10

Empresas desarrollando cumplimiento de requisitos

Guía 10. Solidez

Reglamento Europeo de Inteligencia Artificial

Financiado por la Unión Europea
NextGenerationEU

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

Plan de Recuperación, Transformación y Resiliencia

España | digital 2026

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

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

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

Entre las referencias técnicas relevantes actualmente en desarrollo, destaca el **prEN 18229-2 "AI Trustworthiness Framework – Part 2: Accuracy and Robustness"**, que servirá de base para la evaluación de la precisión y la solidez de los sistemas de IA una vez sea adoptado como norma armonizada en el contexto del cumplimiento del Reglamento Europeo de Inteligencia Artificial.

Fecha de versión: 10 de diciembre de 2025

Contenido General

1. Preámbulo 5
2. Introducción 9
3. Reglamento de Inteligencia Artificial 13
4. ¿Cómo abordar los requisitos? 16
5. Documentación técnica 37
6. Cuestionario de autoevaluación 40
7. Anexos 41
8. Referencias, estánderes y normas 63

Índice Detallado

1. Preámbulo 5
1.1 Objetivo del documento 5
1.2 ¿Cómo leer esta guía? 6
1.3 ¿A quién está dirigido? 7
1.4 Casos de uso utilizados en la guía 8
2. Introducción 9
2.1 ¿Qué es la solidez para IA? 9
3. Reglamento de Inteligencia Artificial 13
3.1 Análisis previo y relación de los artículos 13
3.2 Contenido de los artículos en el Reglamento de IA 14
3.3 Correspondencia del articulado con los apartados de la guía 15
4. ¿Cómo abordar los requisitos? 16
4.1 Evaluación de la solidez 16
4.1.1 Ciclo de vida 18
4.1.2 Selección de métricas 19
4.1.3 Validación y Verificación 21
4.1.4 Eficiencia en sistemas de IA 22
4.1.5 Rendimiento en sistemas de IA 23
4.1.6 Monitorización de la solidez 25
4.2 Solidez y resistencia a errores 26
4.3 Redundancia y planes de respaldo 29
4.4 Solidez para sistemas que continúan aprendiendo 31
4.4.1 Tipos de degradación del sistema de IA y planes de mitigación, estrategias y herramientas 33
4.4.2 Estrategias para mitigar cambios en la degradación de la precisión y solidez del modelo y/o de sus datos 34
5. Documentación técnica 37
5.1 Certificación de la solidez de modelos 38
6. Cuestionario de autoevaluación 40
7. Anexos 41
7.1 Anexo I: Métricas y pruebas para sistemas de IA 41
7.1.1 Métricas para establecer la solidez 41
7.1.2 Métodos de prueba 42
7.2 Anexo II: Solidez e incertidumbre 46
7.2.1 La solidez como capacidad de tratamiento y minimización de la incertidumbre 46

7.2.2 Métodos para medir y cuantificar la incertidumbre de la salida de un modelo de IA 47
7.3 Glosario 47
7.4 Apéndice: Traducciones críticas (inglés - español) 62
8. Referencias, estánderes y normas 63
8.1 Referencias 63
8.2 Estándares 70

1. Preámbulo

1.1 Objetivo del documento

Un sistema de IA de alto riesgo tiene que estar preparado para minimizar y prevenir los comportamientos perjudiciales e indeseables y ser capaz de detectar cuando su funcionamiento tenga lugar fuera de aquel dominio de entrada y ejecución establecido por su finalidad prevista. De igual modo tiene que estar diseñado e implementado para evitar la adopción de decisiones equivocadas o generar una información de salida errónea. Todo ello para evitar consecuencias negativas para la seguridad y los derechos fundamentales.

Es objetivo de esta guía establecer las medidas necesarias que, a nuestro entender, sirven para cubrir los requerimientos del Reglamento Europeo de la IA, en lo relativo a la solidez del sistema de alto riesgo.

El Reglamento Europeo de la IA considera el concepto de **solidez técnica** como un mecanismo **clave** para los sistemas de IA de alto riesgo. Así lo desarrolla a través de su considerando (75), donde establece que:

AI Act

Considerando (75)

La solidez técnica es un requisito clave para los sistemas de IA de alto riesgo, que deben ser resilientes en relación con los comportamientos perjudiciales o indeseables por otros motivos que puedan derivarse de limitaciones en los sistemas o del entorno en el que estos funcionan (p. ej., errores, fallos, incoherencias o situaciones inesperadas).

En las medidas propuestas en esta guía se establece que el sistema de IA deberá considerar en su diseño, desarrollo y validación, las soluciones técnicas y organizativas para prevenir dichas situaciones, incluso, con la posibilidad de que cuando, dicha solidez técnica no opere dentro de parámetros seguros, el sistema pueda interrumpir su funcionamiento. Tal es así que lo establece en el mismo considerando (75):

AI Act

Considerando (75)

> […] los sistemas de IA de alto riesgo, por ejemplo mediante el diseño y desarrollo de soluciones técnicas adecuadas para prevenir o reducir al mínimo ese comportamiento perjudicial o indeseable. Estas soluciones técnicas pueden incluir, por ejemplo, mecanismos que permitan al sistema interrumpir de forma segura su funcionamiento (planes de prevención contra fallos) en presencia de determinadas anomalías o cuando el funcionamiento tenga lugar fuera de determinados límites predeterminados. […]

Dicho de otro modo, la solidez del sistema de IA es tan relevante que en aquellos casos en los que no pueda garantizarse adecuadamente, el sistema debe estar preparado para **interrumpir** su funcionamiento de **manera controlada.**

Es responsabilidad del proveedor del sistema de inteligencia artificial de alto riesgo tomar las medidas adecuadas (tanto organizativas como técnicas) para garantizar que se cumple con los requerimientos de solidez del sistema. Igualmente, dentro de su ámbito de aplicación, el responsable del despliegue del sistema también tiene responsabilidades que se materializarán en medidas concretas (de nuevo organizativas y técnicas).

Desarrollaremos a lo largo de la guía una serie de medidas organizativas y técnicas destinadas primero a validar el sistema de inteligencia artificial y, a continuación, a aplicar los controles de calidad del modelo que ayudan a verificar y validar las razones que llevan a seleccionar tales métricas.

En cuanto a aspectos complementarios que son clave para poder implantar el sistema con el objetivo de ejecutar su finalidad prevista, estas métricas van más allá del resultado inmediato del modelo para un baremo de pruebas, y pueden tener implicaciones de discriminación, sesgo o imprecisión.

A lo largo de la guía abordaremos la relación con estos aspectos complementarios, y su inclusión en el concepto de solidez.

1.2 ¿Cómo leer esta guía?

Esta guía está estrechamente relacionada con las guías relativas al artículo 15 de precisión, solidez y ciberseguridad.

La solidez del sistema de IA tiene, entre otros objetivos, la conservación de la precisión obtenida durante el entrenamiento, prueba y validación del sistema a lo largo del ciclo de vida de este; especialmente en aquellas entradas que no se encuentran en el conjunto de entrenamiento y validación, pero que se encuentran dentro del dominio del sistema de IA.

La relación con la ciberseguridad es también estrecha. Por un lado, la ciberseguridad contribuye a la solidez del sistema al protegerlo frente a ataques adversarios, que podrían

alterar su respuesta (y por ende su precisión). Por otro lado, los mecanismos de solidez deben garantizar que las medidas de ciberseguridad se mantienen en el tiempo.

Para una correcta aplicación de esta guía, se ha estructurado siguiendo el artículo del Reglamento Europeo de la IA, y a la vez proporcionando una serie de medidas que permitan cubrir los requisitos establecidos en el artículo 15 previamente analizado, en relación con la solidez.

La guía aborda en primer lugar en el apartado 4, como establecer unas métricas de solidez para el sistema, su validación y verificación. Es un apartado de peso importante en relación con las responsabilidades del proveedor del sistema de IA, porque establece las bases para definir la solidez de su sistema de IA. Se acompaña un detalle ampliado en el 7.1 Anexo I: Métricas y pruebas para sistemas de IA que puede ser consultado para ampliar la base teórica de lo que se pretende explicar en el citado apartado. En este apartado, la solidez se relaciona con eficiencia, rendimiento y monitorización para ofrecer una visión completa.

Siguiendo las recomendaciones del Reglamento Europeo de la IA en el artículo 15, se presentan en el apartado 4.1 medidas para conservar la solidez establecida cuando el sistema está expuesto a errores.

En el apartado 4, se establece que la solidez, una vez definida y protegida frente a fallos, pueda estar garantizada gracias a una serie de medidas destinadas a establecer su continuidad en el tiempo.

Para aquellos sistemas que continúan aprendiendo con el tiempo, ya sea porque lo hacen de manera automática, o de manera programada a través de recolección de datos (o retroalimentación de usuarios) y progresivas actualizaciones, en el apartado 5 se proporcionan medidas técnicas específicas.

El proveedor debe considerar que esta guía ofrece unas recomendaciones sobre la cobertura de los requisitos establecidos en el artículo 15 del Reglamento Europeo de la IA, pero que no es un compendio exhaustivo, y que puede que, en algunos casos, tenga necesariamente, que considerar adaptar los detalles específicos (especialmente lo referido en los Anexos) a la naturaleza y la finalidad prevista del sistema y los riesgos para la salud y los derechos y libertades de las personas físicas.

Las instrucciones sobre cómo obtener la solidez asociada al sistema, cómo usarla e interpretarla, sus umbrales y métricas asociadas a ella deben documentarse según la guía de Documentación técnica. A continuación, a lo largo de la guía detallaremos dimensiones a tener en cuenta para ello.

1.3 ¿A quién está dirigido?

Dar cabida a todas las cuestiones detalladas en esta guía es responsabilidad del proveedor del sistema de inteligencia artificial de alto riesgo tomando las medidas adecuadas aquí propuestas, tanto organizativas como técnicas. Todo ello con el objetivo de garantizar que se cumple con los requisitos de precisión del sistema.

Dentro de su ámbito de aplicación, el responsable del despliegue del sistema también tiene responsabilidades que se materializarán en medidas concretas, de nuevo organizativas y técnicas. La guía indicará, en cada caso, cuales le son de aplicación y el alcance de estas.

1.4 Casos de uso utilizados en la guía

A lo largo de la guía se utilizarán dos **casos de uso** a modo de **ejemplo** de cómo **elaborar** la documentación técnica. Los ejemplos estarán centrados únicamente en el proveedor que es el responsable de generar y conservar la documentación. La descripción detallada de los casos de uso utilizados podrá encontrarse en la Guía práctica y ejemplos para entender el Reglamento de IA**.**

Nota: Siempre que se ponga un **ejemplo**, se hará **de manera ilustrativa**. **Proveedor y responsable del despliegue** han de considerar la aplicación de **todas las medidas** indicadas en esta guía.

Los casos de uso se han seleccionado atendiendo a su capacidad para explicar la información y procedimientos detallados en esta guía, por continuidad con la guía de precisión, se han seleccionado los mismos casos.

Los casos seleccionados en este caso para la elaboración de la guía son:

- **Detección de denuncias falsas**
- **Sistema de promoción de empleados.**

2. Introducción

2.1 ¿Qué es la solidez para IA?

La solidez se puede expresar a partir de una serie de propiedades asociadas al sistema de IA que permitirán definir y establecer claramente diferentes aristas del sistema de IA.

Las palabras y descripciones **destacadas** se corresponden con términos que son desarrollados en el glosario (ver 7.3 Glosario).

Podemos decir que la relación entre estas propiedades, la solidez y el sistema de IA no pueda entenderse los unos, sin las otras. Estas propiedades que hemos considerado deseables son:

- **Fiabilidad**: Se refiere a la consistencia entre los valores inicialmente estimados y los valores estimados subsecuentes, o la coherencia general de una medida en estadística. La solidez suele considerarse una subcaracterística de la fiabilidad [ISO/IEC 25059:2023]. En el marco del método SQuaRE [75], la fiabilidad y la solidez se integran en la evaluación de la calidad de los sistemas de software, lo que permite una medición sistemática y alineada con estándares de calidad [ISO/IEC 25010].
- **Estabilidad:** En sistemas de inteligencia artificial, la estabilidad expresa el grado en que se permite que la salida del sistema cambie de forma controlada cuando sus entradas varían dentro de un dominio específico. Esta propiedad es crucial en aplicaciones que requieren consistencia en las respuestas ante pequeñas fluctuaciones en los datos de entrada, como ocurre en tareas de clasificación, identificación o toma de decisiones automatizada. La estabilidad se basa en la disponibilidad de información previa sobre la salida esperada del sistema, ya sea conocida por el usuario, derivada de simulaciones o definida mediante sistemas de optimización. Esta característica es relevante en redes neuronales, donde ayuda a asegurar que las variaciones en las entradas no conduzcan a salidas drásticamente diferentes. Sin embargo, también es aplicable a otros modelos de IA, como sistemas de reglas, árboles de decisión o métodos de aprendizaje estadístico, siempre que se espere que el sistema mantenga una respuesta predecible y coherente dentro de un rango de variabilidad de las entradas.
- **Sensibilidad**: En sistemas de inteligencia artificial, la sensibilidad expresa el grado en que se permite que la salida del sistema varíe cuando sus entradas experimentan cambios. Similar a la estabilidad, el análisis de sensibilidad requiere verificar que el sistema esté restringido dentro de ciertos límites de respuesta y es particularmente relevante en aplicaciones que exigen coherencia en condiciones específicas de entrada, como en tareas de interpolación o regresión. Este criterio se establece normalmente en términos cuantificables (como un umbral de variación permitido dentro de un rango específico de entradas), lo cual facilita la comparación entre distintas arquitecturas o enfoques de IA. La sensibilidad es crucial en sistemas que requieren precisión directa con respecto a un valor de referencia conocido (*ground truth*), como los sistemas de IA que controlan dispositivos médicos. Por ejemplo, en el caso de una bomba de insulina controlada por IA, la sensibilidad del sistema es un

factor determinante para su seguridad y fiabilidad, ya que la administración de insulina debe mantenerse dentro de los límites de las dosis recomendadas por el fabricante, sin exceder esos parámetros.

- **Relevancia**: Expresa la magnitud del impacto que las distintas entradas tienen sobre las salidas en un sistema de inteligencia artificial. Dado que diferentes modelos de IA, incluidas las redes neuronales, pueden arrojar criterios de relevancia distintos y, aun así, ser ambos válidos, cualquier protocolo de comparación entre modelos debe incluir un mecanismo para resolver posibles discrepancias (por ejemplo, mediante un sistema de votación). Este criterio es especialmente útil en aplicaciones donde la IA realiza una tarea que un humano podría realizar, y por lo tanto debe ser capaz de entender y verificar las salidas. Evaluar la relevancia es esencial para confirmar que la precisión del sistema está sustentada en los factores correctos, y este proceso de verificación puede ser realizado tanto por un operador humano como de forma automática mediante referencias previamente validadas.
- **Alcanzabilidad**: Esta propiedad se refiere a la precisión multi-paso de un sistema de IA en interacción con su entorno de operación. Aplica a sistemas que funcionan bajo el paradigma de agentes, donde existe un ciclo de percepción-acción-entorno en el cual el agente percibe el estado del entorno, ejecuta acciones, y evalúa el impacto de estas hasta lograr ciertos objetivos. La propiedad de alcanzabilidad evalúa si el sistema, al aplicar un modelo de IA (como una red neuronal u otro tipo de algoritmo), puede alcanzar un conjunto de estados definidos, ya sean objetivos a lograr o estados de fallo a evitar, en función de sus decisiones. Esta propiedad es útil en diversos enfoques de IA, proporcionando una métrica de precisión y rendimiento en sistemas de bucle cerrado, es decir, aquellos donde el sistema responde de manera continua a las condiciones del entorno en tiempo real.

Ejemplo – Detención de denuncias falsas

En el análisis de riesgos para el sistema de denuncias falsas se establece que el sistema debe ser capaz de mantener una respuesta uniforme para pequeñas variaciones en los textos de la denuncia, para evitar que determinados datos o palabras clave puedan disparar la categoría de una denuncia, causando que se clasifique incorrectamente. Esta clasificación incorrecta es especialmente grave, en el caso de una **clasificación** de una denuncia **verdadera como falsa**, causando que no se trate adecuadamente.

Por ello se considera especialmente relevante la **sensibilidad** del sistema. Para verificar la **sensibilidad del sistema** de IA, respecto a variaciones de entradas, se establece su valoración local, considerada que su valoración local, para medir la variación **alrededor** de denuncias de los conjuntos de datos, modificadas en su entorno, pero no muy alejadas entre sí.

La evaluación local del sistema se va a realizar sobre el conjunto de datos verificación, y posteriormente extender al proceso de validación (ver apartado 4.1.3).

El procedimiento seguido para esta verificación consiste en los siguientes pasos:

- En el conjunto de verificación se selecciona el mismo número de denuncias etiquetadas como falsas y verdaderas.

- Se utiliza una bolsa de palabras genérica, suficientemente amplia, del castellano.
- Para cada denuncia con la bolsa de palabras se generan nuevas denuncias, con cambios de palabras aleatorios del 5%, es decir se crean denuncias, *próximas* (locales) a las seleccionadas.
- Se establece que, durante la fase de verificación, se estudiará la **sensibilidad** del sistema frente a estas denuncias generadas y aquellas que se han seleccionado originalmente como base para estas

Las propiedades de solidez pueden ser **locales o globales**. Es más común verificar la solidez local que global, pues la primera es más fácil de especificar. Con local nos referimos a verificar las propiedades de solidez dentro de los dominios de entrada y salida del sistema, y siempre en relación con su finalidad prevista. Por ejemplo, un clasificador de imágenes predice correctamente "un coche". La propiedad de solidez local puede especificar que todas las imágenes generadas rotando la imagen original 5 grados deben ser calificadas como un coche.

Un inconveniente de verificar propiedades de solidez local es que las garantías son locales a la muestra testeada y no se extienden a otras muestras en la base de datos, este inconveniente se puede soslayar, aumentando y consolidando las pruebas de solidez local para ser lo más extensas posible, y cubrir variaciones, dentro del dominio esperado. Esto permitirá encontrar puntos del dominio de entrada que son mucho menos robustos que sus vecinos, o encontrar zonas de frontera de dominio en las que el sistema es menos robusto y abordar a lo largo del proceso de diseño e implementación su mejora paulatina.

Por otro lado, las propiedades de solidez global definen garantías que se cumplen de manera determinista sobre todas las posibles entradas (éstas deben especificarse definiendo un rango válido para características de entrada, lo cual es más difícil en configuraciones donde las características individuales no tienen significado semántico).

Ambos entornos de solidez (globales y locales) tienen igual importancia para un sistema de IA de alto riesgo y acorde al estado del arte del tipo de modelo, la finalidad prevista y los riesgos para la salud, los derechos y libertades fundamentales, ambas deben ser verificadas. La solidez local, por permitir encontrar y actuar sobre puntos menos robustos y mejorar el sistema de IA en el proceso de su diseño, prueba y validación. La global por ofrecer garantías del comportamiento del modelo en los rangos de entrada del dominio esperado. Especial interés tiene (tal y como se detalla en la guía de documentación) considerar dentro de las propiedades de solidez, los usos indebidos razonablemente previsibles del sistema de IA, para que sean consideradas en el análisis global y local de la misma.

Propiedades deseables al diseñar sistemas de IA incluyen la solidez, resiliencia, fiabilidad, precisión, seguridad, protección, privacidad, etc. Puesto que la solidez es una propiedad crucial que supone nuevos desafíos en el contexto de sistemas de IA: la calidad de los datos, la degradación del modelo, la estabilidad de las características de entrada, Precisión frente a *Recall* y perturbaciones en los datos de entrada.

