").
- **Guía de Ciberseguridad:** Se indica que el proveedor del sistema inteligente deberá de aplicar las medidas de ciberseguridad de dicha Guía durante todo el ciclo de vida del sistema. Por esta razón, se deberán de establecer medidas de vigilancia para los aspectos descritos por los epígrafes de la guía de ciberseguridad en forma de indicadores (Véase Anexo A.IV "Indicadores de seguridad" de la presente guía) o a través de la revisión de registros de forma manual o automatizada.
- **Guía de evaluación de conformidad:** El proceso de evaluación de conformidad se extiende durante todo el ciclo de vida del sistema inteligente para asegurar que dichos criterios se mantienen en el tiempo. En particular, el proveedor deberá verificar que el sistema de vigilancia poscomercialización es coherente con la documentación técnica para poder completar la evaluación de conformidad.
- **Guía de Documentación técnica:** En el apartado "Sistema de evaluación post comercialización" de la guía de Documentación técnica se indica que se deberá generar:
    - Documentación detallada del sistema de vigilancia posterior a la comercialización.
    - Documentación sobre las medidas tomadas para vigilar y cubrir los riesgos detectados deberán de ser documentadas adecuadamente.

6. Documentación técnica

De acuerdo con el Reglamento de Inteligencia Artificial, y con lo dispuesto en su Anexo IV y en el artículo 9, la documentación técnica deberá incorporar una descripción detallada del sistema de vigilancia poscomercialización y del plan de vigilancia.

Documentación detallada sobre el sistema de vigilancia. Deberá de incluir al menos los siguientes aspectos:

a. Indicadores seleccionados: dato del cual es obtenido, escala de normalidad y riesgo monitorizado asociado en su caso.
b. Sistemas de captación y envío: marco o tecnología empleada para la captación y envío, correspondencia de indicadores con sistema de captación y envío y periodicidad de los envíos.
c. Registro de los indicadores: sistema de registro seleccionado, formato de registro y tiempos de almacenamiento.
d. Sistema de alertas y la interfaz de análisis para supervisores: listado con todas las alertas definidas incluyendo el indicador o indicadores sobre los que basan su disparador y descripción de la interfaz de análisis implementada.

2. **Documentación detallada sobre las medidas del plan de vigilancia poscomercialización**. Deberá de incluir al menos:
   a. Listado de acciones desarrolladas en la vigilancia continua del sistema: descripción de la tarea, objetivo de esta, periodicidad de ejecución, responsable vinculado y método de registro/comunicación del resultado.
   b. Listado de acciones desarrolladas en la vigilancia periódica del sistema: descripción de la tarea, objetivo de esta, periodicidad de ejecución, responsable vinculado y método de registro/comunicación del resultado.
   c. Borrador del informe de incidentes: incluir una copia de dicho borrador.
   d. Listado de actividades formativas de los involucrados en el plan de vigilancia: fecha, descripción de la actividad, objetivo, duración, roles involucrados y método de registro/comunicación del resultado.

7. Anexos

7.1 Anexo A - Indicadores de vigilancia

La definición de indicador depende en gran medida del contexto de su aplicación. No obstante, en el ámbito de un sistema de vigilancia, un indicador se puede considerar como un dato dentro de una escala que permite medir su impacto respecto a una o varias consecuencias concretas. A modo de ejemplo, nuestro sistema inteligente responsable de administrar la insulina al usuario registra un **dato** del nivel actual de azúcar en sangre: 115 mg/dL. ¿Cómo se transforma dicho dato en un indicador? En este caso, para ser un **indicador**, se debe conocer:

1. La escala del dato: el nivel de azúcar mínimo (70 mg/dL) y máximo (99 mg/dL) aceptable.
2. El impacto de dicho dato: evaluar qué implicaciones tiene la medición actual (115 mg/dL) en las consecuencias o escenarios que queremos controlar: a partir de 99 mg/dL el nivel de azúcar es inusualmente alto y se debe de realizar una evaluación del estado del usuario en profundidad.

