11

Empresas desarrollando cumplimiento de requisitos

Guía 11. Ciberseguridad

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 aplicables, destacan las normas **ISO/IEC 27001:2022 "Information security, cybersecurity and privacy protection – Information security management systems – Requirements"** e **ISO/IEC 27017:2015 "Information technology – Security techniques – Code of practice for information security controls based on ISO/IEC 27002 for cloud services"**, que servirán de base para garantizar la seguridad de la información y la protección de los servicios en la nube en el desarrollo y despliegue de sistemas de inteligencia artificial en el contexto del cumplimiento del Reglamento Europeo de Inteligencia Artificial.

Fecha de revisión: 10 de diciembre de 2025

Contenido general

Índice detallado

1. Preámbulo

1.1 Objetivo del documento

Esta guía está destinada a desarrollar el cumplimiento de las medidas necesarias en materia de ciberseguridad para sistemas de IA, **específicamente en los aspectos relativos** a **inteligencia artificial**, de manera que esta se integre en un esquema de ciberseguridad más amplio. El objetivo principal de esta guía es proporcionar a las empresas los conocimientos y pasos necesarios para implementar los requisitos de ciberseguridad en sistemas de IA, conforme a lo establecido en el artículo 15 del Reglamento Europeo. Su aplicación es responsabilidad tanto del proveedor como del responsable del despliegue del sistema.

1.2 ¿Cómo leer esta guía?

En el apartado introductorio, se abordan lo que entendemos por ciberseguridad para sistemas de IA de alto riesgo, en relación con el Reglamento Europeo de IA. Este apartado sienta las bases conceptuales de la guía.

En el segundo apartado, se establecen cómo entendemos los requisitos establecidos en el Artículo 15: precisión, solidez y ciberseguridad, del Reglamento Europeo de IA. Es importante aclarar que esta guía se centra exclusivamente en los aspectos de ciberseguridad, y no en los requisitos de precisión o solidez, los cuales se tratan de manera específica en otras guías. A lo largo de esta guía, el lector (proveedor o, dependiendo de su alcance, el responsable del despliegue) deberá abordar el apartado con el enfoque siguiente:

- Identificar los activos y actores de su sistema de IA, en relación con el ciclo de vida.
- Asociar activos y actores para poder establecer sus relaciones.
- Identificar las vulnerabilidades a las que están expuestos los activos del sistema de IA.
- Definir y poner en marcha los controles de seguridad para proteger al sistema.
- Revisar periódicamente la efectividad de esos controles.

Para cada una de estas vulnerabilidades y en relación con el objetivo de alcanzar los requisitos.

En el apartado 3 se revisa el articulado relevante del Reglamento Europeo de IA, así como los requisitos que exige.

En el apartado 4 se discute cómo abordar los requisitos enumerados en el apartado 3. En la apartado 5 se establecen las bases de qué debe ser documentado en relación con la ciberseguridad, en línea con el proceso descrito para el apartado 4.

En el siguiente, apartado 6, se establecen una serie de cuestiones genéricas de autoevaluación para el cumplimiento de los requisitos exigidos por el Reglamento Europeo de IA.

En el apartado 7, se plantean las recomendaciones sobre formación y sobre políticas de acceso, que se mencionan a lo largo de las otras secciones de la guía. Además, este apartado incluye el subapartado 7.3 que es la relación de todos los términos y aspectos técnicos indicados a lo largo de la guía.

La guía se cierra con el apartado 8, donde se reúnen las fuentes consultadas para la elaboración de la guía, y recomendado para profundizar en aquellos aspectos más específicos y propios del sistema de IA y su finalidad prevista. Específicamente en el apartado 8.1, se proporciona una relación de estándares, que pueden ayudar a la organización en la consideración un esquema de ciberseguridad más amplio, y que se encuentra fuera del alcance y objetivo de esta guía, centrada en la ciberseguridad para IA, en el marco del Reglamento Europeo de IA.

1.3 ¿A quién está dirigido?

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 protección en el aspecto de ciberseguridad para la IA. 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).

Esta guía está dirigida a sistemas de IA de alto riesgo en fases avanzadas de desarrollo (a partir del nivel de madurez tecnológica TRL 6) y a sistemas que ya están en operación. Por tanto, aplica tanto a sistemas que están próximos a su puesta en servicio como a aquellos que requieren medidas de ajuste o refuerzo durante su funcionamiento.