La compresión de los riesgos especialmente relacionados con la solidez del sistema de IA es esencial para su adopción, por lo que se deberán evaluar los riesgos especialmente

ligados a la solidez de los sistemas de IA en su relación con aquellos localizados en el plan de riesgos, especialmente aquellos relacionados con los daños a la salud y los derechos y libertades de las personas físicas.

Aunque la verificación de *software* es una parte esencial en cualquier proceso industrial, el objetivo de asegurar la seguridad y la precisión del *software* usado en todas las partes del sistema hace que el proceso de verificación *software* sea una parte importante de la certificación del sistema. Por tanto, las técnicas usadas en sistemas de IA también están sujetas a validación. Como no todas las técnicas que validan *software* son directamente aplicables a sistemas de IA (como modelos de redes neuronales), nuevos enfoques son necesarios para la verificación de sistemas de IA, abordaremos esos enfoques en esta guía sobre la solidez.

Los cambios que conlleva la transformación digital a nivel de organización deberán contemplar procesos de **gestión del cambio** de cultura para alcanzar los principios de IA confiable. Un ejemplo de metodologías de la teoría de juegos que puede inspirar tales cambios incluye la teoría de diseño de mecanismos (*mechanism design*)[1] [113].

Además, una manera resumida y visual de ver los cambios asociados al cumplimiento de los requisitos de solidez es:

[1] merecedora del Nobel de Economía en 2007, en que el objetivo es obtener resultados deseados de acuerdo con un concepto de solución específico, cuando nos dan una función objetivo y desconocemos el mecanismo.

3. Reglamento de Inteligencia Artificial

3.1 Análisis previo y relación de los artículos

Este artículo establece los requerimientos que deben cumplirse en materia de tres aspectos fundamentales *"Precisión, solidez y ciberseguridad"*. Ciberseguridad y precisión son tratadas de manera específica en sus guías.

En este apartado vamos a realizar énfasis en los párrafos de dicho artículo que están orientados específicamente a la solidez en sistemas de IA de alto riesgo, que se concentran en los apartados 1 y 4.

Artículo 15 Precisión, solidez y ciberseguridad apartado 1 → establece la necesidad de mantener adecuados niveles de precisión durante todo el ciclo de vida del sistema de IA, por lo que indicaremos una serie de medidas destinadas a que los sistemas de IA **no degraden** sus especificaciones de rendimiento y solidez una vez puestos en marcha, durante todo su ciclo de vida.

Los sistemas de IA no deben de presentar problemas de funcionamiento (compatibilidad con antiguas librerías que usen o datos que procesen) ni de calidad en cuanto a la solidez de éstos conforme se usan en el tiempo. Para ello, deberán **alcanzar** y **respetar** niveles mínimos razonables de **solidez** (y métricas asociadas) preestablecidos en los requisitos de estos y concretos a la tarea que resuelven, que garanticen su seguridad y buen funcionamiento.

Artículo 15 Precisión, solidez y ciberseguridad apartado 4 → establece la necesidad de medidas de protección frente a errores, fallos o incoherencias.

Estos aspectos descritos en estos apartados amplían el foco requerimientos del Reglamento Europeo de la IA relativos a la solidez abordando los fallos y sistemas de redundancia como un requerimiento importante. Del mismo modo, el artículo refleja que la solidez para aquellos sistemas que continúan aprendiendo tras su puesta en marcha debe ser considerada de manera específica y con medidas particulares.

El concepto e interpretación de la solidez, en términos de medidas técnicas y métricas de evaluación, es inseparable y dependiente del dominio de aplicación y, por tanto, su finalidad prevista.

3.2 Contenido de los artículos en el Reglamento de IA

AI Act

Art.15 – Precisión, solidez y ciberseguridad

1. Los sistemas de IA de alto riesgo se diseñarán y desarrollarán de modo que alcancen un nivel adecuado de precisión, solidez y ciberseguridad y funcionen de manera uniforme en esos sentidos durante todo su ciclo de vida.

2. Para abordar los aspectos técnicos sobre la forma de medir los niveles adecuados de precisión y solidez establecidos en el apartado 1 y cualquier otro parámetro de rendimiento pertinente, la Comisión, en cooperación con las partes interesadas y organizaciones pertinentes, como las autoridades de metrología y de evaluación comparativa, fomentará, según proceda, el desarrollo de parámetros de referencia y metodologías de medición.

3. En las instrucciones de uso que acompañen a los sistemas de IA de alto riesgo se indicarán los niveles de precisión de dichos sistemas, así como los parámetros pertinentes para medirla.

4. Los sistemas de IA de alto riesgo serán lo más resistentes posible en lo que respecta a los errores, fallos o incoherencias que pueden surgir en los propios sistemas o en el entorno en el que funcionan, en particular a causa de su interacción con personas físicas u otros sistemas. Se adoptarán medidas técnicas y organizativas a este respecto.

La solidez de los sistemas de IA de alto riesgo puede lograrse mediante soluciones de redundancia técnica, tales como copias de seguridad o planes de prevención contra fallos.

Los sistemas de IA de alto riesgo que continúan aprendiendo tras su introducción en el mercado o puesta en servicio se desarrollarán de tal modo que se elimine o reduzca lo máximo posible el riesgo de que los resultados de salida que pueden estar sesgados influyan en la información de entrada de futuras operaciones (bucles de retroalimentación) y se garantice que dichos bucles se subsanen debidamente con las medidas de reducción de riesgos adecuadas.

5. Los sistemas de IA de alto riesgo serán resistentes a los intentos de terceros no autorizados de alterar su uso, sus resultados de salida o su funcionamiento aprovechando las vulnerabilidades del sistema.

Las soluciones técnicas encaminadas a garantizar la ciberseguridad de los sistemas de IA de alto riesgo serán adecuadas a las circunstancias y los riesgos pertinentes.

Entre las soluciones técnicas destinadas a subsanar vulnerabilidades específicas de la IA figurarán, según corresponda, medidas para prevenir, detectar, combatir, resolver y controlar los ataques que traten de manipular el conjunto de datos de entrenamiento («envenenamiento de datos»), o los componentes entrenados previamente utilizados en el entrenamiento («envenenamiento de modelos»), la información de entrada diseñada para hacer que el modelo de IA cometa un error («ejemplos adversarios» o «evasión de modelos»), los ataques a la confidencialidad o los defectos en el modelo.

3.3 Correspondencia del articulado con los apartados de la guía

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

Artículo Reglamento	Requerimiento Reglamento	Sección guía
15.1 párrafo 1	Nivel adecuado y consistente de solidez durante todo el ciclo de vida	Apartado 4.1
15.4 párrafo 1	Resistencia a errores, fallos e incoherencias	Apartado 4.2
15.4 párrafo 2	Planes de redundancia, respaldo y a prueba de fallos	Apartado 4.3
15.4 párrafo 3	Sistemas que siguen aprendiendo tras su puesta en marcha.	Apartado 4.4

4. ¿Cómo abordar los requisitos?

4.1 Evaluación de la solidez

Las métricas de solidez aplicables garantizadas por el proveedor en la documentación deben reflejar y ser una señal de la calidad del sistema. De otro modo, cuando no se puedan garantizar los requisitos establecidos en la documentación en cuanto a nociones o métricas de solidez establecidas mínimas, el proveedor pondrá a disposición del responsable del sistema mecanismos para notificar al humano que se requiere su supervisión (según guía de supervisión humana), tanto para una variación temporal de estas métricas de solidez, como en el caso de una variación persistente en forma de una degradación de la respuesta.

Los **proveedores** del sistema de IA deberán implementar capacidades que lo provean de la solidez necesaria acorde a la finalidad prevista y con el objetivo de mitigar los riesgos encontrados en el plan de riesgos (ver guía de riesgos).

Durante cualquier momento del ciclo de vida válido del sistema, el proveedor deberá:

- Facilitar al responsable del despliegue mecanismos para **observar, supervisar y reportar** diferentes tipos de degradación del modelo (salidas del modelo fuera de rango, posibles ataques por adversarios o perturbaciones que datos o modelo puedan sufrir) que sobrepasen los límites documentados como razonables para cada métrica de solidez, con el fin de hacerlos reproducibles para su corrección.
- Cuando los requisitos de solidez garantizados **cambien, corregirlos** para garantizar los estándares o métricas provistas según la documentación. Cuando irremediablemente esto no sea posible, y se incumplan las garantías documentadas, tanto la documentación como la tarjeta de base de datos como la tarjeta de modelo de IA responsable deberán actualizarse y se deberá notificar a los responsables del despliegue implicados (ver guía de precisión). Si este ajuste se realiza con una actualización, se debe explicar la motivación del cambio en la documentación técnica, tal y como ésta indica para las actualizaciones.

Si surgieran conflictos entre la aplicación de los requisitos del capítulo 3 sección III del Reglamento Europeo de la IA (en definitiva, las guías que acompañan este sandbox), deberá demostrarse que éstos son formalmente el resultado de hacer respetar principios éticos, e informar del conflicto que tiene como consecuencia un grave daño a la solidez del modelo, este tipo de contraprestación debe documentarse adecuadamente, tal y como establece la guía de documentación técnica, indicando los motivos y las explicaciones detalladas de las preferencias realizadas. Todas las partes implicadas deberán velar por resolverlos. Por ejemplo, usando procedimientos de conformidad establecidos como estándar, buenas prácticas y aquellos procesos guiados por certificaciones existentes y venideras.

Las medidas que debe tomar el proveedor en marcha incluirán aquellas que miden la calidad de un sistema basada en su capacidad de mantener cierto nivel de solidez asociado a su precisión bajo cualquier circunstancia [SC42 N483 ISO IEC TR 24029-1]. Para demostrar esta capacidad se pueden hacer análisis estadísticos, pero para probarla se requiere alguna forma de análisis formal (Ver enfoques en [ISO SC42 N938 ISO IEC CD 24029-2], ISO/IEC 25059:2023], método SQuaRE [75] y métodos en [ISO/IEC TR 24029-1:2021 para evaluar la solidez sin el uso de métodos formales).

En este apartado vamos a desarrollar, por tanto, las medidas que son responsabilidad del proveedor, y las vamos a agrupar por aspectos específicos del sistema de IA, para poder poner la solidez en perspectiva para cada una de ellas.

Los siguientes puntos representan una indicación de los temas que se van a abordar, relacionados con la evaluación de la solidez:

- La relación del ciclo de vida y la solidez, considerada como una vertical para el sistema de IA.
- Se presenta un enfoque para seleccionar las métricas de solidez del sistema de IA.
- Se establece el proceso de verificación y validación de las métricas seleccionadas.
- Se considera que tanto eficiencia, como rendimiento, son elementos asociados a la solidez, por lo que proporcionamos indicaciones para como alcanzarlos.
- Finalmente se aborda el aspecto de monitorizar la solidez

Organizativamente, el **responsable del despliegue** deberá tener formación suficiente para detectar cuando se produce una degradación del sistema y así poder retroalimentar al proveedor cuando así lo requiera la documentación del producto. Así, el responsable del despliegue deberá:

1. Conocer los conceptos básicos asociados a la solidez y familiarizarse con la documentación que le da acceso para:
    a. Visualizar cuadros de mando de monitoreo de todas las métricas de solidez. Estos cuadros de mando deben ser integrados en el sistema de AI por el proveedor, pueden desarrollarse utilizando soluciones comerciales, de código abierto o desarrollos propios.
    b. Interpretar y usar la salida del sistema en cuanto a su solidez, a nivel de funcionalidad, modelo y sistema.
2. Acceder a los tutoriales disponibles (en línea, tutoriales, material didáctico adicional) más allá de la documentación del producto para poder comprender la funcionalidad del producto de IA en su totalidad cuando sea necesario para conseguir el punto anterior. Dado que se trata de sistemas de alto riesgo, plantearse planes específicos de formación para el personal cuando la cualificación o la formación de los operadores pueda no ser suficiente.

Debe entenderse que los medios técnicos integrados en el sistema de IA de alto riesgo han sido puestos a disposición del responsable del despliegue por parte del proveedor del sistema, proporcionando conocimiento de estos durante la fase de comercialización y a lo largo de todo el ciclo de vida del propio sistema.

Técnicamente, para que el responsable del despliegue pueda evaluar la salida del sistema de IA en su conjunto, y para una entrada de datos concreta, debe conocer con suficiente nivel de detalle la documentación para poder actuar según la guía de supervisión, si la solidez se degrada bajo los niveles mínimos garantizados. Así, es responsabilidad del responsable del despliegue:

Saber usar y acceder la interfaz de obtención de métricas de solidez y eficiencia computacional asociadas. Para poder usar el sistema, el responsable del despliegue deberá familiarizarse con:

- La interfaz del sistema y de las salidas de este, para cada funcionalidad del sistema, y para el sistema en su conjunto. Dentro de su finalidad prevista y acorde al uso de éste.
- Estadísticas que sirvan para comparar el funcionamiento y métricas del sistema provisto con otros modelos existentes del estado del arte o usados como algoritmo/modelo base.
- Mecanismos de gestión de notificaciones de potenciales errores de funcionamiento documentados, procedimientos de inspección, supervisión, corrección, notificación de potenciales salidas erróneas, o la ausencia de estas.
- Posibles acciones que tomar cuando la salida no esté dotada de explicabilidad coherente con su solidez, o cuando presencie niveles no aceptables de solidez o incertidumbre en la salida.

Además, podrá mejorar su experiencia de usuario estudiando la relevancia de las métricas de solidez aplicables y las elegidas, según el caso de uso, para observar y comprender las diferencias en significancia estadística de los modelos.

1. Saber explorar las métricas de solidez y otras métricas de calidad asociadas a la buena funcionalidad y supervisión de éste.
2. Conocer cómo notificar o corregir potenciales errores de funcionamiento (salidas erróneas, ausencia de éstas, etc.) al proveedor, e inspeccionar datos de entrada y salida.

4.1.1 Ciclo de vida

En cada elemento del ciclo de vida de un sistema de IA de alto riesgo, se establece un nivel de solidez acorde a la finalidad prevista del sistema, estableciendo unas medidas adecuadas. Dependiendo de la fase del ciclo de vida las medidas deberán de ser realizadas por proveedor (diseño, desarrollo) o responsable del despliegue (durante su puesta en producción o funcionamiento), evitando la degradación, y cualquier problema de funcionamiento que aminore la solidez del sistema durante su ciclo de vida.

La solidez de un sistema de IA debe ser verificada durante todo el ciclo de vida definido en este orden:

- **Diseño y desarrollo**. Se deben identificar las características reconocidas y se debe verificar la separabilidad. Se diseñarán experimentos para evaluar las capacidades del sistema de IA en cuanto a un conjunto deliberado de variables a las cuales el modelo debe mostrar solidez. Por ejemplo, se considerarán enfoques de aprendizaje

por conjuntos (técnica que combina múltiples modelos individuales para crear un modelo final, *ensemble learning*) para incrementar la solidez de los modelos, incluyendo la resistencia a ataques adversarios o de evasión. En este aspecto, y relacionado con los datos, podemos considerar el siguiente ejemplo: en un sistema de reconocimiento del habla automático, se demostrará el uso de datos lo suficientemente grandes y diversos, que llevan al modelo a soportar y tratar, adecuadamente, acentos, ruido de fondo y vocabulario técnico.

- **Verificación y validación (V&V).** Se deben verificar las partes cubiertas del dominio de entrada (por ejemplo, con criterios de sensibilidad, o usando métodos formales) y se debe medir el impacto de la perturbación forzando en el proceso de verificación y validación la perturbación intencionada, formal y relacionada con la finalidad prevista adecuada. En el apartado 4.1.3 se presentan diferentes medidas para abordar este aspecto.
- **Puesta en producción / implantación.** En este caso, se puede verificar el impacto de problemas causados por la precisión numérica (compiladores que reorganizan o reemplazan operaciones, reducen precisión numérica o cambian el proceso de redondeo). En estos escenarios, y tal como se ha definido en la guía de ciberseguridad, para evitar los fallos originados por problemas de configuración, los entornos de validación y producción del sistema de IA tienen que ser idénticos, o dotados de los mecanismos que les permitan ser idénticos (como se indica en la citada guía, utilizar tecnologías de contenedores o equivalentes). En el apartado 4.1.6, abordamos en detalle aspectos centrados en la solidez y los sistemas en producción.

4.1.2 Selección de métricas

Hasta ahora hemos visto el concepto de solidez y sus propiedades deseables, en el marco organizativo del sistema de IA. Este marco organizativo es el andamiaje necesario para que sobre él establezcamos el detalle de la operativa técnica sobre como cuantificar la solidez. En sucesivos apartados vamos a abordar como se enfoca la solidez, técnicamente, con metodologías aplicables y métricas asociadas.

La metodología para evaluar la solidez de un modelo se basa en los siguientes pasos:

1. **Establecer requisitos u objetivos de solidez y métricas asociadas**: ¿Hasta qué punto el sistema requiere ser sólido? ¿Cuáles son las características de solidez de interés? Las **propiedades de solidez** demuestran el grado de **generalización** del modelo a datos nuevos, en comparación con su precisión en los datos con que fue entrenado, o con datos esperados en operaciones típicas. Se debe constituir el conjunto de criterios de decisión de propiedades de solidez. Dados estos, se identificarán las métricas que cuantifican los elementos que demuestran conseguir la solidez.
2. **Planificación de experimentos que demuestran solidez**. Éstos se basarán en métodos estadísticos, formales o empíricos, o en la práctica, en una combinación de varios.
3. **Conducir los experimentos:** Los experimentos definidos en el paso 1 deben ejecutarse acorde al plan establecido y se deben registrar los resultados, los

conjuntos de datos utilizados y los valores de salida, con los que se calculan las métricas de manera agregada.

4. **Analizar resultados** contra las métricas establecidas seleccionadas en el punto 1.
5. **Interpretar** los **resultados** para informar la decisión, estos deben ser interpretados no solo en el valor de la propia métrica, si no en relación de su evaluación con todos los experimentos, tanto de manera progresiva (evolución en el tiempo) como agregada (estadísticamente).
6. **Decidir la solidez** del sistema dados los criterios e interpretación identificados antes. Si la solidez es insuficiente, volver a la etapa que alivie sus deficiencias: añadiendo objetivos de solidez, métricas, otras medidas, replantear experimentos o modificar el protocolo de recogida de datos para el experimento o corregir el análisis. Esta iteración se debe repetir hasta que se alcance el objetivo de solidez establecido.

Ejemplo – Sistema de promoción de empleados

Durante la fase de diseño del sistema de IA de cara a establecer los criterios de solidez y cómo estos se van a aplicar al sistema, el proveedor del sistema sigue las siguientes consideraciones, alineadas con los puntos destacados en este apartado:

1. Se seleccionan las siguientes características para el sistema de IA: fiabilidad, estabilidad, sensibilidad y relevancia. De acuerdo con el equipo de diseño, se establecen los umbrales para cada una de ellas, y como se deben de mantener a lo largo del ciclo de vida del sistema. Se generan conjuntos datos de validación y verificación para cada una de las diferentes características, datos que no son usados durante el entrenamiento. Además, se dispone de un grupo de control para todos los demás que también se evalúa frente a esas características.
2. Dado que no se puede desarrollar un método formal adecuado para el sistema de IA (ver apartado 7.2.1) que pueda construir un sistema al que aplicar un optimizador matemático, se opta por una combinación de métodos estadísticos y empíricos para realizar los experimentos
3. Se establece el plan de experimentación y los mecanismos de registro. El proveedor opta por mantener dentro de su sistema de control de versiones, asociado a las versiones del modelo, el acumulado de los experimentos para cada paso de las iteraciones de validación y verificación.
4. El análisis de los resultados se realizará utilizando las técnicas **AUC** y **Lift.**
5. Con el análisis se interpretan los resultados en relación con los valores de especificaciones definidos en el punto 1.
6. Se iterará hasta alcanzar la solidez deseada en el proceso de validación y verificación.