Otro ejemplo relacionado con el mismo sistema sería el indicador de número de predicciones. El **dato** sería el número de predicciones generadas por el sistema durante el último minuto: 128 predicciones. Para transformarlo en **indicador** debemos de establecer:

1. La escala del dato: se espera un mínimo de 90 predicciones y un máximo de 180 predicciones por minuto.
2. El impacto de dicho dato: En este caso el dato se encuentra dentro del rango de actividad usual del sistema. Sin embargo, si fuera más alto de 180 podría alertarnos sobre potenciales errores del sistema no notificados o si por el contrario fuera más bajo de 90 podría alertarnos de una posible carga de trabajo del sistema inusualmente alta.

A continuación, se detalla un **listado con diferentes ejemplos de indicadores** que deberán de ser seleccionados en función de la evaluación de riesgos realizada (Véase sección 5 de la Guía de Gestión de Riesgos: "¿Qué elementos debo implantar y cómo debo hacerlo para desarrollar un adecuado sistema de gestión de riesgos?") y transformados en registros (Véase sección 5 de la Guía de Conservación de Registros: "¿Qué elementos debo implantar y cómo debo hacerlo para desarrollar un adecuado sistema de gestión de registros?"):

Algunas notas aclaratorias sobre los listados de indicadores propuestos:

- No se trata de un listado exhaustivo sino de un listado mínimo sobre el que ampliar indicadores en función del sistema inteligente concreto.
- En los casos en los que el indicador obtenga un dato promedio también se aconseja obtener el valor mínimo y máximo para registrar valores puntuales anómalos de los indicadores. Por ejemplo, si se obtiene el número de predicciones por minuto promedio también se aconseja obtener el mínimo de predicciones realizadas por

minuto y el máximo. En definitiva, se aconseja incluir todas las medidas estadísticas necesarias para una correcta vigilancia del indicador.

- Resulta de especial importancia resaltar que los indicadores de ciberseguridad son un listado de ejemplos que debe de ser adaptado y ampliado por los responsables de seguridad del sistema inteligente en base a su contexto de aplicación. Por lo tanto, no se tratan de indicadores mínimos sino de ejemplos generales.
- Los umbrales mínimos y máximos establecidos en la escala deben de ir acompañados de un sistema de envío de alertas para actuar en tiempo real ante cualquier tipo de contingencia. También se aconseja contar con un *dashboard* de vigilancia en tiempo real de las mediciones obtenidas del sistema.
- En algunos casos contar solo con indicadores no es suficiente y se debe de considerar una revisión manual de los registros en función de la complejidad de estos. Por ejemplo, el número de orígenes de sesión diferentes para un usuario (supongamos 15 inicios de sesión en lugares diferentes) no solo debe de ser evaluado a través de un indicador sino en muchos casos a través de una revisión manual o automatizada de dichos orígenes. El objetivo es que el indicador actúe como alerta o disparador para dicha revisión.

7.1.1 Anexo A.I - Indicadores sobre el sistema inteligente

INDICADOR	DATO	ESCALA
Predicciones realizadas por unidad de tiempo	Número de predicciones realizadas. Ejemplo: 1.567 predicciones en el último minuto.	Mínimo esperado de predicciones y número máximo de predicciones sin pérdida de rendimiento. Ejemplo: mínimo 200 y máximo 22.000.
Tiempo medio de predicción	Tiempo entre la introducción de input en el sistema inteligente y su respuesta. Ejemplo: 0.65 segundos.	Tiempo mínimo esperado para una predicción y tiempo máximo aceptable. Ejemplo: 0.2 segundos / 2.5 segundos
Cola de procesamiento	Número de tareas en espera de ser procesadas por el modelo en tiempo real. Ejemplo: 10 tareas en cola.	Rango aceptable de cola de procesamiento. Ejemplo: 0 - 50 tareas.
Tasa de errores de predicción*	Porcentaje de predicciones incorrectas sobre el total de	Rango de tolerancia para errores. Ejemplo: 0% - 10%