A lo largo de la guía, se detallan los pasos y recomendaciones necesarios para implementar los requisitos de ciberseguridad específicos establecidos en el Reglamento, tanto por parte del proveedor como del responsable del despliegue del sistema de IA. Es importante señalar que esta guía no aborda todas las medidas generales de ciberseguridad, sino aquellas más destacadas o específicas para sistemas de IA de alto riesgo.

Para temas complementarios, como gestión de riesgos, transparencia, documentación técnica o gestión de calidad, se recomienda consultar las guías correspondientes.

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 en el proveedor del sistema de IA, quien es el responsable de generar y conservar la documentación. Sin embargo, es importante destacar que la obligación de cumplir con los requisitos de ciberseguridad no solo recae en las empresas que desarrollan la IA, sino también en aquellas que la comercializan, implementan o despliegan.

La descripción detallada de los casos de uso utilizados podrá encontrarse en **Guía práctica y ejemplos para entender el Reglamento de IA**

Siempre que se ponga un **ejemplo**, se hará **de manera ilustrativa**. Tanto **proveedores como responsables del despliegue del sistema** deben aplicar las medidas y controles indicados en esta guía.

Los casos de uso se han seleccionado atendiendo a dos motivos:

- La capacidad para explicar la información y procedimientos detallados en la guía, para establecer los criterios necesarios, en el entendimiento de aplicación de los requisitos establecidos en el artículo 15 de precisión solidez y ciberseguridad.
- La relevancia de ambos casos de uso en cuanto al impacto en sufrir un ataque que pueda generar incidentes graves en los derechos de las personas físicas incluyendo personas con riesgo de exclusión social o minorías.

Los casos seleccionados han sido, con las consideraciones indicadas:

- **Sistema automático de concesión de ayudas.**
- **Asistencia al trabajo.**

2. Introducción

2.1 ¿Qué es la ciberseguridad para IA?

La ciberseguridad en sistemas de Inteligencia Artificial de alto riesgo no es una opción, y todos los sistemas se encuentran expuestos a amenazas específicas que requieren medidas de protección rigurosas y adaptadas a su contexto. Estas amenazas incluyen la manipulación de los datos de entrenamiento, que puede comprometer la integridad de los modelos; los ataques diseñados para forzar errores en los resultados, como los ejemplos adversarios; o aquellos dirigidos a obtener información privada de los datos utilizados durante el entrenamiento del sistema. Además, defectos propios del modelo o errores en su integración pueden ser aprovechados por terceros para alterar su funcionamiento o rendimiento. Como se explicará más adelante, siempre se deberá estar vigilante ante la aparición de nuevas amenazas que puedan aparecer de las categorías, mencionadas, o de otras nuevas y en ningún caso se debe considerar esta guía contiene un catálogo cerrado de amenazas y vulnerabilidades.

Vivimos inmersos en un contexto tecnológico en el que las ciberamenazas a los sistemas de la información son variadas, y los riesgos que estas contemplan son amplios. Las organizaciones de toda índole tienen en su más profundo carácter interiorizado el concepto de ciberamenaza, y lo que es más relevante, de ciberseguridad. Todos hemos visto alguna noticia que hace referencia a un *ciberataque* recibido por alguna empresa, o institución que, en el mejor de los casos, detiene el servicio prestado desde unas horas a algunos días. En el peor de los casos, la pérdida de datos, los riesgos para las personas y sus derechos son irreparables.

El ámbito de los sistemas de inteligencia artificial de alto riesgo **no escapa a las amenazas de ciberseguridad generales, *que no son objetivo de esta guía*** y que deberán ser tenidas en cuenta, de la manera adecuada. Estos sistemas de inteligencia artificial además suman y agravan, por su naturaleza, una serie de ciberamenazas propias que exponen sus vulnerabilidades y que son vectores de ataques específicos de los **sistemas de inteligencia artificial,** especialmente en el contexto de los sistemas de alto riesgo. No tener en consideración estas amenazas, y por tanto estar expuesto a ellas, supone **un grave riesgo para cualquier sistema** y, sin lugar a duda, una acción **inadmisible en el contexto de los sistemas de alto riesgo**.

Los daños potenciales que pueden causar los sistemas de alto riesgo para la seguridad y salud de las personas, daños materiales, de privacidad, limitaciones de derechos, discriminación, acceso al empleo y un largo etc., son de una gran magnitud, dado el contexto de consideración de alto riesgo que estos tienen.

Las medidas de ciberseguridad para sistemas de IA presentadas en esta guía tienen como objetivo **mitigar** los **riesgos** que amenazan los **derechos y libertades de las personas físicas** y de la sociedad en general. Estos se pueden ver seriamente amenazados por la existencia de puertas traseras en los modelos, posibilidad ataques de exfiltración o vulnerabilidad a ataques adversarios.