Aunque las especificaciones pueden implementarse de formas diferentes, durante el desarrollo se podrá decidir, por ejemplo, algunos valores específicos de umbrales para finalizar la optimización. Este valor no es necesariamente especificado de antemano y, por tanto, el sistema será más o menos preciso, dependiendo de los datos. Es por esto por lo que es necesario validar el modelo testeando en datos concretos para observar cómo se comporta (**validación del rendimiento**), y estos datos dependerán de la finalidad prevista.

Para la selección de métricas, el 7.1 Anexo I: Métricas y pruebas para sistemas de IA, se presentan una lista de referencia que puede ser consultada, con la consideración que el proveedor deberá seleccionar la métrica más adecuada para la medición de la solidez siempre considerando la finalidad prevista y los riesgos.

4.1.3 Validación y Verificación

La evaluación de la solidez requiere de la intervención de dos pasos muy importantes que se pueden identificar claramente, **validación y verificación**:

- **Verificación**: Confirmación, mediante la provisión de evidencia objetiva, de que se han cumplido los requisitos especificados (ver ISO/IEC 25000:2014 - *Systems and software engineering* 4.43 e ISO/IEC 25030:2019 - *Systems and software engineering* 3.22) de acuerdo con la forma en que se especificaron[2]. Esta fase, no requiere ejecutar el programa con datos reales y se ejecuta antes de la validación.
- **Validación**: Confirmación, a través de la provisión de evidencia objetiva, de que se han cumplido los requisitos de la finalidad prevista del sistema de IA (ver ISO/IEC 25000:2014 - *Systems and software engineering* 4.41 e ISO/IEC 25030:2019 - *Systems and software engineering* 3.21). La validación incluye probar el sistema de IA, previamente verificado, utilizando conjuntos de datos reales, ejecutar el código para validar que hace lo que debería hacer y funciona como se esperaba.

La mayoría de los tipos de datos procesados por diferentes arquitecturas, por ejemplo, de redes neuronales, pueden ser analizados por al menos un método formal (ver ISO/IEC 24029-2:2023. Artificial intelligence (AI) – Assessment of the robustness of neural networks – Part 2: Methodology for the use of formal methods). Cada método, con sus ventajas e inconvenientes (principalmente la escalabilidad, etc.) puede atajar uno o más criterios de entre los deseados en el modelo de evaluación de solidez: estabilidad, sensibilidad, relevancia o alcanzabilidad. En el 7.1 Anexo I: Métricas y pruebas para sistemas de IA, se ofrece una serie de mecanismos que pueden ser utilizados para el análisis de estos tipos de datos.

Ejemplo – Detección de denuncias falsas

Para ampliar los mecanismos de verificación y validación del modelo de detección de denuncias falsas se establece un método formal incompleto de caja negra y determinista (ver en el Anexo I).

Para ello entre las denuncias establecidas como ciertas y las denuncias establecidas como falsas, se extrae un grupo de control. Del conjunto total original, se extrae un listado de palabras relacionadas con ambos tipos de denuncias. Con esta lista de

[2] Cumpliendo especificaciones especificadas, verificaciones tempranas de errores de programación, verificación de programas y documentos a través de un recorrido del sistema con un tutorial paso a paso, etc.

palabras, para cada denuncia de este grupo de control, se establece el conteo estadístico de frecuencia de aparición.

Con estos datos se construye una función que proporciona en relación con la distribución de palabras, el resultado exacto para cada una de las denuncias del grupo de control.

En la fase de validación y verificación se comparan las predicciones del sistema de IA, con las predicciones de la función establecida, tanto para ejemplares del grupo de control como fuera de éste. De esta manera este método formal **complementa** el proceso de verificación y validación.

4.1.4 Eficiencia en sistemas de IA

Los sistemas de IA de alto riesgo pueden ejecutarse en entornos con diferentes capacidades de cómputo, tanto para aquellos que se encuentran en etapa de inferencia como para aquellos que continúan aprendiendo con el tiempo, la capacidad de cómputo disponible va a afectar profundamente a la solidez del sistema.

Es muy importante, por tanto, que las métricas escogidas para validar las características de la solidez se validen y verifiquen (ver apartado 4.1.3) en entornos *hardware* que repliquen exactamente las capacidades de cómputo (memoria, CPU, precisión numérica, velocidad de procesado etc.) finales a las que tendrá acceso el sistema de IA. Si el sistema va a ser instalado por el **responsable del despliegue** las instrucciones de instalación, deberán tener claramente indicadas las necesidades de ese *hardware* y su relación directa con la precisión del sistema.

En aquellos casos en los que además puedan existir restricciones de cómputo o memoria, proporcionamos ejemplos de medidas:

1. Aprendizaje máquina en el filo, cuando se requiera obtener precisión y solidez con recursos computacionales limitados, tales como ejecutar un modelo en un dispositivo móvil como un teléfono en tiempo real o incluso en dispositivos IoT (sensores inteligentes) u otros dispositivos (cámaras).
2. **Destilación** [47], cuando existen limitaciones en la memoria del sistema que ejecutará el sistema de IA.
3. **Aprendizaje privilegiado destilado** [92], cuando no se dispongan de todas las modalidades de datos que se usaron durante el entrenamiento en tiempo de inferencia.
4. **Aprendizaje continuo** [25], cuando el modelo no pueda crecer en arquitectura y número de parámetros conforme aumenten las tareas de aprendizaje, y deba mantener niveles mínimos de precisión para todas las tareas para las que fue entrenado.
5. Técnicas de compresión y poda de redes (cuando la memoria ocupada por los modelos se exceda de un umbral de memoria) u otros.

4.1.5 Rendimiento en sistemas de IA

Las métricas de solidez asociadas al *hardware* y rendimiento del sistema deben describirse en la documentación del sistema antes de diseñar las pruebas de precisión y rendimiento de la etapa de ejecución del sistema. De esta manera se establecerá el rendimiento y eficiencia objetivo para, por ejemplo, tratar un número específico de consultas o solicitudes por hora.

- Se deberá considerar la solidez de la precisión, rendimiento y las estrategias de su implementación dentro del sistema de IA como un todo, y cómo están restringidas por limitaciones en la disponibilidad de recursos, tales como la memoria o energía. Por ejemplo, una métrica puede ser el rendimiento por potencia en vatios o rendimiento por vatio (6.6.5 ISO SC42 N1011), aunque julios por vatio o nano julios por píxel puede ser más intuitivos para calcular estadísticas agregadas [38]. A nivel de *hardware*, la capacidad de memoria requerida (por ejemplo, en Python, se puede usar Memray para el seguimiento de reservas de memoria y producir informes).
- Métricas de complejidad computacional: pueden medir la degradación en la parte de solidez. Por ejemplo, en modelos de aprendizaje automático, se traducirían a latencia de inferencia (por ej. latencia de clasificación, eficiencia de clasificación, rendimiento de clasificación y consumo de energía).
- Métricas de rendimiento: como capacidad de procesado en cortos periodos de tiempo, por ej. en FLOPS (*Floating Point Operations Per Second*), o eficiencia, capacidad de procesamiento mantenida durante una unidad representante de un largo periodo de tiempo, por ej. en FLOPY (*Floating Point Operations Per Year*). Por ejemplo, aunque la latencia se suele usar para medir interacciones de cara al cliente, y rendimiento en aplicaciones de servidor, en IA deben adaptarse al problema concreto, por ejemplo, en un clasificador de imágenes, la eficiencia puede medirse en el número de imágenes procesadas en un segundo para un tamaño de lote mucho mayor de 1, o en latencia cuando el tamaño de lote es uno (la latencia es inversamente proporcional a la eficiencia).

Otras consideraciones adicionales que contribuyen a alcanzar la solidez adecuada del modelo se basan en el *hardware* y su monitorización durante diferentes puntos del ciclo de vida del modelo de IA:

- *Hardware* para el uso de arquitecturas más complejas. Para alcanzar los valores deseables de las métricas de precisión y solidez especificadas en los requerimientos del sistema, los proveedores podrán utilizar medidas de aceleración de cómputo para *hardware* especializado, unidades de procesamiento gráfico (GPUs), circuitos integrados de aplicación específica, o conjuntos de instrucciones embebidos en unidades de procesamiento central (CPUs)[3].

[3] Otros aceleradores pueden aplicarse a funciones simples (como la multiplicación de matrices) o complejas (una función ResNet). Además, debido a restricciones de medidas de ahorro debido a recursos energéticos, la prolongación de entrenamiento de ciertos modelos podría favorecer la precisión sobre otros entrenados durante menor tiempo.

- El *hardware* y arquitecturas asociadas a un nivel de precisión y solidez deben ser transparente, fijo, y actualizado en su caso, en la documentación establecida (según guía de Documentación Técnica).
- Se deberán considerar las posibles incoherencias en el sistema debido a diferencias latencias[4] de clasificación.
  - Reforzamos de nuevo una continuidad exacta entre entornos *hardware* de desarrollo, verificación y validación, que permita que las métricas obtenidas sean entendibles en contextos de idéntica capacidad de cómputo y arquitectura. En aquellas situaciones en las que no sea posible disponer del mismo entorno entre fases del ciclo de vida, **será necesario**, establecer unas condiciones de **solidez específicas** para cada entorno, de cara a permitir el paso de diseño a verificación y de ésta a validación. En ese caso los valores y consideraciones objetivo (definidas como especificaciones) deben estar establecidas en el hardware y su arquitectura del sistema final. De nuevo recordamos que, si el sistema va a ser instalado en las instalaciones del responsable del despliegue, en cualquier formato de comercialización posible, se deben definir claramente unos requisitos mínimos (e idealmente recomendados) en los que se pueden garantizar los requisitos de solidez.

Ejemplo – Sistema de promoción de empleados

El sistema de promoción de empleados en uno de sus formatos de comercialización se entrega al responsable del despliegue para que este realice la instalación de éste. El proveedor decide establecer una serie de métricas adicionales a la solidez: latencia de una evaluación individual y capacidad de procesamiento por segundo de evaluaciones, considerando en el diseño del sistema que estos dos elementos son claves para evitar que un procesado demasiado lento pueda perjudicar la posición en la clasificación de una persona por retraso en el cálculo de su puntuación.

Dado que estos dos puntos son fuertemente dependientes de la relación con el *hardware*, y durante las fases de diseño y validación/verificación no se tendrá acceso a los mismos recursos *hardware* que durante la producción:

- Se establecen los valores mínimos para el *hardware* de desarrollo y preproducción, que deben alcanzarse siempre en cualquier iteración del algoritmo.
- Se establecen unas especificaciones mínimas de *hardware* sobre las que se alcancen los valores esperados en un entorno productivo. Estas pruebas se ejecutan siempre que se considera una versión lista para producción y son condición necesaria para que pueda darse la versión como correcta (siempre unida a otros elementos significativos de la solidez, no como condición única).
- El sistema incluye un registro especial de las dos características (latencia y evaluaciones por segundo) durante su ejecución para informar de su posible

[4] La latencia de clasificación mide la duración entre la entrada de datos del usuario al modelo y la salida de la inferencia del modelo, y determina la calidad (en este caso, la precisión y solidez) del servicio provisto, la cual debe caer dentro de un intervalo paraguas que defina las restricciones operacionales de la implementación de este, y reporte cuando la ejecución de la inferencia del modelo sobrepasa límites no razonables (ISO SC42 N1011).

degradación, tanto por cambios en los datos, como por modificaciones en las volumetrías iniciales.
- Se genera en las instrucciones del sistema las indicaciones de los requisitos mínimos y las volumetrías consideradas, para que el usuario pueda establecer unos requisitos *hardware* en sus instalaciones que le garanticen el funcionamiento del sistema de IA.

4.1.6 Monitorización de la solidez

Los sistemas de IA son artefactos *software* con una naturaleza muy especial, eso hace que su comportamiento a lo largo del tiempo pueda variar, no solo en aquellos casos en los que el sistema continúa aprendiendo con el tiempo, si no en general en todos los escenarios.

La aparición de espacios de dominio no esperados en los datos de entrada, o incluso la existencia de errores o fallos no detectados que pueden afectar a la precisión, solidez o incluso a la finalidad prevista causando daño a las personas o afectar derechos y libertades, son aspectos que no pueden ser descartados.

Este tipo de monitorización encaja en el escenario de una monitorización post comercialización (consultar dicha guía).

Centrado en la solidez, un monitoreo adecuado involucra observar estadísticas, distribución de los datos y cambios en el uso del negocio. Otras nociones para tener en cuenta son:

- Se deberán proveer mecanismos de retroalimentación para distinguir entre errores del responsable del despliegue haciendo mal uso del sistema de los del propio sistema siendo inflexible para satisfacer las necesidades del responsable del despliegue (errores del sistema), o el sistema haciendo suposiciones erróneas sobre el responsable del despliegue (Errores de contexto).
- Se anotarán los tipos de errores en un proceso de resolución de problemas que incluya un registro de problemas, su razonamiento o causa y tipo de arreglo, permanente o a corto plazo y su efectividad (y según guía de registros) y se documentará siguiendo las indicaciones para cambios y actualizaciones tal y como se indica en la documentación técnica.
- Los proveedores deberán implementar la capacidad de auditar los modelos basados en principios éticos (por ejemplo, [57]), de forma que se señalen las tensiones clave responsables bajo las consecuencias éticas, y reconocer éstas como compromisos. Para ello, se deberán usar metodologías de evaluación de solidez (basadas en la evaluación de la Estabilidad, sensibilidad, Relevancia o Alcanzabilidad [ISO/IEC 24029-2:2023 SC42 N938 ISO IEC CD 24029-2]), o la evaluación de conformidad para guiar los procesos que preservan los principios de IA responsable de CapAI [57].

Existen diferentes plataformas que permitan realizar este control o seguimiento, tanto comerciales como de código abierto. El proveedor puede seleccionar cualquiera de ellas para formalizar este seguimiento, u optar por un desarrollo propio siempre que la solución elegida cumpla con los requisitos del Reglamento Europeo de Inteligencia Artificial.

La solidez no se puede entender sin estar relacionada con la supervisión y la transparencia que establecen como se comunica la información del estado del sistema y el impacto que esta tiene en la toma de decisiones, se recomienda consultar o revisitar las guías respectivas de estos aspectos. Es importante considerar se debe **informar** a un **experto en el dominio** de la finalidad prevista del sistema, que no necesariamente pueda tener conocimientos de IA.

4.2 Solidez y resistencia a errores

En el apartado introductorio de la guía hemos visto, apoyados en las consideraciones que proporciona el Reglamento Europeo de IA, que el sistema debe disponer de **solidez técnica** (tal y como lo indica el considerando (75) del Reglamento) para evitar que los fallos, errores o incoherencias puedan tener consecuencias para la seguridad, o afectar de manera negativa derechos fundamentales.

Igualmente, a lo largo del apartado 4.1, se han desarrollado las medidas organizativas y técnicas necesarias para establecer la solidez del sistema y como medirla en el tiempo.

Sobre este aspecto, es el propio Artículo 15 del Reglamento Europeo de la IA en el párrafo 4, el que indica que:

AI Act

Art.15.4 – Precisión, solidez y ciberseguridad

Los sistemas de IA de alto riesgo serán lo más resistentes posible en lo que respecta a los **errores, fallos o incoherencias** que pueden surgir en los propios sistemas o en el entorno en el que funcionan, en particular a causa de su interacción con personas físicas u otros sistemas. Se adoptarán medidas técnicas y organizativas a este respecto.

Por tanto, es evidente que es necesario, asociado al concepto de solidez, establecer estrategias de predicción de fallos, consecuencias no intencionadas de efecto negativo y efectos que afecten a la seguridad de las personas o que afecten de manera negativa derechos fundamentales tendrán que ser estudiadas y valoradas para evitar su alcance en posibles mal ejecuciones del sistema de IA.

El proveedor es el responsable de poner en marcha las medidas destinadas a mitigar la falta de solidez en sistemas de IA frente a errores, fallos e incoherencias.

Organizativamente, el proveedor:

- Facilitará la implementación de técnicas de alerta (señalando qué casos especiales necesitan cambios en las reglas para corregir esas decisiones en el futuro) a las partes responsables o, en último caso, si no se recibe respuesta de éstas en un tiempo establecido como seguro, se implementará funcionalidad que permita el apagado automático del sistema. En el caso específico de solidez que nos ocupa, el sistema deberá estar preparado para cuando la solidez este comprometida, lanzar dichas alertas. La implantación automática basada en IA debe usar protocolos de fallos seguros, como por ejemplo permitir en cualquier momento la intervención del humano. Este aspecto se detalla en la guía de supervisión humana.
- Introducir protocolos de fallos seguros. El desarrollo de tecnologías de IA con protocolos de fallo seguro debe permitir a los humanos predecir percances catastróficos, aunque su aparición sea rara, deberá reservar una planificación de diseño que conllevará un dedicado esfuerzo a intentar predecir futuros riesgos y trampas en que puede caer el sistema.

Medidas **técnicas** para preservar la solidez como resistencia a fallos, errores o incoherencias técnicas incluyen los aspectos que se detallan a continuación.

Previo al desarrollo del sistema de IA, se deberán establecer diálogos con diferentes partes involucradas, maximizando la diversidad e inclusión de diferentes perfiles del dominio durante el diseño, desarrollo, mantenimiento, aplicación, supervisión y uso del sistema. El objetivo es prevenir consecuencias catastróficas del posterior uso del sistema cuando haya sido implantado.

Una aproximación de este enfoque es elaborar diferentes comités de evaluación del diseño del modelo que serán involucrados, con diferentes puntos de vista de expertos con diversa trayectoria, para poder predecir (y atajar, en fase de diseño), incoherencias de diseño o implementación del sistema que puedan conducir a consecuencias inesperadas no deseadas. En pymes y empresas de nueva creación, estos comités pueden enfocarse a grupos de trabajo de toda la compañía donde se pueda evaluar una operativa transversal del proceso.

El pensamiento para instaurar consiste en la idea de la "*IA segura basada en el sufrimiento*, a favor del fallo seguro": Es más fácil esforzarse en evitar desencadenamientos muy trágicos que podrían ocurrir el desplegar un sistema de IA dado, que intentar asegurarnos de que todos los aspectos del sistema funcionarán siempre correctamente y como es esperado [40].

Otro enfoque es el clásico en IA segura del compromiso o contraprestación entre la probabilidad de éxito y la controlabilidad. En este caso, si es difícil controlar los valores del resultante agente artificial, entonces el riesgo que se supone es un callejón sin salida, mientras que en la IA del fracaso seguro no requiere preocuparse de la controlabilidad del sistema al mismo nivel, mientras haya suficiente controlabilidad para predecir que se pueden evitar.

En la IA segura centrada en el *sufrimiento*, se proponen estrategias de intervención tales como:

- Influenciar las salidas del sistema con IA controlada, a través de la propagación de los valores fundamentales del mismo, o estrategias de IA segura dirigidas como el **progreso diferencial** para alejar diferencialmente al sistema del peor caso. También identificar IA segura donde más se necesita, identificando el peor evento que podría ocurrir por un fallo controlado, y trabajar hacia identificar un control general o mecanismos de apago del sistema.
- Diseñando maneras de hacer fallar al sistema de IA con fallos agraciados o "benignos", de manera que hubiese aún peores fallos que el diseño de este fallo agraciado[5].