	predicciones realizadas. Ejemplo: 3%	
Precisión del modelo*	Porcentaje de predicciones correctas sobre el total de predicciones realizadas. Ejemplo: 97%	Rango de precisión deseado. Ejemplo: 85% - 100%
Recall (Exhaustividad)*	Porcentaje de verdaderos positivos sobre el total de casos positivos reales (verdaderos positivos y falsos negativos). Especialmente importante en el caso de problemas desbalanceados. Ejemplo: 90%	Rango aceptable de *Recall*. Ejemplo: 80% - 100%
F1 Score del modelo*	Media armónica de precisión y exhaustividad. Ejemplo: 0.85	Rango aceptable de F1 Score. Ejemplo: 0.7 - 1.0
Variación en las medidas estadísticas de los inputs	Depende del input en particular, pero conceptualmente se trata del registro de mediciones estadísticas como la media, mediana y desviación típica de los valores de entrada para detectar variaciones considerables en tiempo real. Ejemplo para un sistema que recibe un texto como input: Media de 6.8 tokens / Desviación típica de 2.1	El rango aceptable de las medidas estadísticas seleccionadas para el input en concreto. Ejemplo: Media de mínimo 1 y máximo 100.

* En muchos casos las medidas relacionadas con errores de predicción del sistema no se pueden obtener en tiempo real por lo que no podrán aplicarse en la vigilancia continua sino en la vigilancia periódica descrita en la sección 4.2 de la presente guía "Vigilancia periódica". A modo de ejemplo, no podemos saber a tiempo real si el sistema de promoción de empleados ha cometido un error en la asignación de un nuevo puesto hasta no comprobar los resultados y KPIs obtenidos por el trabajador con el tiempo. En este caso no tiene sentido obtener indicadores de tasa de error del sistema de forma continuada sino realizar evaluaciones periódicas del mismo.

7.1.2 Anexo A.II - Indicadores sobre la infraestructura

INDICADOR	DATO	ESCALA
Uso de CPU	Porcentaje de utilización de la CPU en el sistema. Ejemplo: 60%	Rango aceptable de uso de CPU. Ejemplo: 20% - 90%
Media de procesos en el sistema	Número medio de procesos en el sistema por unidad de tiempo. Ejemplo: 433 procesos de media / minuto.	Rango aceptable de procesos activos de media. Ejemplo: 53 - 2500 procesos / minuto.
Uso de memoria RAM	Porcentaje de utilización de la memoria en el sistema. Ejemplo: 70%	Rango aceptable de uso de memoria. Ejemplo: 30% - 95%
Uso de GPU y sistemas similares	Porcentaje de utilización de la GPU. Ejemplo: 55%	Rango aceptable de uso de GPU. Ejemplo: 20% - 85%
Uso de almacenamiento	Porcentaje de utilización de almacenamiento en el sistema. Ejemplo: 80%	Rango aceptable de uso de almacenamiento. Ejemplo: 10% - 90%
Tiempo de actividad	Porcentaje de tiempo que el sistema permanece operativo sin interrupciones. Ejemplo: 26 horas.	Rango deseado de tiempo de actividad. Ejemplo: Máximo de 72 horas.
Ancho de banda de red	Capacidad de la red para transmitir datos por segundo. Ejemplo: 1.62 Gbps	Rango deseado de ancho de banda de red. Ejemplo: 100 Mbps - 10 Gbps
Temperatura del sistema	Temperatura promedio del sistema. Ejemplo: 22°C	Rango aceptable de temperatura ambiente. Ejemplo: 15°C - 30°C

7.1.3 Anexo A.III - Indicadores sobre las acciones de los usuarios

Los indicadores relacionados con las acciones de usuarios se pueden aplicar tanto de forma general para todos los usuarios como a nivel de usuario en función de las necesidades de vigilancia y riesgos del sistema inteligente. Por ejemplo, el número de inicios de sesión se puede aplicar como una métrica para todos los usuarios del sistema y también se puede aplicar para cada uno de los usuarios de este.