Ejemplo – Sistema de promoción de empleados

Durante el análisis de riesgos del sistema de IA se ha considerado como un riesgo que el sistema no cuente con la misma cantidad de información para todos los empleados, lo que puede causar que la evaluación no se realice correctamente causando que aquellos empleados que tengan menos información dentro del sistema resulten discriminados. Los errores que pueden causar que la información no se incorpore en el sistema para un empleado dado, son muy variados, y de difícil inventario pues el sistema se integra con varias fuentes.

Para subsanarlo se considera **la posibilidad de falta de información** (es decir el fallo en la entrada de datos) como posible para todos los empleados. Para ello se crean dos formas de tratar con este *potencial error*:

- El sistema elabora un listado de todos los empleados que tienen menos información que la media, y elabora alertas sobre ellos para que pueda ser revisado por el personal y subsanada la falta de información, por cualquiera de los orígenes.
- El sistema es capaz de proporcionar una lista **alternativa** de valoración de empleados, en los que se añade una valoración sobre la información faltante para aquellos empleados que no está completa, basada en un análisis de los datos de aquellos que, si cuentan con la información completa, para disponer de una aproximación del estado de promoción para aquellos empleados con fallo de incorporación de información.

Organizativamente, los responsables del despliegue deberán familiarizarse con casos pasados y las lecciones aprendidas en casos de uso similares reportados, documentados y mantenidos al día por el proveedor en la documentación actualizada del sistema.

[5] La implementación de valores seguros del sistema puede hacerse preguntándose si la corregibilidad reduce los riesgos de que el sistema provoque sufrimiento [41].

El responsable del despliegue es el responsable de utilizar los canales proporcionados por el proveedor para notificar adecuadamente la presencia de errores, dado que vigilarlos es su responsabilidad. Esto implica, por tanto, que el proveedor debe proporcionar medios para gestionar la recepción de notificación de la presencia de errores, por parte del responsable del despliegue.

Técnicamente, el responsable del despliegue deberá estar preparado para:

1. Conocer o disponer -y formar, si es necesario- de personal que pudiera cumplir con lo establecido por el proveedor los conceptos y posibles implementaciones de "fallo seguro".
2. Actuar ante la falta de solidez como concepto de degradación del modelo.

4.3 Redundancia y planes de respaldo

El camino de conocer, garantizar y conservar la solidez del sistema como característica del sistema, tiene la derivada adicional de disponer de técnicas de redundancia para el funcionamiento del sistema de IA, de tal manera que esta pueda garantizarse.

AI Act

Art.15.4 – Precisión, solidez y ciberseguridad

La solidez de los sistemas de IA de alto riesgo puede lograrse mediante soluciones de **redundancia** técnica, tales como **copias de seguridad** o **planes de prevención contra fallos.**

Por tanto, una parte de los mecanismos para alcanzar la solidez se fundamentan en las técnicas de redundancia, que incluyan planes de respaldo o a prueba de fallos.

El **proveedor** el principal responsable de establecer las técnicas de redundancia en el Sistema de IA y que estos estén acorde a la finalidad prevista del sistema y prestando atención también a los resultados no deseados previsibles.

El proveedor deberá documentar y facilitar el monitoreo del nivel de cobertura de las pruebas de calidad realizadas sobre el sistema (ver guías de ciberseguridad, precisión y esta misma, así como la guía de sistema de gestión de calidad), análisis de correlación y redundancia que puedan desvelar potenciales ataques al sistema o la degradación de la calidad de este. Ejemplos de técnicas a utilizar son: **correlación de controles y medidas**, por ejemplo, restauraciones de datos, del Sistema, medidas KPI (número de restauraciones de datos incorrectas), grado de cobertura de defensas preventivas, test de cobertura, pruebas de calidad y restauración de datos.

El proveedor proveerá acciones para el diseño continuo del sistema basado en la salvaguarda de riesgos, para avalar el sistema de IA con resiliencia a ataques, solidez a cambios en el entorno operativo, minimizar el daño inesperado no intencionado, y reproducibilidad.

El proveedor deberá proporcionar:

1. Mecanismos de **copia de datos automático** (según guía de registros) que permitan la redundancia de modelos, algoritmos, logs o registros provistos por las librerías de MLOps y DevOps, datos y ejecuciones del modelo.
2. Mecanismos para el **fallo seguro** de los mismos componentes durante todo el ciclo de vida del sistema de IA.
3. Herramientas que permitan ejecutar **un plan de actuación** cuando se haya producido el fallo, para la recuperación efectiva de datos, modelos, algoritmos y el sistema de IA.
4. **Herramientas de reproducibilidad**, trazado (según guía de registros), procedencia o linaje de los datos que garanticen la solidez del modelo utilizado y permitan detectar modificaciones en los datos que lleven a cambios significativos en la solidez o rendimiento del modelo.

Tanto la redundancia como el plan de actuación de los mismos componentes arriba señalados deberá ser local y global; idealmente, utilizando un segundo dispositivo o mecanismo, como por ej., un alojamiento seguro en la nube.

Para ejecutar **técnicamente** los puntos anteriores 1, 2, y 3, se recomienda seguir las siguientes metodologías y herramientas de MLOps:

- Medidas de **redundancia geográfica.**
- **Sistemas escalables en capacidad** según la demanda, sistema de monitorización de disponibilidad y rendimiento.
- Implementaciones de **sistemas de copias de seguridad y seguimiento** de versiones de datos.
- Modelos y algoritmos que se incluyan en **plataformas de DevOps** tales como de todas las existentes en el mercado, tanto comerciales como código abierto (gratuitas). Éstas proveen características clave como automatización de operaciones de desarrollo y distribución, cómputo en la nube centralizado, monitoreo de errores continuo, y tecnología de contenedores para ejecutar microservicios o aplicaciones con recursos limitados en entornos múltiples, que permitirían orquestas los escenarios de redundancia geográfica y la escalabilidad acorde a la demanda.

Ejemplo – Detección de denuncias falsas

El tramitado de denuncias en las comisarías, es un proceso que requiere una continuidad y una disponibilidad en el tiempo, que permita que un ciudadano pueda presentar una denuncia en cualquier comisaria en cualquier momento.

Para garantizar que el sistema de IA puede realizar las operaciones de evaluación de las denuncias, con alta disponibilidad el sistema de IA se encuentra distribuido en CPDs (centros de procesos de datos), con sistemas de redundancia en diferentes ubicaciones físicas, tanto en almacenamiento como en servidores de aplicaciones, para evitar que un fallo en una de las ubicaciones del CPD pueda suspender la ejecución del servicio.

Introducen un sistema de respuesta auto escalable, que permite al sistema aumentar el rendimiento de la base de datos y aplicaciones acorde a la demanda.

Para la integración de todo ello con el proceso de distribución de nuevas versiones, se conecta una herramienta de DevOps de código abierto, que antes de liberar una nueva versión verificada y validada, realiza una batería de pruebas unitarias, de integración y comprueba la precisión del sistema y si todo es correcto distribuye la actualización sobre la infraestructura.

Con estas consideraciones, el sistema de IA tiene una infraestructura que garantiza la redundancia y un mecanismo de despliegue que garantiza que se cumplen los requisitos necesarios para su puesta en marcha establecidos desde el diseño y a través del análisis de riesgos.

En cuanto a los datos (punto 4 anterior), parámetros de configuración o del modelo aprendido, deben poder descargarse para fomentar su reproducibilidad. Para promover la publicación de resultados y el linaje de modelos, cuando los pesos de un modelo, modelos pre-entrenados o conjuntos de datos no sean posibles almacenarlos en el repositorio de código principal que almacene el mismo, se promoverá el uso de herramientas abiertas Open Access y repositorios de confianza.

Para mantener un seguimiento y registro de las bases de datos usadas y las modificaciones que éstas sufren, se podrán usar herramientas gratuitas de control de versiones específicas para bases de datos.

Por su parte, el **responsable del despliegue** en este aspecto deberá conocer las medidas de redundancia que se proporcionan, así como los planes de respaldo y a pruebas de fallos.

Especialmente lo será si el sistema de IA es instalado por el responsable del despliegue en unas instalaciones propias. En ese caso, es responsabilidad del responsable del despliegue replicar con precisión las instrucciones de instalación y mantenimiento que el proveedor haya proporcionado, asociadas a las medidas descritas en el apartado anterior.

Organizativamente, la responsabilidad del usuario será conocer, aplicar y mantener vigiladas las métricas de precisión y estadísticas asociadas, pudiendo reflejar cambios entre las distribuciones de los datos de entrenamiento y de inferencia (pudiendo exhibir desplazamiento de distribución o datos (*shift* del conjunto de datos o *distribución shift*)).

En cuanto a formación **técnica**, el responsable del despliegue tiene la responsabilidad de conocer, aplicar y mantener vigilado el sistema de respaldo de fallos en caso de fallo, si así lo especifica la documentación del proveedor.

4.4 Solidez para sistemas que continúan aprendiendo

El comportamiento de los sistemas de IA que continúan aprendiendo después de ser comercializados, requiere un foco y tratamiento especial para garantizar su solidez especialmente en la posibilidad de aparición de sesgos.

AI Act

Art.15.4 – Precisión, solidez y ciberseguridad

> Los sistemas de IA de alto riesgo que continúan **aprendiendo tras su introducción en el mercado** o puesta en servicio se desarrollarán de tal modo que se elimine o reduzca lo máximo posible el riesgo de que los resultados de salida que pueden estar sesgados influyan en la información de entrada de futuras operaciones (bucles de retroalimentación) y se garantice que dichos bucles se subsanen debidamente con las **medidas de reducción de riesgos** adecuadas.

Es importante considerar que el aprendizaje continuo debe extenderse, conceptualmente en lo que a cobertura de las medidas propuestas se supone, cuando este se **actualice** al haber sido reentrenado siguiendo un proceso de actualización, y no solo los sistemas que continúan aprendiendo de manera automática durante su utilización.

Proveedor y responsable del despliegue deberán asegurar que el entrenamiento constante y continuado en el tiempo del sistema de IA no deteriora sus métricas de solidez, precisión y rendimiento.

El proveedor, como responsable del diseño, implementación, verificación y validación del sistema de IA, es el principal responsable para cubrir los requerimientos que establecemos en este apartado. A nivel de **organizativo** deberán considerar:

- El proveedor del sistema de IA debe tener expertos que puedan cubrir el monitoreo de la degradación del sistema durante todas y cada una de las diferentes etapas por las que el modelo puede sufrir degradación a nivel de científico de datos, ingeniero de datos, así como a cambios en el funcionamiento de las MLOps y DevOps.
- A nivel de diseño y concepción del modelo, estrategias de aprendizaje continuo, compresión de modelos de aprendizaje profundo deberán ser consideradas para atajar el posible olvido catastrófico. Otras medidas de solidez a estudiar en estas etapas de incepción del sistema el diseño de mecanismos (*Mechanism Design* [113, 53, 51]).
- Se deberá asegurar que el entrenamiento constante y continuado en el tiempo del sistema de IA no deteriora sus KPIs y métricas de precisión, solidez, rendimiento, y resto de métricas que alcancen los principios de IA confiable.
- El proveedor deberá indicar técnicas formales de anotación en la documentación del sistema, sobre casos concretos de ejecución del sistema de IA cuando éste deba o tenga la opción de interactuar, interoperar con otros sistemas o humanos, reusarse o integrarse con otros sistemas. Especialmente, se **deberá anotar la retroalimentación** o mejoras **provistas** por **humanos** para que la versión futura del modelo actualizado las tenga en cuenta, dentro del paradigma e IA centrada en el humano, Computación Humana, *Human on the Loop* (HOTL) y *Human in the Loop* (HITL).

El proveedor, en relación con esos datos notificados, deberá mantener pública una base de datos de lecciones aprendidas de fallos, que debe servir para mantener una base de conocimiento actualizada continuamente y que integra la seguridad en el diseño y el paradigma del diseño de salvaguarda. Éste puede evocar, situaciones o visualizaciones de ejemplos que pueden ayudar a discernir y prevenir escenarios no favorables de HRAIS a evitar.

El responsable del despliegue debe tener acceso y estar familiarizado con la funcionalidad del sistema para poder visualizar y reportar, en cualquier momento, si es necesario, al proveedor, de anomalías que no aseguren que el entrenamiento constante y continuado en el tiempo del sistema de IA no deteriora sus KPI y métricas de calidad, solidez, eficiencia, rendimiento e incertidumbre, entre otras.

4.4.1 Tipos de degradación del sistema de IA y planes de mitigación, estrategias y herramientas

La calidad en términos de solidez de un sistema de IA es tan buena 1) como buenos sean los datos (ver Guía de Datos), y 2) como sólido sea el modelo y su infraestructura una vez en producción.

Una vez entrenado el modelo, el proveedor se puede ver obligado a encontrar un compromiso en la contraprestación entre precisión y estabilidad. Sin embargo, un sistema sólido debe evitar la degradación de este, para ello es necesario que para las taxonomías de degradación que puedan existir, podamos trasladar mecanismos de corrección o de detección.

Existen tres tipos principales de posible degradación de un modelo de IA, conforme éste se sigue entrenando o usando para inferencia; Todas deberán controlarse y evaluarse desde la fase de desarrollo, hasta la fase de ejecución durante el ciclo de vida:

1. La **desviación del modelo**: Se refiere a la diferencia entre las predicciones promedio del modelo y los valores reales. Representa el error sistemático que indica la capacidad del modelo para capturar las relaciones subyacentes en los datos. Una desviación alta sugiere que el modelo es demasiado simple y no logra captar patrones complejos, lo que conduce a un subajuste. Para mitigar la desviación, es recomendable aumentar la complejidad del modelo o mejorar la calidad y cantidad de los datos de entrenamiento. Es esencial equilibrar la desviación y la varianza para evitar tanto el subajuste como el sobreajuste del modelo.
2. La desviación o **deriva del modelo en el tiempo**: es un fenómeno en que las predicciones del modelo aprendido se degradan debido a alteraciones en el entorno. Ocurre cuando P(Y|X) cambia: las mismas entradas producen diferentes salidas. Estrategias para mitigar y monitorear este tipo de derivas: la batería de pruebas debe encargarse de detectar fallos del modelo y del sistema silenciosamente, el impacto del cambio en los datos con respecto a los datos de entrenamiento original del modelo y su impacto en la precisión y métricas asociadas.

3. La **desviación de los datos en el tiempo**: ocurre cuando los datos (o variable independiente) de entrada de un modelo (la distribución conjunta, P(X)) cambian. Es una de las mayores razones por las que un modelo degrada su exactitud en el tiempo. Se produce cuando P(X) cambia, o P_t1 (X) ≠ P_t2 (X)). Cuando P(X) cambia, ocurre un **desplazamiento del dominio**.

4.4.2 Estrategias para mitigar cambios en la degradación de la precisión y solidez del modelo y/o de sus datos

Tanto para mitigar la **desviación del modelo como la desviación de los datos,** el reentreno de la red suele ser una solución, pero no siempre viable, especialmente en contextos de aprendizaje continuo.

Caracterizar la desviación del modelo y adaptar el modelo al mismo [58], usar estrategias de desviación del conjunto de datos [60], o de generalización fuera de distribución (**out-of-distribution**, OOD) o **zero-shot generalization** [61]. Para que el modelo continúe aprendiendo y adaptándose al cambio en la distribución, estrategias de aprendizaje continuo pueden ser útiles para la desviación de datos [62] y para aprender continuamente con modelos que no paran de entrenar. Por ejemplo, la utilización de ***Elastic Weight Consolidation (EWC)***, *Piggyback* [112], modelos de ***Experience Replay*** *(replay buffer, prototype experience replay, rehearsal)*, *replay* con modelos generativos (*pseudo rehearsal*), o *Memory networks*, entre muchos otros [25].

Ejemplo – Detección de denuncias falsas

En el diseño del sistema de IA de alto riesgo, el proveedor ha considerado que, aunque la base de entrenamiento es amplia (ver ejemplos en guía de precisión), el modelo puede degradarse por cambios en los datos presentes en las denuncias a lo largo del tiempo pudiendo sufrir: desviación de los datos y desplazamiento del dominio. El concepto que se establece no es que puedan ocurrir, si no que ocurrirán, por lo que se hace necesario mitigarlos (en la medida de lo posible) y proporcionar un mecanismo de retroalimentación que amplie la base de entrenamiento con la evaluación final de un agente humano.

Con el objetivo de mitigar la desviación de los datos y desplazamiento del dominio en las denuncias (tanto provenientes de su contenido estructurado: provincias, direcciones; como del contenido textual) se utiliza la técnica de la adaptación del dominio (*domain adaptation)* propuesta en esta guía sobre los conjuntos de denuncias de entrenamiento y verificación/prueba.

Dado que el reentrenamiento del sistema no es viable de manera continua. El sistema de IA se dota de un proceso que permite a los agentes añadir la veracidad/falsedad real de una denuncia en la que el sistema ha cometido un error, con esta información el sistema de IA prepara un ejemplar de entrenamiento despersonalizado y anonimizado que podrá incorporarse a los conjuntos de entrenamiento en las actualizaciones futuras sucesivas. Esta información se incorpora al sistema de MLOps descrito en esta guía para este mismo sistema de IA.

Para mitigar la desviación de los datos y desplazamiento del dominio, se pueden usar técnicas de adaptación del dominio (*domain adaptation)* [59] o transferencia de conocimiento (*transfer learning)*.

Cuando los conjuntos de datos sean visuales, se deberá además estudiar la potencial ocurrencia de sesgo general, de selección, de marco o de etiqueta y todos sus subtipos (Tabla 1 en [88] y lista de requerimientos asociado en apéndice).

Para detectar **relaciones causales** que puedan ser una **fuente de sesgo**, se pueden utilizar clasificadores para discernir la presencia de ciertos objetos o atributos en un clasificador, para detectar el sesgo de selección [93], a través del *Neural Causation Coefficient* (NCC, [92]).

- Cuando no se tiene acceso al objetivo real: ***Confidence-based Performance estimation (CBPE)*** para modelos de clasificación, y ***Direct Loss Estimation (DLE)*** *para problemas de regresión*.
- Se deberá procurar establecer mecanismos para detectar el fallo silencioso del modelo, por ejemplo, con sistemas de alertas. Por ejemplo: NannyML afirma ser a día de hoy el único algoritmo de código abierto capaz de capturar completamente el impacto de la desviación de los datos en la precisión, y AzureML de Microsoft con el diagnóstico de errores a nivel de MLOps.
- Para estimar relaciones causales que pueden comprometer la solidez, el sesgo de variable omitida que explica las variables dependiente e independiente conlleva a estimaciones sesgadas. Para ello se pueden usar métodos que reducen el riesgo de este sesgo: variable instrumento o *least squares estimator*, o el *regression discontinuity design*.

Para monitorizar el buen conocido fenómeno del **olvido catastrófico** y la incapacidad de aprender de manera continua [25], se recomienda usar modelos de aprendizaje profundo continuo (*continual learning* o *lifelong learning*), considerar y reportar métricas de control del olvido y de monitorización del aprendizaje continuo, por ejemplo, ***Backward transfer, forward transfer, model size efficiency, samples storage size efficiency, y computational efficiency*** [28].