INDICADOR	DATO	ESCALA
Inicios de sesión	Número de inicios de sesión realizados por los usuarios en un periodo dado. Ejemplo: 500 inicios de sesión diarios.	Rango aceptable de inicios de sesión. Ejemplo: 100 - 1.000 diarios.
Número de interacciones por usuario por unidad de tiempo	Cantidad promedio de predicciones solicitadas por cada usuario en un periodo dado. Ejemplo: 50 predicciones por usuario al día.	Rango aceptable de interacciones por usuario. Ejemplo: 0 - 1000 al día.
Tiempo promedio entre interacciones	Duración promedio de las sesiones de usuario en la plataforma. Ejemplo: 15 minutos	Rango deseado de tiempo promedio en la plataforma. Ejemplo: 5 - 60 minutos.
Interacciones con la interfaz	Número de interacciones realizadas por los usuarios con la interfaz en un periodo dado. Ejemplo: 2.000 interacciones diarias	Rango deseado de interacciones con la interfaz. Ejemplo: 500 - 5.000 diarias.
Medidas estadísticas de los inputs introducidos por usuario	Mismos datos que "Variación en las medidas estadísticas de los inputs" de la tabla de indicadores sobre el sistema inteligente, pero asociando	Misma escala que "Variación en las medidas estadísticas de los inputs" de la tabla de indicadores sobre el sistema inteligente, pero asociando

	dichas métricas con un usuario en particular.	dichas métricas con un usuario en particular.
Número de orígenes de sesión diferentes	En algunos casos resulta fundamental controlar la dispersión de inicios de sesión tanto bajo la perspectiva técnica como de ciberseguridad. Ejemplo: 3 orígenes diferentes.	Rango aceptable de orígenes diferentes de inicios de sesión por usuario. Ejemplo: 1 - 20.

7.1.4 Anexo A.IV - Indicadores de seguridad

Los siguientes indicadores de ciberseguridad son algunos ejemplos que deben ser ampliados por los responsables de seguridad del sistema inteligente en base a su contexto de aplicación siguiendo las recomendaciones de la guía de ciberseguridad. Además, los indicadores deben de combinarse con análisis manuales de los registros obtenidos para una comprensión completa de cada escenario.

INDICADOR	DATO	ESCALA
Intentos de inicio de sesión fallidos	Número de intentos fallidos de inicio de sesión en un periodo dado. Ejemplo: 50 fallidos diarios.	Rango aceptable de intentos fallidos. Ejemplo: 0 - 100 diarios
Cambios no autorizados en archivos	Número de modificaciones no autorizadas en archivos del sistema en un periodo dado. Ejemplo: 1 cambio detectado.	Importancia alta a partir de un solo cambio.
Número de usuarios activos en el sistema	Número de usuarios logueados en el sistema en un momento dado. Ejemplo: 10 usuarios.	Rango aceptable de usuarios logueados. Ejemplo: 2 - 3 usuarios.

Volumen de datos transferidos en la red	Cantidad de datos transferida por unidad de tiempo en la red. Ejemplo: 103.2 MB / segundo.	Rango aceptable de cantidad de datos. Ejemplo: 10 - 750 MB / segundo.
Número de comunicaciones abiertas	Cantidad de conexiones de red establecidas por unidad de tiempo. Ejemplo: 2.498 conexiones / hora.	Rango aceptable de conexiones. Ejemplo: 200 - 20.000 conexiones / hora.
Este listado solo representa un ejemplo con diferentes indicadores y debe de ser completado en función de las características del sistema inteligente y el análisis de riesgos realizado.		