Organizativamente, el sistema de IA deberá implementar funcionalidad que permita el cómputo, monitoreo y notificación automática o por parte del **responsable del despliegue** de las degradaciones a las que pueda estar expuesto el sistema. Por tanto, será responsabilidad del usuario conocer dicho sistema proporcionado por el proveedor, y acceder a la documentación técnica del mismo.

En cuanto a medidas **técnicas**:

- El responsable del despliegue tiene la responsabilidad de conocer, aplicar y mantener vigilado el modelo a través de las interfaces provistas para transmitir la precisión, sesgos (según Guía de Datos) e incertidumbre de este.
- Así, deberá usar los manuales de instrucciones de las interfaces de las herramientas que proveen todas las métricas relacionadas, para estar en posición de interpretar (ver guía de Transparencia), verificar que la salida del modelo se corresponden con su precisión y métricas asociadas, y que la explicación señala las razones correctas

esperadas [113], para así evaluar si el sistema está funcionando correctamente, sin sesgo (evitando por ejemplo alucinación de elementos [114,115], ver Sesgos en Guía de Datos). Una vez identificados los sesgos, actuará en consecuencia según guía de supervisión humana).

Más información sobre medidas para monitorizar y controlar la solidez de sistemas de IA en la ISOS:

- [SC42_N483_ISO_IEC_TR_24029-1] ISO/IEC TR 24029-1, Artificial Intelligence (AI) — Assessment of the robustness of neural networks — Part 1: Overview
- [ISO/IEC 24029-2:2023 SC42 N938 ISO IEC CD 24029-2], Artificial intelligence (AI) — Assessment of the robustness of neural networks — Part 2: Methodology for the use of formal methods
- (ISO/IEC 24029-2:2023, ISO /IEC JTC 1/SC 42/WG 3 Secretariat: ANSI Date: 2021-09-06).

5. Documentación técnica

Tan importante como la propia solidez, sus métricas asociadas y los mecanismos necesarios para garantizarla, es la documentación de la misma. La documentación técnica, tal y como se detalla en la guía correspondiente, debe contener la siguiente información, al menos, en relación con la solidez:

- Las instrucciones sobre cómo obtener e interpretar las medidas y métricas de solidez, incluyendo su explicación y monitorización, deben documentarse según la guía de Documentación para todos los posibles usuarios del sistema que deban monitorizar, desarrollar o intervenir en el sistema.
- Para que la solidez reportada sea accionable, tal documentación debe incluir los rangos posibles de los parámetros configurables del modelo, rangos de datos de entrada y salida, así como las medidas de latencia y eficiencia estimadas para obtener la solidez establecida.
- Los valores mínimos aceptables para las métricas de solidez aplicables (y garantizadas por el proveedor en la documentación) deben reflejar y ser una señal de la calidad del sistema. Cuando éstas no se puedan garantizar, el proveedor facilitará mecanismos al responsable del despliegue para notificar al humano (según guía de supervisión humana) y/o posible interrupción/apagado del sistema.
- Dado que la solidez debe mantenerse a lo largo de todo el ciclo de vida del sistema, la documentación deberá actualizarse tras cada modificación relevante del modelo, reentrenamiento o cambio de contexto operativo, reflejando las evidencias y resultados de las nuevas evaluaciones realizadas.
- Además, la documentación de la solidez debe explicar a los diferentes responsables del despliegue el funcionamiento del sistema en cuanto a desafíos comunes entre los principios de IA responsable. Por ejemplo, indicando garantías mínimas que provee el sistema para conocidos compromisos: entre imparcialidad global de grupo y privacidad individual (ya demostrado), el potencial compromiso entre solidez y privacidad. Puesto que estos compromisos entre requisitos han demostrado la posibilidad de estar en conflicto, la tarjeta del modelo de IA deberá afirmar las mínimas garantías esperadas, así como aquellos compromisos que se hayan probado pueden o no garantizarse.
- Para testear el modelo como caso *sandbox* se recomiendan metodologías con las que el usuario deberá familiarizarse, especialmente si éste se convierte en proveedor reusando y modificando el producto: como la implantación canaria (donde el modelo se pone en producción para un pequeño segmento de usuarios, siendo monitorizado y controlado ante posibles problemas señalados por pruebas de integración y regresión).
- Se deberá reportar junto con las fichas del conjunto de datos y cartas del modelo, la *Responsible AI scorecard* como base aglutinadora de los principios de IA responsable que afectan la solidez del sistema.

En cualquier caso, la citada guía de documentación establece como organizar dicha documentación y como relacionarla con las medidas descritas en esta guía. Para disponer

de un escenario completo y por tanto un marco de trabajo correcto, se recomienda consultar ambas en conjunto para establecer cómo documentar la solidez del sistema.

5.1 Certificación de la solidez de modelos

Previamente a la puesta en operación y tras la implantación del modelo, se podrá certificar la solidez del modelo. Dada finalidad prevista del sistema de IA y tipo de datos que usa, se recomienda ver tipo de ataques y defensas (en ese particular, la guía de ciberseguridad realiza una revisión de los escenarios aplicables).

En términos de metodologías, y que por tanto pueden ayudar al proveedor a documentar adecuadamente la solidez del sistema de IA, se pueden establecer mecanismos para establecer ciertas garantías de solidez. Para ello, se pueden considerar las nuevas licencias y aquellas bajo preparación:

- Hippocratic License https://firstdonoharm.dev/
- RAIL (https://www.licenses.ai)
- Ethicas foundation (certificación/sello)

Ejemplos:

- Para estabilizar la solidez de modelos que procesan imagen contra distorsiones como compresión, re-escalado o corte (*cropping*) se puede usar entrenamiento de estabilidad (*Stability Training* [79]).
- Para evaluar y robustecer modelos de procesamiento de lenguaje natural, se puede usar aumento de datos y entrenamiento adversarial con *TextAttack* [12] [13].
- Para funciones objetivo-suaves como la *population loss*, se puede usar entrenamiento adversarial por principios, por ejemplo, considerando una formulación de penalti Lagrange perturbando la distribución de los datos en una bola *Wasserstein*.

Según el momento del ciclo de vida, la certificación de la solidez requiere de diferentes medidas organizativas.

1 Operación y monitoreo. Es fundamental evaluar la solidez de los sistemas de inteligencia artificial (IA) en su ámbito operativo y supervisar su evolución. Cuando no se alcanza el nivel de robustez requerido, se deben implementar acciones correctivas, como alertar al operador o activar modos de seguridad para prevenir fallos. Los métodos formales pueden aplicarse en este contexto; sin embargo, suelen demandar mayores recursos de procesamiento, memoria y energía en comparación con otros enfoques.

- Para hacer las explicaciones en el tiempo más robustas se puede usar aumento de datos contrafactual para entrenar el modelo con muestras de datos y sus contras fácticas y así evitar eventos contrafactuales desafortunados (***unfortunate counterfactual events***: UCEs [82]) o usar contra

fácticos plausibles [85, 86], los cuales no solo proveen mejor solidez que explicaciones contra fácticas más cercanas al ejemplo a explicar, pero también suponen una mayor imparcialidad individual [85].

- Para explicar la deriva de los datos de entrada en el sentido de diferenciales de datos, se pueden usar ontologías e inducción de conceptos sobre conocimiento de fondo [87].

2 Reevaluación: El monitoreo de las pruebas de validación y verificación de los principios éticos de la UE podrían hacerse a través de los requisitos establecidos por licencias actualizadas regularmente.

6. Cuestionario de autoevaluación

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

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

7. Anexos

7.1 Anexo I: Métricas y pruebas para sistemas de IA

7.1.1 Métricas para establecer la solidez

Además de las identificadas en la guía de precisión, se podrán reportar métricas específicas de solidez de los modelos de inteligencia artificial:

Clasificación multiclase:

- **Coeficiente Kappa de Cohen**: Mide el acuerdo entre anotadores, ajustando por la probabilidad de acuerdo aleatorio.
- **Matriz de confusión y métricas derivadas**: Proporciona una visión detallada del rendimiento del modelo, incluyendo verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos.
- **Hinge Loss**: Utilizada principalmente en máquinas de soporte vectorial, evalúa la pérdida basada en el margen de clasificación.
- **Coeficiente de correlación de Matthews (MCC)**: Ofrece una medida equilibrada del rendimiento del modelo, especialmente útil en conjuntos de datos desbalanceados.

Es importante destacar que, en conjuntos de datos desbalanceados, métricas como el Área Bajo la Curva ROC (AUROC) pueden no ser adecuadas, ya que no reflejan correctamente la proporción de falsos positivos y verdaderos positivos.

Clasificación binaria:

- **Curva Precisión-Recall**: Útil cuando las clases están desbalanceadas, ya que se centra en la capacidad del modelo para identificar correctamente las instancias positivas.
- **Curva ROC**: Muestra la relación entre la tasa de verdaderos positivos y la tasa de falsos positivos a diferentes umbrales.
- **Exactitud balanceada**: Promedia la sensibilidad y la especificidad, proporcionando una medida más equilibrada en casos de clases desbalanceadas.

Clasificación multietiqueta:

- **Puntuación F1**: Combina precisión y recall en una sola métrica, siendo especialmente útil cuando se manejan múltiples etiquetas por instancia.

Sistemas de puntuación:

- **Área Bajo la Curva (AUC)**: Evalúa la capacidad del modelo para distinguir entre clases.

- **Lift**: Mide la mejora de las predicciones del modelo en comparación con una estrategia aleatoria.

Métodos estadísticos:

- **Error Cuadrático Medio (RMSE)**: Calcula la raíz cuadrada del promedio de los errores al cuadrado, proporcionando una medida de la magnitud del error.
- **Error Máximo (MaxError)**: Indica la mayor diferencia entre los valores predichos y los reales.
- **Correlación entre valores actuales y predichos**: Evalúa la relación lineal entre las predicciones del modelo y los valores reales.
- **Validación cruzada**: Incluye técnicas como K-fold, Group K-fold, Leave-One-Out, Leave-One-Group-Out y validación cruzada de Monte Carlo, que ayudan a evaluar la generalización del modelo.

Además, para medir la robustez asociada a la precisión del modelo, se pueden utilizar:

- **AUROC**: Aunque debe emplearse con precaución en conjuntos de datos desbalanceados. [ISO SC42 N1011]
- **Curva de Respuesta Acumulativa (CRC)** o **Curva de Ganancia**: Alternativas a la curva ROC cuando el objetivo se centra en una proporción específica de la base de datos. [7]
- **Hamming Loss**: Mide la fracción de etiquetas incorrectamente predichas en problemas de clasificación multietiqueta; valores más bajos indican mayor robustez.
- **Esquemas de validación cruzada 5x2**: Proporcionan una evaluación más robusta del rendimiento del modelo.

La selección de métricas debe alinearse con los objetivos específicos del modelo y las características del conjunto de datos, garantizando una evaluación justa y precisa de su rendimiento y robustez.

7.1.2 Métodos de prueba

Los métodos estadísticos confían en procesos de pruebas matemáticas capaces de ilustrar cierto nivel de confianza en los resultados. Permiten a los ingenieros verificar si las propiedades del sistema alcanzan un umbral deseado objetivo. Por ejemplo, ¿Cuántas predicciones producidas son defectuosas o presentan incoherencias?

Para ello hay los siguientes métodos de prueba:

- Los métodos formales se basan en pruebas formales firmes, para demostrar una propiedad matemática sobre un dominio. Los métodos formales incluyen análisis de incertidumbre para probar la estabilidad de la interpolación, usar optimizadores (**solvers**, por ej. un **optimizador SMT** para probar la ausencia o existencia de ejemplos adversarios y la capacidad de exhibirlo), técnicas de optimización (como **Branch and Bound** o programación por restricciones -**constraint programing**) o interpretaciones

abstractas para probar una propiedad de espacio estable máximo (**Maximum stable space property**).

- Los métodos empíricos confían en la experimentación, observación y juicio experto. para evaluar el punto hasta el que las propiedades del sistema son ciertas en los escenarios testeados. Los ensayos de campo (field trials) que pueden usar estándares de testeo de software (por ej. ISO/IEC/IEEE 29119-3:2021, donde se indica el riesgo residual con relación a lo evaluado por las pruebas para poder encontrar defectos en el elemento testeado), testeado a posteriori (con evaluación entrada-salida, o validación por testeo).

Realizar pruebas de perturbación de las entradas, para consolidar la solidez del sistema, es un mecanismo de prueba que permite evaluar el comportamiento del sistema de IA en condiciones de producción, donde las entradas son menos controladas. A continuación, mostramos tres ejemplos de perturbaciones según los tipos de sistemas de IA:

- Para evaluar la solidez de un modelo contra una perturbación específica (por ejemplo, presencia de gotas de líquido causantes de defectos de imagen en lentes) usando un método formal debemos tener una función describiendo el proceso de aplicación de la perturbación que se base en un parámetro limitado que puede ser variado para presentar variaciones aceptables de perturbación, y es preferible tener una composición conmutativa de funciones. Otras perturbaciones podrían ser vibración, rotación, brillo, turbulencias atmosféricas, ruido homogéneo, borrosidad, sobreexposición *(***blooming bright spots***)* o manchas (smear).
- En datos de sonido, se pueden estudiar los rangos de frecuencia audibles por humanos o no (como las perturbaciones basadas en ultrasonidos).
- En aprendizaje por refuerzo (RL) se puede: aplicar fuerzas disturbantes al agente hace el agente más sólido, influenciar a un agente para que tome siempre la peor acción posible puede verse como un ataque que conlleva a una mejor solidez del agente entrenado en entornos con propiedades físicas diferentes. Reducir la precisión o eficiencia de un agente es también posible incluso si el adversario es solo parte del entorno de la víctima y, por tanto, parte de sus observaciones. Un ejemplo de técnica en este caso es entrenamiento de liga (**league training**) [81].

7.1.2.1 Métodos formales

Los métodos formales son técnicas matemáticas para la especificación rigurosa y verificación de sistemas de software y hardware que tienen como objetivo probar la corrección del sistema. Los métodos formales permiten hacer corresponder el sistema de IA en relación con la finalidad prevista con un modelo matemático. Pueden complementar a otros métodos e incrementar la confianza en la solidez de los modelos, razonar sobre estos y demostrar si satisfacen las propiedades de solidez relevantes (ver apartado 2.1 Propiedades asociadas a la solidez).

Generalmente, establecen un enlace en el modelo (por ej. red neuronal) a través de las entradas a las salidas sobre un dominio, que permite saber cuánto impacta cada entrada a cada salida. Hay diversos tipos de métodos formales. Generalmente se distingue entre:

- **Completos** (proveen respuestas exactas) e **incompletos**. Los últimos pueden resultar más realistas para su aplicación a la verificación de redes neuronales que consiguen alta precisión en tareas de datos complejas, pues usan técnicas de abstracción que escalan a éstas. Sin embargo, pueden fracasar en demostrar una propiedad de solidez que en realidad se satisface.
- **Deterministas o no**. Los últimos pueden verificar modelos como **mixture density networks** o **autoencoders variacionales** que no producen una salida determinista sino una distribución. Para demostrar garantías formales de su solidez con alta probabilidad, los métodos no deterministas pueden establecer los parámetros de esta distribución de salida determinísticamente.
- **De caja blanca** (conscientes) o de **caja negra** (agnósticos al modelo). Los verificadores de caja blanca requieren acceso a la representación interna de la red, su arquitectura y parámetros aprendidos, pero no a los datos de entrenamiento ni al algoritmo usado para entrenar la red neuronal. Los verificadores de caja negra pueden ser usados cuando el modelo está encriptado y no es accesible. Sin embargo, requieren la capacidad de ejecutar el modelo para entradas concretas, lo cual puede hacerlos menos precisos que los de caja blanca.
- **Aritmética de computador o real**. Aunque en muchos casos no es así, los verificadores pueden tener en cuenta los errores de redondeo hechos con la aritmética de computador y razonar, bajo la semántica establecida para esa aritmética concreta, sobre las salidas de una red neuronal[6].

7.1.2.2 Métodos formales y su aplicabilidad para verificar la solidez de sistemas de IA basados en redes neuronales artificiales

La aplicación de métodos formales, como mecanismo de verificación de la solidez, permiten, en las redes neuronales establecer dos propiedades formales de la solidez de un sistema:

- La **estabilidad** de la interpolación calcula la incertidumbre de una red neuronal como una manera de detectar cuando la red tendrá capacidad de interpolación insuficiente y, por tanto, insuficiente solidez.
- Un **espacio estable máximo** (**Maximum stable space**): calcula el tamaño del dominio donde la red neuronal tendrá una predicción de clasificación estable. Los métodos formales pueden dividirse según al tipo de modelo al que se aplican:
    - Métodos de verificación de **redes neuronales lineales** por partes (**Piecewise linear neural networks -PLNN)**: Se puede verificar la solidez de estas redes neuronales con optimizadores **satisfiability modulo theories** (SMT) solver, con programación lineal en enteros mixta (**PLEM, o mixed integer linear program, MILP**) u otros reflejados en ISO /IEC DTR 24029:2021 como **Fast-Lin – Fast-Lip, CROWN y formal safety analysis.**

[6] Para evitar que las garantías de solidez no se cumplan para los cómputos realmente hechos con aritmética en coma flotante u otras, los verificadores formales consideran la semántica del tipo de aritmética de cómputo, estándar o no, y garantizan que la salida captura las salidas posibles de la red bajo esa semántica. Solamente cuando los operadores redondean correctamente según lo establecido en el estándar IEEE 754, los verificadores pueden tener en cuenta cambios en el orden de los cómputos y pueden aproximar el redondeo hecho en cada operador.

- Métodos para verificar **redes neuronales binarizadas** (binarized neural nets, **BNNs**): Crean una representación exacta de la BNN con fórmulas booleanas tal que todos los pares válidos de entrada y salida sean soluciones de esta fórmula, y la verificación se consigue usando satisfacibilidad booleana, **Integer Linear Programming, ILP**).
- Métodos de verificación a través de optimizadores (**solvers**): **MILP**, **SMT** (Satisfiability Modulo Theories): Todos son **determinísticos** y de **caja blanca**. Codifican todo cómputo de una red neuronal como una restricción y, dependiendo de la arquitectura, pueden ser verificadores completos o incompletos[7]. Si los elementos en los límites de la restricción de solidez satisfacen las restricciones, la propiedad queda demostrada.
- Verificadores de Interpretación Abstracta. Marcos para el análisis escalable de sistemas probabilísticos y determinísticos grandes y complejos que, en el contexto de redes neuronales, proveen un método de caja blanca determinista, incompleto que puede verificar la solidez de arquitecturas grandes a través de la abstracción de las entradas del modelo por medio de formas geométricas, cajas, zonotopos, poliedros u otros, por lo que hay un compromiso inherente entre precisión y escalabilidad. Por ejemplo, usando cajas para acotar las entradas escala a millones de neuronas por segundo, pero es impreciso para verificar propiedades de solidez; mientras que las relajaciones semidefinidas son más precisas, pero no escalan a arquitecturas grandes.

7.1.2.3 Métodos para verificar formalmente la solidez de otros modelos particulares de redes neuronales

Los **modelos de tipo transformers**, se pueden verificar descomponiendo capas en subcapas donde se calculan límites que actúan de garantía desde la primera subcapa (de transformaciones lineales, funciones no lineales y operaciones de *self-attention*) a la última [99]. Verificando y validando la solidez por tanto en cada una de las etapas establecidas.

Para el caso de redes recurrentes (**RNNs**), éstas se pueden considerar como máquinas de estados infinitos o autómatas de estados finitos, por lo que se puede medir la solidez local de las RNNs para clasificación con **model checking** y **abstract interpretation**. Abstract interpretation es un marco para el análisis escalable de sistemas deterministas de gran escala, complejos y probabilísticos que proporciona un método incompleto, determinista y de caja blanca que puede verificar la solidez de grandes redes neuronales.

En sistemas de IA basados en **aprendizaje por refuerzo** (RL): La estabilidad y solidez debe ser considerada a la hora de documentar la solidez de los sistemas, haciendo pruebas tanto en simulación como en el mundo real [68], y usando métricas adaptadas a la tarea y contexto [69, 105].

[7] Por ej. funciones de activación hiperbólicas como la sigmoide y tanh son demasiado complejas para codificarlas precisamente, por lo que los optimizadores las aproximan con abstracciones formales.

Para otros tipos de datos como las redes neuronales basadas en grafos requerirán métodos específicos especializados[8].

7.2 Anexo II: Solidez e incertidumbre

En este Anexo presentamos la relación existente entre la incertidumbre y la solidez para el sistema de IA. El proveedor debe considerar este anexo en relación con la evaluación de la solidez, como detalle y extensión de las medidas indicadas en el apartado 4.

7.2.1 La solidez como capacidad de tratamiento y minimización de la incertidumbre

La fiabilidad de la precisión y solidez reportada para un modelo de IA depende de la fiabilidad de los datos y la estimación de los factores que generan los datos, su hipótesis en relación con la finalidad prevista, etc. [1]. Sí los datos se degradan (cuando los datos que procesa ya no siguen la distribución de los datos usados cuando se entrenó), el modelo también lo hará. (ver Guía de Datos).

Una representación fiable de la incertidumbre de un modelo es clave para cualquier modelo de IA. Al menos, hay **dos** tipos principales de **incertidumbre** asociados a un modelo predictivo: 1) **Aleatoria**, y 2) **Epistémica**. La aleatoria no es reducible, pues se debe a la inherente naturaleza estocástica de la dependencia entre instancias x y salidas y. La epistémica es reducible por el modelo: podría eliminarla recogiendo más datos para los conjuntos de trabajo [1].

El proveedor deberá proveer infraestructura para monitorizar y notificar, durante todo el ciclo de vida del sistema de IA, de potenciales:

1. Cambios en la distribución de datos de entrenamiento y test (***shift* del conjunto de datos**);
2. Disparidades en cuanto al cambio del dominio o tarea de aprendizaje automático requiriendo el ajuste al nuevo dominio de datos por medio de fine-tuning con congelación de pesos en redes neuronales, o con procesos de transferencia de aprendizaje (transfer learning).
3. **Cambios** en el mecanismo de adquisición de **nuevos datos**, involucrando, por ejemplo, la introducción de artefactos no deseados, datos que incluyen indicadores que pueden introducir un sesgo indirectamente, o un simple desequilibrio de estos.
4. Olvido catastrófico y la incapacidad de aprender de manera continua [25].

En cuanto a medidas **técnicas**, detallamos en las secciones siguientes, diferentes medidas para tratar diferentes dimensiones que conducen a la solidez de un sistema de IA.

[8] Por ejemplo, cuando se tengan datos tabulares que combinan datos de varios tipos (numéricos, simbólicos, textuales) y expresen relaciones entre ellos, esto puede impedir el uso de métodos formales debido a la gran cantidad de filas y a la dificultad de estimar la varianza dentro de cada fila.

7.2.2 Métodos para medir y cuantificar la incertidumbre de la salida de un modelo de IA

Técnicas para cuantificar la incertidumbre asociada a un modelo incluyen:

- Calibrar los modelos de clasificación (con regresión isotónica o **Sigmoid / Platt scaling**), es decir usar una técnica de post-procesamiento con el fin de mejorar la estimación de la probabilidad o mejorar el error de la distribución.
- Evaluar la incertidumbre indirectamente (evaluando su utilidad para una predicción mejorada y toma de decisiones), por ejemplo, con curvas (*accuracy-rejection curves*) u otras estrategias [1].
- Usando otras medidas de calibración: **Macro-average accuracy, proportion of classes, MSE, LogLoss, CalBin, CalLoss, Avg error, relative error, Anderson-Darlin (A2) test** [2].

Herramientas de estimación y monitoreo (ver guía de Supervisión Humana) de la degradación de la precisión y solidez del modelo a través de la incertidumbre asociada a la misma:

- IBM Uncertainty Quantification 360 Toolkit (para estimar, comunicar y usar la incertidumbre en modelos predictivos y así mejorar tanto la precisión como la solidez del modelo).
- NannyML (para estimar la incertidumbre de modelos tras la puesta en producción del modelo, antes y después de que los datos de salida objetivo dejen de estar disponibles, detecta data y model drift que posiblemente lo causan).

7.3 Glosario

Término	Definición
Abstract interpretation	Es una técnica que simplifica el análisis de programas al crear representaciones abstractas que capturan propiedades esenciales. Ayuda a comprender y razonar sobre el comportamiento de los programas sin necesidad de ejecutarlos realmente, lo que facilita la detección de errores y la optimización en la inteligencia artificial.

Término	Definición
Anderson-Darling (A2) test	La Prueba de Anderson-Darling es una prueba estadística utilizada para determinar si un conjunto de datos sigue una distribución específica. Se utiliza para evaluar si los datos se ajustan a una distribución teórica, como la distribución normal, exponencial o uniforme. La prueba se basa en la comparación de los valores observados con los valores esperados bajo la distribución teórica. Calcula una medida de ajuste llamada estadística de Anderson-Darling, que tiene en cuenta las diferencias entre los valores observados y esperados, dando más peso a las desviaciones en las colas de la distribución. La hipótesis nula de la prueba es que los datos siguen la distribución teórica, mientras que la hipótesis alternativa es que los datos no se ajustan a la distribución teórica. Si el valor de la estadística de Anderson-Darling es mayor que un valor crítico dado, se rechaza la hipótesis nula y se concluye que los datos no siguen la distribución teórica.
Aprendizaje privilegiado destilado	Es una técnica de aprendizaje automático que utiliza información adicional o conocimiento privilegiado disponible durante la etapa de entrenamiento para mejorar el rendimiento del modelo final. Se entrena un modelo secundario para aprender a predecir esta información privilegiada y se utiliza para enseñar al modelo base a aprovecharla y mejorar sus predicciones.
AUROC	El AUROC es una métrica que resume la calidad de la curva ROC y proporciona una medida cuantitativa del rendimiento del modelo. Cuanto mayor sea el valor del AUROC, mejor será el rendimiento del modelo.
Autoencoders variacionales	Son modelos de aprendizaje automático que generan representaciones latentes siguiendo una distribución específica. Estos modelos permiten reconstruir los datos originales y también generar nuevos datos similares a los originales. Son útiles en tareas de generación de datos y otras aplicaciones donde se requiere la captura de la incertidumbre asociada con la generación de datos.

Término	Definición
Mean error / relative error	El error medio (ME, por sus siglas en inglés) es una medida utilizada para evaluar el sesgo o la tendencia sistemática de un modelo o estimación al predecir los valores. Se calcula tomando la diferencia promedio entre las predicciones y los valores reales. Él ME es una medida que tiene en cuenta la dirección del error. Si las predicciones son sistemáticamente mayores que los valores reales, el ME será un número positivo, lo que indica un sesgo positivo. Por otro lado, si las predicciones son sistemáticamente menores, el ME será un número negativo, indicando un sesgo negativo. El error relativo (RE), por otro lado, proporciona una medida del error en relación con el valor real. Al expresarlo como un porcentaje, es útil para evaluar la precisión relativa del modelo o estimación en diferentes escalas de datos. El RE permite una comparación más directa de la precisión entre diferentes modelos o estimaciones, independientemente de la magnitud de los valores.
Backward transfer, forward transfer	En el contexto del aprendizaje automático y el entrenamiento de modelos, el backward transfer (transferencia hacia atrás) y el forward transfer (transferencia hacia adelante) se refieren a la influencia que puede tener el aprendizaje de una tarea en el rendimiento de otras tareas relacionadas. El backward transfer se produce cuando el aprendizaje de una tarea anterior mejora el rendimiento en tareas posteriores. Es decir, el conocimiento adquirido al aprender una tarea previa ayuda a mejorar el rendimiento en tareas subsiguientes. Por otro lado, el forward transfer ocurre cuando el aprendizaje de una tarea posterior mejora el rendimiento en tareas previas. En este caso, el conocimiento adquirido al aprender una tarea posterior influye en la mejora del rendimiento en tareas anteriores. Estas transferencias pueden ocurrir cuando las tareas tienen cierta relación o similitud en términos de características o estructura subyacente. Si el conocimiento aprendido en una tarea puede ser aplicado de manera efectiva en tareas relacionadas, se produce una transferencia de aprendizaje, ya sea hacia atrás o hacia adelante.
Blooming bright spots	Se refiere a un fenómeno visual que puede ocurrir en imágenes digitales, especialmente en imágenes capturadas por cámaras digitales o sensores CCD (dispositivos de carga acoplada). Está relacionado con la presencia de ruido en las imágenes.
Branch and Bound	Es una técnica utilizada para resolver problemas de optimización explorando sistemáticamente todas las posibles soluciones. Utiliza estrategias de ramificación para dividir el espacio de búsqueda y técnicas de acotamiento para eliminar soluciones no prometedoras. El objetivo es encontrar una solución óptima o aproximarse a ella de manera eficiente.

Término	Definición
CalBin	Es una medida de calibración basada en el agrupamiento solapado. Para cada clase se debe ordenar todos los casos según la predicción *p(i, j)* reasignando los índices. Se tomarán los 100 primeros elementos como primer bin, se calculará el porcentaje de casos de la clase *j* en esta casilla como probabilidad real, f_j. $$\Sigma_{i\epsilon\|\ldots_{100}}\|p(i,j)-fj\|$$ El error para este bin es, Posteriormente se calculare el error para el resto de bins. Finalmente se calculará la media de los errores.
CalLoss	Un clasificador perfectamente calibrado siempre da una curva ROC convexa. Sin embargo, un clasificador puede producir puntuaciones muya latas (AUC alto), pero las probabilidades quizas difieren de las probabilidades reales. Un metodo para calibar un clasificador es calcular el casco convexo o, lo quees lo mismo utilizar la regresión isotonica. Flach y Matsubara (2007) descomponen la puntuación de Brier en pérdida de calibración y pérdida de refinamiento. CalLoss es definida como la desviación media al cuadrado de las probabilidades empiricas derivadas de la pendiente de los segmentos ROC.
Cohen Kappa	Es una medida de acuerdo, utilizada para evaluar la concordancia entre dos evaluadores o clasificadores. Proporciona una medida más robusta que la simple coincidencia en la clasificación y tiene en cuenta el acuerdo esperado por casualidad. Un valor de kappa cercano a 1 indica un alto grado de acuerdo, mientras que un valor cercano a 0 indica un acuerdo similar al esperado por casualidad.
Computational efficiency	La eficiencia computacional en el aprendizaje automático se refiere a la capacidad de los algoritmos y modelos de machine learning para realizar tareas de manera rápida y eficiente en términos de recursos computacionales, como el tiempo de ejecución y la capacidad de almacenamiento. En el contexto del aprendizaje automático, la eficiencia computacional es un aspecto importante debido a la creciente complejidad de los conjuntos de datos y los modelos. Se busca optimizar el uso de recursos computacionales para lograr un equilibrio entre el rendimiento del modelo y los recursos requeridos.

Término	Definición
Confidence-based Performance estimation (CBPE)	Es una técnica que utiliza las medidas de confianza o probabilidad asignadas por un modelo de clasificación para evaluar su rendimiento. Al establecer un umbral de confianza, se calcula la precisión del modelo solo para las predicciones con una medida de confianza superior a ese umbral, lo que permite obtener una estimación más precisa del rendimiento del modelo en situaciones críticas.
Correlating controls and measures	La correlación entre controles y mediciones se refiere a la relación o dependencia que existe entre los valores de los controles utilizados en un sistema y las mediciones resultantes. En el contexto de sistemas de control y monitoreo, es importante comprender cómo los cambios en los controles afectan las mediciones y viceversa. Cuando existe una correlación fuerte entre los controles y las mediciones, significa que los cambios en los valores de los controles tienen un impacto significativo en las mediciones obtenidas. Esto puede ser deseable en algunos casos, ya que indica que los controles tienen un efecto predecible y pueden utilizarse para influir en las mediciones de manera controlada. Sin embargo, también puede haber situaciones donde una correlación fuerte entre los controles y las mediciones no sea deseable. Por ejemplo, si existen factores externos que pueden afectar las mediciones independientemente de los controles, una correlación fuerte puede dificultar la identificación de la verdadera causa de las variaciones en las mediciones. La comprensión de la correlación entre los controles y las mediciones es esencial en muchos campos, como la ingeniería, la física y la ciencia de datos. Permite analizar y modelar adecuadamente los sistemas, identificar relaciones causales, optimizar los controles para lograr los resultados deseados y entender el comportamiento de los sistemas en diferentes condiciones.
Coverage tests	Son técnicas utilizadas para evaluar qué parte del código fuente ha sido ejecutada durante las pruebas. Hay diferentes tipos de cobertura, como cobertura de línea, cobertura de rama, cobertura de condición y cobertura de camino, que se enfocan en diferentes aspectos del código. Las pruebas de cobertura son útiles para medir y mejorar la calidad del software, aunque no garantizan la ausencia de errores.

Término	Definición
Crammer-Singer	Es un método iterativo utilizado para resolver el problema de clasificación multiclase. A través de una serie de transformaciones y optimizaciones, busca maximizar la separabilidad entre las clases en un espacio de características. Este enfoque evita la necesidad de descomponer el problema en clasificadores binarios independientes y puede mejorar el rendimiento de la clasificación al considerar información adicional.
CROWN	Siglas de "Convex Relaxation Of Nonlinear layer" (Relajación Convexa de Capas No Lineales), es un enfoque utilizado en el campo del aprendizaje automático y las redes neuronales. Es un enfoque que utiliza la programación matemática convexa para aproximar la función de propagación hacia adelante de una red neuronal no lineal mediante una formulación convexa. Proporciona límites superiores e inferiores para la salida de la red neuronal en función de los límites de las entradas, lo que permite realizar análisis de robustez y certificar propiedades de seguridad para el modelo.
Data differentials	Diferencias o cambios en los datos entre diferentes puntos o momentos en el tiempo. Son una medida utilizada para analizar cómo los datos varían o difieren en diferentes situaciones y pueden proporcionar información valiosa sobre las tendencias, patrones y cambios en los datos.
Data shift	Cambio en la distribución de los datos entre el conjunto de entrenamiento y el conjunto de prueba o entre diferentes momentos en el tiempo. Puede afectar negativamente el rendimiento del modelo y requiere enfoques específicos para abordar el problema y mitigar su impacto.
Dataset shift	Se produce cuando hay una discrepancia o cambio en la distribución de los datos entre conjuntos de datos relacionados. Puede tener un impacto negativo en el rendimiento del modelo y requiere técnicas específicas para mitigar su efecto y mejorar la generalización del modelo.
Destilación	Se refiere a transferir el conocimiento de un modelo maestro más grande y complejo a un modelo alumno más pequeño y simplificado. Se utiliza para entrenar al modelo alumno de manera más eficiente y efectiva, utilizando etiquetas suaves generadas por el modelo maestro en lugar de etiquetas categóricas. Esto permite que el modelo alumno capture el conocimiento y la estructura subyacente de los datos y generalice mejor.

Término	Definición
Direct Loss Estimation (DLE)	Es un enfoque en el aprendizaje automático que busca estimar directamente la función de pérdida asociada a un problema en lugar de utilizar la función de pérdida convencional. El objetivo es obtener estimaciones más precisas de la pérdida y mejorar el rendimiento del modelo en tareas específicas, teniendo en cuenta aspectos particulares del problema.
Elastic Weight Consolidation (EWC)	Es un método utilizado en el aprendizaje automático para mitigar el olvido catastrófico de conocimientos previamente aprendidos al entrenar modelos en nuevas tareas. Protege selectivamente los pesos importantes para tareas anteriores mediante una función de penalización que evita cambios drásticos en esos pesos durante el entrenamiento en nuevas tareas. Esto permite que el modelo conserve y utilice eficientemente los conocimientos previos adquiridos.
Ensemble Learning	Es un enfoque en el aprendizaje automático que combina múltiples modelos de aprendizaje para mejorar la precisión y el rendimiento general. Al aprovechar la diversidad de los modelos y combinar sus predicciones individuales, se obtiene una predicción final más sólida y confiable. Este enfoque se basa en la premisa de que la combinación de modelos independientes puede superar las limitaciones individuales y mejorar la capacidad de generalización.

Término	Definición
Experience Replay (replay buffer, prototype experience replay, rehearsal)	Experience Replay, también conocido como replay buffer, prototype experience replay o rehearsal, es una técnica utilizada en el entrenamiento de modelos de aprendizaje profundo, especialmente en algoritmos de aprendizaje por refuerzo. La idea central de Experience Replay es almacenar y reutilizar experiencias pasadas durante el entrenamiento en lugar de utilizar solo las experiencias más recientes. Esto se logra mediante el almacenamiento de transiciones de experiencia, que consisten en pares de estados y acciones tomadas en un entorno, en una memoria de almacenamiento llamada replay buffer. Durante el entrenamiento, en lugar de utilizar solo las transiciones más recientes, se muestrean aleatoriamente las transiciones almacenadas en el replay buffer. Esto permite al modelo de aprendizaje profundo volver a experimentar una variedad de situaciones pasadas, lo que ayuda a evitar la dependencia excesiva de las experiencias más recientes y a mejorar la eficiencia y la estabilidad del entrenamiento. El uso de Experience Replay ofrece varias ventajas. En primer lugar, ayuda a decorrelacionar las transiciones de experiencia, ya que las experiencias adyacentes en el tiempo pueden estar altamente correlacionadas. Esto evita que el modelo se entrene solo en situaciones similares y mejora su capacidad para generalizar a diferentes situaciones. Además, el replay buffer permite una exploración más efectiva, ya que las transiciones se muestrean aleatoriamente y se pueden evitar comportamientos subóptimos. Existen variaciones de Experience Replay, como Prototype Experience Replay, que se enfoca en almacenar y utilizar ejemplos representativos o prototipos para mejorar el rendimiento del modelo.
F1-Score	El F1 score es una medida estadística que combina la precisión y el recuerdo (recall) de un modelo de clasificación. Es útil cuando se tienen clases desequilibradas en los datos. Proporciona una única métrica que representa la precisión y el recuerdo de manera equilibrada, y su valor oscila entre 0 y 1, siendo 1 el mejor resultado posible. Un F1 score alto indica un equilibrio entre la precisión y el recuerdo del modelo en la clasificación de las clases.
Fast-Lin	Algoritmo computacionalmente eficiente. Calcula un límite inferior certificado sobre las perturbaciones de entrada permitidas para redes ReLU utilizando un enfoque capa por capa y búsqueda binaria en el dominio de entrada.