8. Referencias

[1] Akenine-Möller, T., & Johnsson, B. (2012). Performance per what? Journal of Computer Graphics Techniques, 1, 37-41.
[2] Amazon AI. (2021). Sagemaker Clarify: Amazon AI Fairness and Explainability Whitepaper. https://pages.awscloud.com/rs/112-TZM-766/images/Amazon.AI.Fairness.and.Explainability.Whitepaper.pdf
[3] Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety.
[4] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82-115.
[5] Artelt, L. et al. (2021). Evaluating Robustness of Counterfactual Explanations.
[6] Bella, A., Ferri, C., Hernández-Orallo, J., & Ramírez-Quintana, M. J. (2010). Calibration of Machine Learning Models. Departamento de Sistemas Informáticos y Computación, Universidad Politécnica de Valencia.
[7] Ben-David, S., Blitzer, J., Crammer, K., & Pereira, F. (2007). Analysis of representations for domain adaptation. Advances in neural information processing systems.
[8] Bennetot, A. (2022). A Neural-Symbolic learning framework to produce interpretable predictions for image classification (Doctoral dissertation).
[9] Besmira, N., Ece, K., & Eric, H. (2018). Towards Accountable AI: Hybrid Human-Machine Analyses for Characterizing System Failure. HCOM.
[10] Blouw, P., Choo, X., Hunsberger, E., & Eliasmith, C. (2019). Benchmarking keyword spotting efficiency on neuromorphic hardware. Proceedings of the 7th Annual Neuro-inspired Computational Elements Workshop, 1-8.
[11] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners.
[12] Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In S. A. Friedler & C. Wilson (Eds.), Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Vol. 81). Proceedings of Machine Learn Research.
[13] Burns, K., Hendricks, L. A., Saenko, K., Darrell, T., & Rohrbach, A. (2018). Women also Snowboard: Overcoming Bias in Captioning Models. ECCV'18, 771-787.
[14] Catalog of Bias. (n.d.). Retrieved from https://catalogofbias.org
[15] Chen, N. (2018). Metrics for Deep Generative Models.
[16] Chen, S. F., Beeferman, D., & Rosenfeld, R. (1998). Evaluation Metrics for Language Models.
[17] De Laat, P. B. (2021). Companies Committed to Responsible AI: From Principles towards Implementation and Regulation? Philosophy & Technology, 34, 1135–1193.
[18] Del Ser, J. et al. (2022). Exploring the Trade-off between Plausibility, Change Intensity and Adversarial Power in Counterfactual Explanations using Multi-objective Optimization.

[19] Deloitte. (2020, August 26). Deloitte introduces trustworthy AI framework to guide organizations in ethical application of technology.
[20] Díaz-Rodríguez, N., Lomonaco, V., Filliat, D., & Maltoni, D. (2018). Don't forget, there is more than forgetting: new metrics for Continual Learning. NeurIPS workshop on Continual Learning.
[21] Díaz-Rodríguez, N., Vellido, A., & Moreno, A. (2021). Questioning causality on sex, gender and COVID-19, and identifying bias in large-scale data-driven analyses: the Bias Priority
[22] Dieterich, D. et al. (2016). COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity.
[23] Dietterich, T. G. (2017). Steps Toward Robust Artificial Intelligence.
[24] Dietterich, T. G., & others. (1998). Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. Neural Computation, 10(7), 1895-1923.
[25] DiveDeep AI. (2022). Data drift vs concept drift. https://divedeep.ai/2022/03/17/data-drift-vs-concept-drift/
[26] Eigner, P. (2021). Towards Resilient Artificial Intelligence: Survey and Research Issues.
[27] ENISA. (2021). SECURING MACHINE LEARNING ALGORITHMS.
[28] Epstein, Z., et al. (2018). Turingbox: an experimental platform for the evaluation of AI systems. IJCAI International Joint Conference on Artificial Intelligence, 2018-July, 5826-5828.
[29] Fabrizzi, L. et al. (2022). A survey on bias in visual datasets.
[30] Ferrario, A. et al. (2022). The Robustness of Counterfactual Explanations Over Time.
[31] Franklin, et al. (2022). An Ontology for Fairness Metrics. Retrieved from https://dl.acm.org/doi/pdf/10.1145/3514094.3534137
[32] Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. ACM computing surveys (CSUR), 46(4), 1-37.
[33] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for Datasets.
[34] Gendered Innovations. (2018). Facial Recognition: Analyzing Gender and Intersectionality in Machine Learning. Retrieved from http://genderedinnovations.stanford.edu/case-studies/facial.html#tabs-2
[35] Gloor, L. (2016). Suffering-focused AI safety: In favor of "fail-safe" measures. Center on Long-Term Risk Report.
[36] Google. (2021). Machine Learning Glossary: Fairness. Retrieved November 29, 2021, from https://developers.google.com/machine-learning/glossary/fairness.
[37] HCAI. (2022). Human-centred artificial intelligence. Retrieved from https://scilog.fwf.ac.at/en/environment-and-technology/15317/human-centred-artificial-intelligence
[38] Henderson, P. (2017). Deep Reinforcement Learning that Matters.
[39] Hertweck, C., & Räz, T. (2022). Gradual (In)Compatibility of Fairness Criteria. arXiv preprint arXiv:2109.04399.
[40] Heuillet, A., Couthouis, F., & Díaz-Rodríguez, N. (2021). Explainability in Deep Reinforcement Learning. Knowledge-Based Systems, 214, 106685.

[41] Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the Knowledge in a Neural Network.
[42] Hockert, T. (2010). Safeguard By Design: Lessons Learned from DOE Experience Integrating Safety in Design.
[43] Holzinger, A. (2016). Interactive Machine Learning for Health Informatics: When do we need the human-in-the-loop? Brain Informatics, 3(2), 119-131. doi:10.1007/s40708-016-0042-6.
[44] Holzinger, A. et al. (2022). Digital Transformation in Smart Farm and Forest Operations Needs Human-Centered AI: Challenges and Future Directions.
[45] Holzinger, A., Carrington, A., & Müller, H. (2020). Measuring the quality of explanations: the system causability scale (SCS). KI-Künstliche Intelligenz, 34(2), 193-198.
[46] Holzinger, A., Dehmer, M., Emmert-Streib, F., Cucchiara, R., Augenstein, I., ... (2022). Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence. Information Fusion, 79, 263-278.
[47] Holzinger, A., Kargl, M., Kipperer, B., Regitnig, P., Plass, M., & Müller, H. (2022). Personas for Artificial Intelligence (AI): An Open Source Toolbox. IEEE Access, 10, 23732-23747.
[48] Holzinger, A., Malle, B., Kieseberg, P., Roth, P. M., Müller, H., Reihs, R., & Zatloukal, K. (2017). Machine Learning and Knowledge Extraction in Digital Pathology needs an integrative approach. In Springer Lecture Notes in Artificial Intelligence (Vol. LNAI 10344). Springer International. doi: 10.1007/978-3-319-69775-8_2
[49] Holzinger, A., Malle, B., Kieseberg, P., Roth, P.M., Müller, H., Reihs, R. & Zatloukal, K. (2017). Towards the Augmented Pathologist: Challenges of Explainable-AI in Digital Pathology. arXiv:1712.06657.
[50] Hullermeier, E., Waegeman, W., Pölsterl, S., & Szepesvári, C. (2019). Aleatoric and Epistemic Uncertainty in Machine Learning: A Tutorial Introduction.
[51] Hurwicz, L., & Reiter, S. (2006). Designing Economic Mechanisms.
[52] Huyen, C. (2019). Evaluation Metrics for Language Modeling. The Gradient. https://thegradient.pub/understanding-evaluation-metrics-for-language-models/
[53] IBM. (2021). Uncertainty Quantification 360 Toolkit. Retrieved from https://uq360.mybluemix.net
[54] Information Commissioner's Office. (2020). Guidance on the AI auditing framework: draft guidance for consultation.
[55] ISO/IEC 25000. (2021). Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) and ISO/IEC WD 25059:2021, Software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) Quality Model for AI systems.
[56] Kaczmarek-Majer, K., Casalino, G., Castellano, G. et al. (2022). PLENARY: Explaining black-box models in natural language through fuzzy linguistic summaries. Information, Elsevier.
[57] Kalifou, R. T., Caselles-Dupré, H., Lesort, T., Sun, T., Diaz-Rodriguez, N., & Filliat, D. (2019). Continual reinforcement learning deployed in real-life using policy distillation and sim2real transfer. ICML Workshop on Multi-Task and Lifelong Learning.
[58] Kusters, R., Misevic, D., Berry, H., Cully, A., Le Cunff, Y., Dandoy, L., ... (2020). Interdisciplinary Research in Artificial Intelligence: Challenges and Opportunities. Frontiers in Big Data, 3, 45.
[59] Le Quy, T., Roy, A., Iosifidis, V., Zhang, W., & Ntoutsi, E. (2022). A survey on datasets for fairness-aware machine learning.
[60] Lesort, T., Díaz-Rodríguez, N., Goudet, O., & Filliat, D. (2022). Understanding Continual Learning Settings with Data Distribution Drift Analysis. https://www.youtube.com/watch?v=WFhozvAgnsU