Término	Definición
Fast-Lip	Algoritmo computacionalmente eficiente. Depende de Fast-Lin para calcular los límites de las funciones de activación y, además, estima la constante local de Lipschitz de la red. En general, Fast-Lin es más escalable que Fast-Lip, mientras que Fast-Lip proporciona mejores soluciones para límites ℓ_1.
Formal safety analysis	Se requiere un análisis formal de las redes neuronales debido a los riesgos asociados con las predicciones erróneas en situaciones adversas. Este enfoque permite verificar propiedades de seguridad y encontrar contraejemplos concretos en redes más grandes, lo que mejora significativamente la capacidad de análisis existente. Además, este enfoque puede contribuir a mejorar la explicabilidad y la robustez de las redes neuronales.
Group K fold cross validation	Es una técnica de evaluación de modelos que tiene en cuenta la estructura grupal de los datos. Divide los datos en grupos predefinidos y evalúa el rendimiento del modelo en conjuntos de prueba y entrenamiento que son representativos de diferentes grupos. Esto permite una evaluación más precisa del rendimiento del modelo en situaciones del mundo real donde los datos tienen una estructura grupal.
K-fold cross validation	Es una técnica que divide los datos en k pliegues y realiza k iteraciones para evaluar y seleccionar modelos. Proporciona una estimación más precisa del rendimiento del modelo y ayuda a evitar problemas de sobreajuste o subajuste al utilizar todos los datos de manera más efectiva.
League training	Es una técnica de entrenamiento en la que múltiples agentes son entrenados en paralelo y compiten entre sí en un entorno de juego. A través de la competencia interna, los agentes aprenden y mejoran sus estrategias con el objetivo de alcanzar un rendimiento óptimo en la liga.
Leave-one-group-out cross validation	Es una técnica de validación cruzada que deja fuera un grupo completo en cada iteración como conjunto de prueba. Es útil cuando los datos tienen una estructura de grupos relacionados y permite evaluar el rendimiento del modelo teniendo en cuenta las características específicas de cada grupo.
Leave-one-out cross validation	Es una técnica de validación cruzada en la que se deja fuera una única muestra como conjunto de prueba en cada iteración. Permite una evaluación exhaustiva del modelo utilizando todos los datos disponibles, aunque puede ser computacionalmente costoso.

Término	Definición
LogLoss	También conocida como pérdida logarítmica o pérdida de entropía cruzada, es una métrica de evaluación común para los modelos de clasificación binaria. Mide el rendimiento de un modelo cuantificando la diferencia entre las probabilidades predichas y los valores reales. La pérdida logarítmica es indicativa de cuán cerca está la probabilidad de predicción del valor real / verdadero correspondiente (0 o 1 en el caso de la clasificación binaria), penalizando las predicciones inexactas con valores más altos. Una menor pérdida de registro indica un mejor rendimiento del modelo. Log Loss es la métrica de clasificación más importante basada en probabilidades. Es difícil interpretar los valores de pérdida de registro sin procesar, pero la pérdida de registro sigue siendo una buena métrica para comparar modelos. Para cualquier problema dado, un valor de pérdida logarítmica más bajo significa mejores predicciones.
Macro-average accuracy	Es una métrica utilizada para evaluar el rendimiento de un modelo de clasificación en problemas multiclase. En lugar de calcular la precisión para cada clase individualmente y promediar los resultados, la precisión promedio macro calcula la precisión por clase y luego realiza un promedio no ponderado de estas precisiones. Toma en cuenta el rendimiento de todas las clases por igual y no se ve afectada por el desequilibrio de clases en los datos. Cada clase contribuye de manera equitativa al cálculo de la precisión promedio macro, lo que la hace adecuada cuando se desea evaluar el rendimiento general del modelo en todas las clases por igual.
Matthews correlation coefficient	Es una métrica utilizada para evaluar el rendimiento de un modelo de clasificación binaria. Proporciona una medida del equilibrio entre los verdaderos positivos, los verdaderos negativos, los falsos positivos y los falsos negativos. Es una métrica útil cuando se trabaja con conjuntos de datos desequilibrados, donde las clases positivas y negativas tienen diferentes proporciones. A diferencia de la precisión o el recall, el MCC no se ve afectado por el desequilibrio de clases y proporciona una evaluación más equilibrada del rendimiento del modelo.
MaxError	ME o Error Máximo es el valor absoluto de la diferencia más significativa entre una variable predicha y su valor real.

Término	Definición
Maximum stable space	Es el conjunto de puntos de datos que son robustos frente a perturbaciones y no experimentan cambios significativos en sus etiquetas o clasificaciones. Este concepto es útil para comprender la estabilidad y la robustez de los modelos de aprendizaje automático y puede ser utilizado para mejorar la generalización y la capacidad de adaptación de los algoritmos de aprendizaje automático.
Maximum stable space property	La propiedad del Maximum Stable Space establece que un algoritmo es capaz de mantener la estabilidad de los datos al preservar las clasificaciones o etiquetas de los puntos de datos incluso en presencia de perturbaciones. Esta propiedad es importante para garantizar la robustez y la generalización de los modelos de aprendizaje automático en diferentes escenarios y condiciones.
Método SQuaRE	Se conoce como Evaluación de Riesgos Semi-Cuantitativa (Semi-Quantitative Risk Assessment). Utiliza una combinación de análisis cualitativo y cuantitativo para evaluar y clasificar los riesgos. Proporciona una forma estructurada de identificar, analizar y priorizar los riesgos, teniendo en cuenta tanto su impacto como su probabilidad. Este enfoque ayuda a las organizaciones a tomar decisiones informadas sobre cómo manejar y mitigar los riesgos identificados.
Mixture density networks	Son una arquitectura de redes neuronales utilizada en problemas de modelado de distribuciones de probabilidad. En lugar de predecir un único valor o una clasificación, las MDN son capaces de modelar y predecir distribuciones de probabilidad complejas. Son útiles en aplicaciones como generación de texto, síntesis de voz, predicción de series de tiempo y problemas de regresión donde es necesario modelar y estimar distribuciones de probabilidad complejas. Proporcionan una forma flexible y poderosa de representar la incertidumbre en las predicciones de las redes neuronales.
Model checking	Es una técnica de verificación formal utilizada en ciencias de la computación para analizar y verificar la corrección de sistemas concurrentes o sistemas de software. Consiste en verificar automáticamente si un modelo formal del sistema cumple con ciertas propiedades especificadas. Esta técnica es particularmente útil para sistemas críticos donde la correctitud y la verificación son de suma importancia, como en el diseño de circuitos electrónicos, protocolos de comunicación, sistemas embebidos y software de seguridad crítica. El model checking proporciona una forma rigurosa y automática de asegurar la corrección de los sistemas y

Término	Definición
	detectar posibles problemas antes de su implementación o despliegue.
Model size efficiency	Se refiere a la capacidad de lograr un rendimiento óptimo o aceptable utilizando un modelo de tamaño reducido. En otras palabras, se trata de obtener resultados comparables o cercanos a los obtenidos con modelos más grandes, pero con menos parámetros o recursos computacionales.
Monte Carlo cross-validation	Es un enfoque que combina la validación cruzada tradicional con el muestreo aleatorio repetido para evaluar el rendimiento de un modelo. Ayuda a obtener una estimación más confiable del rendimiento del modelo al considerar múltiples particiones aleatorias de los datos.
MSE	Es una métrica de evaluación que mide el promedio de las diferencias al cuadrado entre las predicciones y los valores reales en un problema de regresión. Es ampliamente utilizado y penaliza más los errores grandes, pero puede ser sensible a la escala de los datos.
ONNX	ONNX (Open Neural Network Exchange) es un formato de intercambio de modelos de aprendizaje automático de código abierto, que proporciona un formato estándar para describir la estructura del modelo y los parámetros asociados, lo que permite a los desarrolladores entrenar modelos en un marco y utilizarlos en otro sin tener que recrear el modelo desde cero. Fue desarrollado en colaboración por Microsoft y Facebook para facilitar la interoperabilidad entre diferentes marcos y herramientas de aprendizaje automático.

Término	Definición
Optimizador SMT	Es una herramienta de software utilizada en el ámbito de la verificación formal y la solución de problemas de satisfacción de restricciones. Combina las capacidades de los solucionadores de satisfacción de restricciones (SAT) con la capacidad de razonamiento sobre teorías específicas, como teorías aritméticas, teorías de conjuntos, teorías de listas, teorías de arrays, entre otras. Los optimizadores SMT son ampliamente utilizados en la verificación de hardware y software, la síntesis de controladores, la planificación automatizada, la optimización de programas y otros dominios en los que es necesario razonar sobre restricciones y teorías combinadas.
Out-of-distribution	La detección de datos fuera de distribución en redes neuronales profundas es un área de investigación que se enfoca en identificar y detectar datos que son significativamente diferentes de los datos de entrenamiento de un modelo. Esto es importante porque los datos fuera de distribución pueden llevar a predicciones incorrectas o poco confiables.
Piecewise linear neural networks -PLNN	Son una arquitectura de redes neuronales que utilizan neuronas lineales en cada segmento lineal del espacio de entrada. Aunque son más simples y fáciles de entrenar, pueden tener dificultades para modelar relaciones no lineales complejas. Son adecuadas para aplicaciones donde una aproximación lineal es suficiente o cuando se desea un modelo más simple y eficiente.
PLEM o MILP (mixed integer linear program)	Es un enfoque de optimización matemática que combina variables continuas y enteras en un modelo de programación lineal. Permite abordar problemas de toma de decisiones discretas y optimización, donde algunas variables deben ser enteras.
Progreso diferencial	Es un enfoque de entrenamiento de redes neuronales que utiliza pequeños ajustes graduales en los parámetros del modelo para lograr una convergencia más rápida y estable. Es una técnica eficiente y efectiva para optimizar modelos de aprendizaje automático.
Proportion of classes	Indica la distribución relativa de las diferentes clases en relación con el total de muestras o instancias en un conjunto de datos.
RMSE	Es una métrica que cuantifica el error promedio entre las predicciones de un modelo de regresión y los valores reales del conjunto de datos. Es utilizado para evaluar y comparar la precisión de diferentes modelos, siendo deseable un valor de RMSE lo más bajo posible.
Safeguard based design	Enfoque de diseño que se centra en incorporar salvaguardias y mecanismos de seguridad desde el principio en los sistemas de IA,

Término	Definición
	con el principal objetivo de garantizar que los sistemas de IA sean seguros, éticos y confiables, y minimizar los riesgos asociados con su implementación. Este diseño implica identificar posibles problemas, riesgos y vulnerabilidades inherentes a los sistemas y tomar medidas para mitigarlos. Esto puede incluir la implementación de controles de seguridad, medidas de privacidad, evaluaciones de riesgos y pruebas rigurosas. Además, se deben considerar aspectos éticos y legales al diseñar estos sistemas, como la equidad, la transparencia y la responsabilidad. El enfoque de safeguard-based design se basa en el principio de que es más efectivo y menos costoso abordar los problemas de seguridad y ética en las primeras etapas de desarrollo, en lugar de intentar solucionarlos una vez que el sistema esté implementado y en funcionamiento. Al integrar salvaguardias desde el diseño inicial, se busca prevenir posibles consecuencias no deseadas o dañinas y garantizar que los sistemas sean seguros y confiables para su uso en diferentes aplicaciones.
Samples storage size efficiency	Se refiere a la capacidad de reducir el espacio necesario para almacenar las muestras sin comprometer significativamente el rendimiento o la calidad del modelo. Esto es importante porque los conjuntos de datos pueden ser masivos, ocupar mucho espacio de almacenamiento y requerir recursos computacionales significativos para su procesamiento.
Satisfiability modulo theories (solver)	Las teorías del módulo de satisfacción (SMT) se refieren al problema de determinar si una fórmula de primer orden es satisfactoria con respecto a alguna teoría lógica. Los solvers basados en SMT se utilizan como motores de back-end en aplicaciones de verificación de modelos, tales como verificación de modelos acotados, basados en interpolación y en abstracción de predicados.
Solvers	Los solvers son programas o algoritmos especializados que resuelven problemas de optimización matemática encontrando soluciones óptimas o subóptimas a través de técnicas y algoritmos específicos. Son herramientas fundamentales para abordar problemas de optimización en diversos campos, incluyendo la economía, la logística, la ingeniería, entre otros.

Término	Definición
Técnicas de domain adaptation	La adaptación profunda del dominio nos permite transferir el conocimiento aprendido por un DNN en particular en una tarea de origen a una nueva tarea objetivo-relacionada. Se ha aplicado con éxito en tareas como la clasificación de imágenes o la transferencia de estilo. En cierto sentido, la adaptación profunda del dominio nos permite acercarnos al rendimiento a nivel humano en términos de la cantidad de datos de entrenamiento necesarios para una nueva tarea de visión artificial en particular. Por lo tanto, creo que el progreso en esta área será crucial para todo el campo de la visión artificial y espero que eventualmente nos lleve a una reutilización efectiva y simple del conocimiento a través de tareas visuales.
Unfortunate counterfactual events	Se refiere a eventos que podrían haber ocurrido, pero no lo hicieron. Estos eventos se denominan "counterfactual" porque son situaciones hipotéticas o contrafácticas que se plantean para evaluar el impacto de decisiones o acciones alternativas. En el contexto del aprendizaje automático y la toma de decisiones, los "unfortunate counterfactual events" se refieren a situaciones en las que se produce un resultado negativo o no deseado como consecuencia de una decisión tomada o una acción realizada. Estos eventos no deseados pueden ser el resultado de diversas circunstancias, como un modelo de aprendizaje automático que no está correctamente entrenado o no ha capturado todas las posibilidades, datos de entrenamiento insuficientes o desequilibrados, o simplemente la incertidumbre inherente en el proceso de toma de decisiones. El estudio de los "unfortunate counterfactual events" es importante en el aprendizaje automático para evaluar y comprender los errores o resultados no deseados de los modelos, y para identificar posibles mejoras o soluciones. Al analizar y considerar los escenarios contrafácticos, se puede aprender sobre las acciones o decisiones alternativas que podrían haber llevado a mejores resultados y utilizar este conocimiento para mejorar los modelos, los algoritmos y las estrategias de toma de decisiones en el futuro.
Zero-shot generalization	La generalización sin entrenamiento (zero-shot generalization) se refiere a la capacidad de un modelo de aprendizaje automático para aplicar su conocimiento previo y comprender tareas o dominios completamente nuevos sin haber sido entrenado específicamente en esos dominios. Permite que los modelos utilicen el conocimiento transferible adquirido en el entrenamiento para realizar tareas más allá de los ejemplos de entrenamiento específicos.

7.4 Apéndice: Traducciones críticas (inglés - español)

- Performance: Precisión, por el contexto más cercano que se suele representar en ML/AI (podría ser accuracy también), aunque normalmente se traduzca por rendimiento o eficiencia en contextos más generales en ciencias de la computación.
- Accuracy: Exactitud o tasa de aciertos (TP+TN)/(TP+TN+FN+FP)
- Precisión: Precisión, Valor positivo predicho: fracción de instancias recuperadas que son relevantes.
- Recall (hit rate, or TPR): Sensibilidad o exhaustividad, Fracción de instancias relevantes que han sido recuperadas.
- Confidence, Trust y Reliance (a veces traducida como dependencia): Todas se traducen por Confianza.
- Trade-off: Compromiso.

8. Referencias, estánderes y normas

8.1 Referencias

[1] Hullermeier et al. 2019 Aleatoric and Epistemic Uncertainty in Machine Learning: A Tutorial Introduction.
[2] Calibration of Machine Learning Models. Antonio Bella, Cèsar Ferri, José Hernández-Orallo, and María José Ramírez-Quintana. Departamento de Sistemas Informáticos y Computación, Universidad Politécnica de Valencia 2010.
[3] Metrics for Deep Generative Models N Chen · 2018 ·
[4] Language Models are Few-Shot Learners T Brown et al. 2020
[5] Physically-Consistent Generative Adversarial Networks for Coastal Flood Visualization Bjorn Lutjens et al. 2021
[6] IBM Uncertainty Quantification 360 Toolkit https://uq360.mybluemix.net
[7] Questioning causality on sex, gender and COVID-19, and identifying bias in large-scale data-driven analyses: the Bias Priority Recommendations and Bias Catalog for Pandemics. Díaz-Rodríguez et al. 2021 https://arxiv.org/abs/2104.14492
[8] https://catalogofbias.org
[9] A survey on datasets for fairness-aware machine learning Tai Le Quy∗1, Arjun Roy†12, Vasileios Iosifidis‡1, Wenbin Zhang§3, and Eirini Ntoutsi 2022
[10] Datasheets for Datasets. Timnit Gebru et al. 2021
[11] "Searching for a Search Method: Benchmarking Search Algorithms for Generating NLP Adversarial Examples",
[12] EMNLP 2020 Blackbox NLP Workshop track proceedings.
[13] TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP, J Morris et al. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020 [In Python]
[14] PLENARY: Explaining black-box models in natural language through fuzzy linguistic summaries K Kaczmarek-Majer, G Casalino, G Castellano… - Information …, 2022 - Elsevier
[15] Companies Committed to Responsible AI: From Principles towards Implementation and Regulation? Paul B. de Laat Philosophy & Technology volume 34, pages 1135–1193 (2021)[9]
[16] COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity 2016
[17] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine

[9] Note: SHAP is not from Amazon nor proprietary in Table 3 (SHAP is Scott Lundberg's creation with a MIT open source license, in Github, and was developed with Microsoft Research teams. AzureML also implements XAI algorithms and Amazon has as ML tool, Amazon SageMaker).

Learn Research), Sorelle A. Friedler and Christo Wilson (Eds.), Vol. 81. PMLR, New ing

[18] An Ontology for Fairness Metrics. Franklin et al. https://dl.acm.org/doi/pdf/10.1145/3514094.3534137

[19] Human-centred artificial intelligence https://scilog.fwf.ac.at/en/environment-and-technology/15317/human-centred-artificial-intelligence

[20] A Holzinger et al. Digital Transformation in Smart Farm and Forest Operations Need Human-Centered AI: Challenges and Future Directions

[21] 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.

[22] 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.