[61] Lesort, T., Lomonaco, V., Stoian, A., Maltoni, D., Filliat, D., & Díaz-Rodríguez, N. (2020). Continual Learning for Robotics: Definition, Framework, Learning Strategies, Opportunities and Challenges. Information Fusion, 220.
[62] Lesort, T., Seurin, M., Li, X., Díaz-Rodríguez, N., & Filliat, D. (2019). Deep Unsupervised state representation learning with robotic priors: a robustness analysis. 2019 International Joint Conference on Neural Networks (IJCNN).
[63] Lopez-Paz, D., Muandet, K., Scholkopf, B., & Tolstikhin, I. O. (2015). Towards a learning theory of cause-effect inference. In F. R. Bach & D. M. Blei (Eds.), Proceedings of the 32nd International Conference on Machine Learning, ICML 2015 (Vol. 37, pp. 1452-1461). JMLR.org. http://proceedings.mlr.press/v37/lopez-paz15.html
[64] Lopez-Paz, D., Nishihara, R., Chintala, S., Scholkopf, B., & Bottou, L. (2017). Discovering causal signals in images. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 58-66). doi: 10.1109/CVPR.2017.14
[65] Lutjens, B., Sutanudjaja, E., Straatsma, M., & Maris, M. (2021). Physically-Consistent Generative Adversarial Networks for Coastal Flood Visualization.
[66] Mallya, A. (2018). Piggyback: Adapting a Single Network to Multiple Tasks by Learning to Mask Weights.
[67] Mauri, L. et al. (2021). STRIDE-AI: An Approach to Identifying Vulnerabilities of Machine Learning Assets. IEEE CSR.
[68] McSherry, F. (2022). Materialize: a platform for building scalable event based systems.
[69] McSherry, F., & Talwar, K. (2008). Mechanism design via Differential Privacy.
[70] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Statt, C., ... & Gebru, T. (2019). Model cards for model reporting. Proceedings of the 2020 conference on fairness, accountability, and transparency.
[71] Morris, J. et al. (2020). TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations.
[72] Nobel Prize Committee. (2007). Mechanism Design Theory. The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel.
[73] Orcaa. (2020). It's the age of the algorithm and we have arrived unprepared.
[74] Papernot, N. (2016). Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks.
[75] Pisoni, G., Díaz-Rodríguez, N., Gijlers, H., & Tonolli, L. (2021). Human-Centered Artificial Intelligence for Designing Accessible Cultural Heritage. Applied Sciences, 11(2), 870.
[76] PwC. (2020). PwC Ethical AI Framework.
[77] Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., & Lawrence, N. D. (2022). Dataset Shift in Machine Learning.
[78] Raffin, A. (2021). Stable-Baselines3 Reliable Reinforcement Learning Implementations. https://stable-baselines3.readthedocs.io/en/master/
[79] Raffin, A., Hill, A., Lesort, T., Traoré, R., & Díaz-Rodríguez, N. (2018). S-RL Toolbox: Environments, Datasets and Evaluation Metrics for State Representation Learning.
[80] Raffin, A., Hill, A., Traoré, R., Lesort, T., Díaz-Rodríguez, N., & Filliat, D. (2018). NeurIPS workshop on Deep Reinforcement Learning.
[81] Raffin, A., Hill, A., Traoré, R., Lesort, T., Díaz-Rodríguez, N., & Filliat, D. (2018). S-RL Toolbox: Environments, Datasets and Evaluation Metrics for State Representation Learning. NeurIPS workshop on Deep Reinforcement Learning.
[82] Raffin, A., Hill, A., Traoré, R., Lesort, T., Díaz-Rodríguez, N., & Filliat, D. (2019). Decoupling feature extraction from policy learning: assessing benefits of state representation learning in goal based robotics. ICLR 2019 Workshop on Structure & Priors in Reinforcement Learning (SPIRL).
[83] Rahtz, D. (2022). Safe Deep RL in 3D environments using human feedback.