[23] 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. Springer Lecture Notes in Artificial Intelligence Volume LNAI 10344. Cham: Springer International, pp. 13-50. doi: 10.1007/978-3-319-69775-8_2

[24] Human-Centered Artificial Intelligence for Designing Accessible Cultural Heritage G Pisoni, N Díaz-Rodríguez, H Gijlers, L Tonolli Applied Sciences 11 (2), 870

[25] Continual Learning for Robotics: Definition, Framework, Learning Strategies, Opportunities and Challenges T Lesort, V Lomonaco, A Stoian, D Maltoni, D Filliat, N Díaz-Rodríguez Information Fusion 220

[26] 2020 A survey on ontologies for human behavior recognition ND Rodríguez, MP Cuéllar, J Lilius, MD Calvo-Flores ACM Computing Surveys (CSUR) 46 (4), 1-33 219 2014 A fuzzy ontology for semantic modelling and recognition of human behaviour ND Rodríguez, MP Cuéllar, J Lilius, MD Calvo-Flores Knowledge-Based Systems 66, 46-60 133 2014

[27] Explainability in Deep Reinforcement Learning A Heuillet, F Couthouis, N Díaz-Rodríguez Knowledge-Based Systems 214, 106685 119 2021

[28] Don't forget, there is more than forgetting: new metrics for Continual Learning N Díaz-Rodríguez, V Lomonaco, D Filliat, D Maltoni NeurIPS workshop on Continual Learning 2018

[29] Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence A Holzinger, M Dehmer, F Emmert-Streib, R Cucchiara, I Augenstein, ... Information Fusion 79, 263-278 2022

[30] Personas for Artificial Intelligence (AI) An Open Source Toolbox A Holzinger, M Kargl, B Kipperer, P Regitnig, M Plass, H Müller IEEE Access 10, 23732-23747 2022

[31] Measuring the quality of explanations: the system causability scale (SCS) A Holzinger, A Carrington, H Müller KI-Künstliche Intelligenz 34 (2), 193-198

[32] Interdisciplinary Research in Artificial Intelligence: Challenges and Opportunities R Kusters, D Misevic, H Berry, A Cully, Y Le Cunff, L Dandoy, ... Frontiers in Big Data 3, 45 2020

[33] S. Stanton, P. Izmailov, P. Kirichenko, A. A. Alemi, A. G. Wilson, Does knowledge distillation really work? (2021). doi:10.48550/ARXIV.2106.05945. URL https://arxiv.org/abs/2106.05945

[34] A Neural-Symbolic learning framework to produce interpretable predictions for image classification, PhD Thesis 2022.
[35] Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. AB Arrieta, N Díaz-Rodríguez, J Del Ser, A Bennetot, S Tabik, A Barbado, ...Information Fusion 58, 82-115
[36] Wilcoxon, Frank. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*. 1945, 1 (6), 1095 pages 80-83.
[37] Dietterich et al. Approximate Statistical Tests for Comparing Supervised Classification 1092 Learning Algorithms. *Neural Computation, Volume 10, Issue 7*. 1998, 10 (7), pages 1895-1923. 1093 https://doi.org/10.1162/089976698300017197
[38] Akenine-Möller, Tomas, and Johnsson, Björn. Performance per what? Journal of *Computer Graphics* 1073 *Techniques*. 2012, **1**, pages 37-41. http://jcgt.org/published/0001/01/03/paper.pdf
[39] Blouw, Peter and Xuan Choo and Hunsberger, Eric and Eliasmith, Chris. Benchmarking keyword 1075 spotting efficiency on neuromorphic hardware. *Proceedings of the 7th Annual Neuro-inspired* 1076 *Computational Elements Workshop*. 2019, pages 1-8. https://arxiv.org/pdf/1812.01739.pdf
[40] Suffering-focused AI safety: In favor of "fail-safe" measures Lukas Gloor Center on Long-Term Risk Report
[41] Superintelligence as a Cause or Cure for Risks of Astronomical Suffering
[42] Kaj Sotala and Lukas Gloor Foundational Research Institute, Berlin, Germany Superintelligence as a Cause or Cure ... Informatica 41 (2017) 389–400
[43] Safe Deep RL in 3D environments using human feedback. 2022.
[44] Safeguard By Design Lessons Learned from DOE Experience Integrating Safety in Design
[45] Sustainability at scale: towards bridging the intention-behavior gap with sustainable recommendations S Tomkins, S Isley, B London, L Getoor - Proceedings of the 12th ACM conference on …, 2018
[46] Green AI. Schwartz et al.
[47] Distilling the Knowledge in a Neural Network
[48] EVALUATION METRICS FOR LANGUAGE MODELS Stanley Chen, Douglas Beeferman, Ronald Rosenfeld.
[49] Chip Huyen, "Evaluation Metrics for Language Modeling", The Gradient, 2019. https://thegradient.pub/understanding-evaluation-metrics-for-language-models/
[50] Model cards for model reporting. M Mitchell, S Wu, A Zaldivar, P Barnes, L Vasserman… - Proceedings of the …, 2019
[51] Frank McSherry Materialize: a platform for building scalable event based systems
[52] Frank McSherry, Kunal Talwar Mechanism design via Differential Privacy, 2008
[53] Nobel Prize Report, Mechanism Design. 2007 Scientific background on
[54] the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2007 Mechanism Design Theory, based in:
[55] L. Hurwicz & S. Reiter (2006) Designing Economic Mechanisms, p. 30
[56] https://divedeep.ai/2022/03/17/data-drift-vs-concept-drift/
[57] University of Oxford researchers have created a tool called capAI, a procedure for conducting conformity assessment of AI systems in line with the EU Artificial

Intelligence Act. CapAI provides organizations with practical guidance on how to translate high-level ethics principles into verifiable criteria that help shape the design, development, deployment and use of ethical AI. This tool can be used to demonstrate that the development and operation of an AI system are trustworthy. The tool is being validated with firms at the moment and the most up-to-date version can be found here

[58] A survey on concept drift adaptation ACM computing surveys (CSUR), 46(4):1–37, 2014, Gama et al.

[59] Analysis of representations for domain adaptation Neurips 2007, Ben-David et al

[60] Dataset Shift in Machine Learning Quiñonero-Candela et al 2022

[61] Generalized out-of-distribution detection: A survey. Yang et al 2022

[62] Understanding Continual Learning Settings with Data Distribution Drift Analysis" Lesort et al. 2022 video: https://www.youtube.com/watch?v=WFhozvAgnsU

[63] Sagemaker Clarify: Amazon AI Fairness and Explainability Whitepaper https://pages.awscloud.com/rs/112-TZM-766/images/Amazon.AI.Fairness.and.Explainability.Whitepaper.pdf

[64] https://aws.amazon.com/blogs/machine-learning/learn-how-amazon-sagemaker-clarify-helps-detect-bias/

[65] Data Privacy and Trustworthy Machine Learning. Strobel et al. 2022

[66] Gradual (In)Compatibility of Fairness Criteria Corinna Hertweck, Tim Räz 2022 https://arxiv.org/abs/2109.04399

[67] Decoupling feature extraction from policy learning: assessing benefits of state representation learning in goal-based robotics A Raffin, A Hill, R Traoré, T Lesort, N Díaz-Rodríguez, D Filliat ICLR 2019 Workshop on Structure & Priors in Reinforcement Learning (SPIRL) 2019

[68] Continual reinforcement learning deployed in real-life using policy distillation and sim2real transfer RT Kalifou, H Caselles-Dupré, T Lesort, T Sun, N Diaz-Rodriguez, D Filliat ICML Workshop on Multi-Task and Lifelong Learning 2019

[69] S-RL Toolbox: Environments, Datasets and Evaluation Metrics for State Representation Learning

[70] A Raffin, A Hill, R Traoré, T Lesort, N Díaz-Rodríguez, D Filliat NeurIPS workshop on Deep Reinforcement Learning 2018

[71] Deep Unsupervised state representation learning with robotic priors: a robustness analysis

[72] T Lesort, M Seurin, X Li, N Díaz-Rodríguez, D Filliat. 2019 International Joint Conference on Neural Networks (IJCNN) 2017

[73] Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence A Holzinger, M Dehmer, F Emmert-Streib, R Cucchiara, I Augenstein, et al. Information Fusion.

[74] Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study Zech et al 2018 (extendido de Confounding variables can degrade generalization performance of radiological deep learning models. Zech et al. 2018.)

[75] ISO/IEC 25000, Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Guide to SQuaRE and ISO/IEC

25059:2023, Software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality model for AI systems.
[76] STRIDE-AI: An Approach to Identifying Vulnerabilities of Machine Learning Assets Lara Mauri et al. IEEE CSR 2021
[77] Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks Papernot 2016
[78] Steps Toward Robust Artificial Intelligence Thomas G. Dietterich 2017
[79] Improving the Robustness of Deep Neural Networks via Stability Training
[80] SECURING MACHINE LEARNING ALGORITHMS. December 2021 ANNEX D: REFERENCES by input data type and lifecycle stages.
[81] Towards Resilient Artificial Intelligence: Survey and Research Issues
[82] The Robustness of Counterfactual Explanations Over Time, A Ferrario et al.
[83] Research priorities for robust and beneficial artificial intelligence.

[84] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. 2016.
[85] Evaluating Robustness of Counterfactual Explanations Artelt et al 2021
[86] Exploring the Trade-off between Plausibility, Change Intensity and Adversarial Power in Counterfactual Explanations using Multi-objective Optimization. Javier Del Ser et al. 2020
[87] Towards Human-Compatible XAI: Explaining Data Di erentials with Concept Induction over Background Knowledge. Widmer et al 2022
[88] A survey on bias in visual datasets 2022. Fabrizzi et al.
[89] Omitted variable bias: A threat to estimating causal relationships. Wilms et al.
[90] Google. Machine Learning Glossary: Fairness. 2021 [cited 29 November, 2021]; available
[91] from: https://developers.google.com/machine-learning/glossary/fairness.
[92] David Lopez-Paz, Krikamol Muandet, Bernhard Scholkopf, and Ilya O. Tolstikhin. Towards a learning theory of cause-e ect inference. In Francis R. Bach and David M. Blei, editors, Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015, volume 37 of JMLR Workshop and Conference Proceedings, pages 1452{1461. JMLR.org, 2015. URL http://proceedings.mlr.press/v37/ lopez-paz15.html.
[93] David Lopez-Paz, Robert Nishihara, Soumith Chintala, Bernhard Sch olkopf, and Leon Bottou. Discovering causal signals in images. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 58{66, 2017. doi: 10.1109/CVPR.2017.14.
[94] ICO, Guidance on the AI auditing framework: draft guidance for consultation. Information Commissioner's Office, 2020.
[95] PwC, PwC Ethical AI Framework. 2020.
[96] Deloitte, Deloitte introduces trustworthy AI framework to guide organizations in ethical application of technology. August 26, 2020. New York.
[97] Orcaa, It's the age of the algorithm and we have arrived unprepared. 2020.
[98] [Epstein18] Epstein, Z., et al., Turingbox: an experimental platform for the evaluation of AI systems. IJCAI International Joint Conference on Artificial Intelligence, 2018. 2018-July: p. 5826-5828. #Discontinued.

[99] Shi et al. Robustness Verification for Transformers. International Conference on Learning Representations. 2020. arXiv:2002.06622
[100] Incremental Bounded Model Checking of Artificial Neural Networks in CUDA Luiz H. Sena et al.
[101] Decoupling feature extraction from policy learning: assessing benefits of state representation learning in goal based robotics A Raffin, A Hill, R Traoré, T Lesort, N Díaz-Rodríguez, D Filliat ICLR 2019 Workshop on Structure & Priors in Reinforcement Learning (SPIRL) 2019
[102] Continual reinforcement learning deployed in real-life using policy distillation and sim2real transfer RT Kalifou, H Caselles-Dupré, T Lesort, T Sun, N Diaz-Rodriguez, D Filliat ICML Workshop on Multi-Task and Lifelong Learning 2019
[103] A Raffin, A Hill, R Traoré, T Lesort, N Díaz-Rodríguez, D Filliat
[104] NeurIPS workshop on Deep Reinforcement Learning 2018
[105] Stable-Baselines3 Reliable Reinforcement Learning Implementations https://stable-baselines3.readthedocs.io/en/master/
[106] T Lesort, M Seurin, X Li, N Díaz-Rodríguez, D Filliat 2019 International Joint Conference on Neural Networks (IJCNN) 2017
[107] Error Analysis tool, part of the Responsible AI Dashboard in Azure: https://learn.microsoft.com/en-us/azure/machine-learning/concept-responsible-ai-dashboard
[108] Evolved from "Towards Accountable AI: Hybrid Human-Machine Analyses for Characterizing System Failure Besmira Nushi Ece Kamar Eric Horvitz HCOM 2018.
[109] Deep Reinforcement Learning that Matters - P Henderson · 2017
[110] L. Hurwicz & S. Reiter (2006) Designing Economic Mechanisms,
[111] Facial Recognition: Analyzing Gender and Intersectionality in Machine Learning. Report. http://genderedinnovations.stanford.edu/case-studies/facial.html#tabs-2
[112] Piggyback: Adapting a Single Network to Multiple Tasks by Learning to Mask Weights. 2018
[113] AS Ross, MC Hughes, F Doshi-Velez. Right for the Right Reasons: Training Differentiable Models by Constraining their Explanations. IJCAI'17.
[114] K Burns, LA Hendricks, K Saenko, T Darrell, A Rohrbach. Women also Snowboard: Overcoming Bias in Captioning Models. ECCV'18, 771-787.
[115] A Rohrbach, LA Hendricks, K Burns, T Darrell, K Saenko. Object Hallucination in Image Captioning. EMNLP'18.
[116] Referencias Glosario
[117] Para el glosario, se han utilizado las siguientes referencias, que pueden ser consultadas para ampliar los aspectos allí descritos.
[118] https://medium.com/analytics-vidhya/abstract-interpretation-for-ai-cc52bc0ddec2
[119] https://arxiv.org/abs/2209.08754
[120] https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc
[121] https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73
[122] http://dmip.webs.upv.es/papers/BFHRHandbook2010.pdf
[123] https://arxiv.org/pdf/1706.08840.pdf

[124] https://support.zivid.com/en/latest/reference-articles/blooming-bright-spots-in-the-point-cloud.html
[125] https://www.geeksforgeeks.org/branch-and-bound-algorithm/
[126] http://dmip.webs.upv.es/papers/BFHRHandbook2010.pdf
[127] http://dmip.webs.upv.es/papers/BFHRHandbook2010.pdf
[128] https://www.statisticshowto.com/cohens-kappa-statistic/#:~:text=What%20is%20Cohen%27s%20Kappa%20Statistic,to%20the%20same%20data%20item.
[129] https://towardsdatascience.com/predict-your-models-performance-without-waiting-for-the-control-group-3f5c9363a7da
[130] https://jmlr.csail.mit.edu/papers/volume2/crammer01a/crammer01a.pdf
[131] https://arxiv.org/abs/1811.00866
[132] https://en.wikipedia.org/wiki/Data_differencing
[133] https://gsarantitis.wordpress.com/2020/04/16/data-shift-in-machine-learning-what-is-it-and-how-to-detect-it/
[134] https://gsarantitis.wordpress.com/2020/04/16/data-shift-in-machine-learning-what-is-it-and-how-to-detect-it/
[135] https://neptune.ai/blog/knowledge-distillation
[136] https://towardsdatascience.com/you-cant-predict-the-errors-of-your-model-or-can-you-1a2e4a1f38a0
[137] https://arxiv.org/abs/2105.04093
[138] https://machinelearningmastery.com/tour-of-ensemble-learning-algorithms/
[139] https://paperswithcode.com/method/experience-replay
[140] https://towardsdatascience.com/the-f1-score-bec2bbc38aa6
[141] http://theory.stanford.edu/~barrett/pubs/LAL+21.pdf
[142] http://theory.stanford.edu/~barrett/pubs/LAL+21.pdf
[143] https://arxiv.org/abs/1809.08098
[144] https://medium.com/geekculture/cross-validation-techniques-33d389897878
[145] https://machinelearningmastery.com/k-fold-cross-validation/
[146] https://medium.com/dataseries/how-modern-game-theory-is-influencing-multi-agent-reinforcement-learning-systems-2a64a3ba0c2c
[147] https://medium.com/geekculture/cross-validation-techniques-33d389897878
[148] https://medium.com/geekculture/cross-validation-techniques-33d389897878
[149] https://www.analyticsvidhya.com/blog/2020/11/binary-cross-entropy-aka-log-loss-the-cost-function-used-in-logistic-regression/#:~:text=Log%20loss%2C%20also%20known%20as,predicted%20probabilities%20and%20actual%20values.
[150] https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwjdgb2Rv5__AhUvUqQEHTxkCnIQFnoECBwQAQ&url=https%3A%2F%2Ftowardsdatascience.com%2Fmicro-macro-weighted-averages-of-f1-score-clearly-explained-b603420b292f&usg=AOvVaw2K03HLxoh-9m77323sHXGE
[151] https://towardsdatascience.com/the-best-classification-metric-youve-never-heard-of-the-matthews-correlation-coefficient-3bf50a2f3e9a
[152] https://www.mydatamodels.com/blog/regression-metrics/

[153] https://moodle.unob.cz/pluginfile.php/42899/mod_resource/content/1/Semiquantitative%20risk%20assessment%2C%20the%20risk%20position%20of%20an%20entity.pdf
[154] https://towardsdatascience.com/a-hitchhikers-guide-to-mixture-density-networks-76b435826cca
[155] https://towardsdatascience.com/cross-validation-k-fold-vs-monte-carlo-e54df2fc179b
[156] https://www.britannica.com/science/mean-squared-error
[157] https://onnx.ai/
[158] https://ieeexplore.ieee.org/document/8514437
[159] https://medium.com/analytics-vidhya/out-of-distribution-detection-in-deep-neural-networks-450da9ed7044
[160] https://www.nature.com/articles/s43586-022-00125-7
[161] https://towardsdatascience.com/mixed-integer-linear-programming-1-bc0ef201ee87
[162] https://towardsdatascience.com/forecast-kpi-rmse-mae-mape-bias-cdc5703d242d
[163] ethics-by-design-and-ethics-of-use-approaches-for-artificial-intelligence_he_en.pdf (europa.eu)
[164] https://homepage.cs.uiowa.edu/~tinelli/papers/BarTin-HBMC-18.pdf
[165] https://www.solver.com/optimization-tutorial
[166] https://towardsdatascience.com/deep-domain-adaptation-in-computer-vision-8da398d3167f
[167] https://christophm.github.io/interpretable-ml-book/counterfactual.html
[168] http://proceedings.mlr.press/v70/oh17a/oh17a.pdf

8.2 Estándares

El presente punto recopila las principales recomendaciones de un conjunto de normas internacionales relacionadas con la solidez de los sistemas de inteligencia artificial, en el marco del Artículo 15 del Reglamento (UE) 2024/1689, que establece la obligación de garantizar que los sistemas de IA sean sólidos, seguros y precisos durante todo su ciclo de vida.

Estas normas proporcionan directrices técnicas para la evaluación de la solidez, la fiabilidad y la calidad de los sistemas de IA, contribuyendo a que los proveedores y desarrolladores puedan demostrar el cumplimiento de los requisitos de solidez mediante prácticas de diseño, pruebas y validación adecuadas.

El conjunto de estándares aquí referenciados ha sido elaborado principalmente en el marco del Comité Técnico ISO/IEC JTC 1/SC 42, especializado en Inteligencia Artificial, así como por otras organizaciones internacionales de normalización (IEEE, ETSI, ITU, DIN y CEN/CENELEC). Algunos de estos estándares han sido ya publicados, mientras que otros se encuentran actualmente en desarrollo o revisión.

Los Estándares Normativos que recogen los contenidos del presente documento son:

[1] ISO/IEC TR 24029-1:2021, Artificial Intelligence (AI) — Assessment of the robustness of neural networks — Part 1: Overview.
[2] ISO/IEC 24029-2:2023, Artificial intelligence (AI) — Assessment of the robustness of neural networks — Part 2: Methodology for the use of formal methods.
[3] ISO/IEC/IEEE 29119-3:2021, Software and systems engineering — Software testing — Part 3: Test documentation.
[4] ISO/IEC 25000:2014, Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Guide to SQuaRE.
[5] ISO/IEC 25030:2019, Systems and software engineering — Systems and software quality requirements and evaluation (SQuaRE) — Quality requirements framework.
[6] ISO/IEC 25059:2023, Software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality model for AI systems.
[7] prEN 18229-2 (en desarrollo), AI Trustworthiness Framework — Part 2: Accuracy and Robustness.

Adicionalmente, también se pueden consultar los siguientes estándares y documentos técnicos relacionados, en los que se abordan aspectos complementarios de robustez, fiabilidad y métricas de evaluación en IA:

[8] DIN SPEC 92001-2:2020-12, Artificial Intelligence — Life Cycle Processes and Quality Requirements — Part 2: Robustness.
[9] ETSI TR 103 821 (INT-008 / AFI), Autonomic network engineering for the self-managing Future Internet (AFI); Artificial Intelligence (AI) in Test Systems and Testing AI Models (Technical Report).
[10] ITU-T F.748.12 (2021), Deep learning software framework evaluation methodology.
[11] ITU-T F.748.11 (2020), Metrics and evaluation methods for a deep neural network processor benchmark.

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