[84] Rodríguez, N.D., Cuéllar, M.P., Lilius, J., & Calvo-Flores, M.D. (2014). A fuzzy ontology for semantic modelling and recognition of human behaviour. Knowledge-Based Systems, 66, 46-60.
[85] Rodríguez, N.D., Cuéllar, M.P., Lilius, J., & Calvo-Flores, M.D. (2014). A survey on ontologies for human behavior recognition. ACM Computing Surveys (CSUR), 46(4), 1-33.
[86] Rohrbach, A., Hendricks, L.A., Burns, K., Darrell, T., & Saenko, K. (2018). Object Hallucination in Image Captioning. Proceedings of the EMNLP'18.
[87] Ross, A.S., Hughes, M.C., & Doshi-Velez, F. (2017). Right for the Right Reasons: Training Differentiable Models by Constraining their Explanations. Proceedings of the IJCAI'17.
[88] Russel, S. (2015). Research priorities for robust and beneficial artificial intelligence.
[89] Schwartz, R. S., Dodge, J., Smith, N. A., & Ettinger, M. (2019). Green AI.
[90] Sena, L. H., et al. (2019). Incremental Bounded Model Checking of Artificial Neural Networks in CUDA.
[91] Shi, et al. (2020). Robustness Verification for Transformers. International Conference on Learning Representations. arXiv:2002.06622
[92] Sotala, K., & Gloor, L. (2017). Superintelligence as a Cause or Cure for Risks of Astronomical Suffering. Informatica, 41, 389–400.
[93] Stanton, S., Izmailov, P., Kirichenko, P., Alemi, A. A., & Wilson, A. G. (2021). Does knowledge distillation really work? Retrieved from https://arxiv.org/abs/2106.05945
[94] Strobel, B., Yoo, S., Papernot, N., & Kumar, S. (2022). Data Privacy and Trustworthy Machine Learning.
[95] Tomkins, S., Isley, S., London, B., & Getoor, L. (2018). Sustainability at scale: towards bridging the intention-behavior gap with sustainable recommendations. Proceedings of the 12th ACM conference on.
[96] Widmer, G. et al. (2022). Towards Human-Compatible XAI: Explaining Data Differentials with Concept Induction over Background Knowledge.
[97] Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. Biometrics Bulletin, 1(6), 80-83.
[98] Wilms, I. et al. (2021). Omitted variable bias: A threat to estimating causal relationships.
[99] Yang, D., Rangwala, H., Johri, A., & Rose, C. P. (2022). Generalized out-of-distribution detection: A survey.
[100] Yoo, S., Yang, E., & Zhang, Y. (2020). Blackbox NLP Workshop track proceedings. EMNLP. Retrieved from https://github.com/QData/TextAttack
[101] Zech, J. et al. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study.
[102] Zheng, S. (2016). Improving the Robustness of Deep Neural Networks via Stability Training.

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