Es por ello importante que esta guía se aborde con la consideración de **riesgos identificados** que amenazan los **derechos y libertades de las personas físicas** y de la sociedad en general, identificados en el plan de riesgos (ver guía de gestión de riesgos) para el sistema de IA, y exista un proceso que comience en los riesgos identificados, se asocie a las vulnerabilidades presentes en esta guía y pueda demostrarse la aplicación de los controles de seguridad aplicables. Todo ello siempre dentro del ciclo de vida del sistema de IA: concepción, diseño, implementación, validación/verificación y puesta en marcha.

La ciberseguridad desempeña un papel crucial a la hora de garantizar que los sistemas de IA sean resistentes a los intentos de alterar su uso, comportamiento y rendimiento o de comprometer sus propiedades de seguridad por parte de terceros malintencionados que exploten las vulnerabilidades del sistema. Los ciberataques contra los sistemas de IA pueden aprovechar los activos específicos de la IA, como los conjuntos de datos de entrenamiento (por ejemplo, envenenamiento de datos) o los modelos entrenados (por ejemplo, ataques de adversarios), o explotar las vulnerabilidades de los activos digitales del sistema de IA o de la infraestructura de TIC subyacente. Por lo tanto, para garantizar un nivel de ciberseguridad adecuado a los riesgos, los proveedores de sistemas de IA de alto riesgo deben adoptar medidas adecuadas, teniendo en cuenta también, según proceda, la infraestructura de TIC subyacente.

Igualmente, el Reglamento Europeo de IA en su artículo 15, indica que la ciberseguridad (orientada a Inteligencia Artificial) deberá ser adecuada y consistente a lo largo del ciclo de vida del sistema, lo que exige de las organizaciones (especialmente del proveedor, pero también del responsable del despliegue) un dimensionamiento adecuado de los esfuerzos en costes y personal. Del mismo modo, considerar la necesidad de **vigilar** adecuadamente las nuevas formas de ataque a los sistemas de inteligencia artificial, es un campo fuertemente cambiante.

2.2 ¿Cuáles son las principales amenazas?

Generalmente agrupadas bajo el concepto de ataques adversarios, a las que se encuentra expuesto un sistema de inteligencia artificial son:

- Ataques de envenenamiento: buscan manipular el sistema de IA durante su fase de entrenamiento o actualización para comprometer su comportamiento. Estos ataques pueden clasificarse en dos tipos principales:

- Envenenamiento de datos: Consiste en la introducción de datos maliciosos o manipulados en el conjunto de datos de entrenamiento. El objetivo es alterar el aprendizaje del modelo para que genere resultados controlados o insertar puertas traseras que puedan ser explotadas posteriormente. Aunque sus efectos suelen manifestarse durante la fase de inferencia, el ataque ocurre principalmente en la fase de desarrollo y entrenamiento del sistema.
- Envenenamiento de modelos: a diferencia del envenenamiento de datos, en este caso el ataque manipula directamente el modelo de IA en sí, generalmente durante su actualización o en entornos de aprendizaje federado. En estos escenarios, el modelo se entrena de manera distribuida en múltiples dispositivos o entidades, lo que aumenta el riesgo de manipulación. El atacante puede modificar los parámetros del modelo o insertar comportamientos maliciosos que comprometan su funcionamiento sin necesidad de actuar sobre los datos de entrenamiento.

- Ataques de evasión: En este tipo de ataque, el adversario intenta que el sistema de inteligencia artificial haga predicciones incorrectas, ya sea de manera global o en casos específicos. Estos ataques se producen en la fase de **inferencia**, cuando el sistema ya está en **producción**, por lo que es esencial que el sistema cuente con mecanismos de seguridad para detectarlos. Las contramedidas deben diseñarse durante las etapas de diseño y entrenamiento para maximizar la resiliencia del sistema ante estos ataques.
- Ataques de inversión: La privacidad es el principal riesgo en este tipo de ataques, pero combinados con otros tipos de ataques (por ejemplo, de extracción) tienen efectos devastadores. El objetivo del atacante es obtener conocimiento de los datos de entrenamiento del modelo. Al igual que otros ataques, se produce en fase de **inferencia** con el **sistema** de IA en **producción**, pero el sistema deberá ser diseñado y entrenado con el objetivo de resistirlos en la mayor medida posible.
- Ataques de extracción: Estos ataques intentan obtener información sobre el modelo, con el objetivo de replicar o entrenar un modelo similar. La forma más común es a través de interacciones con el sistema, aunque también pueden implicar mecanismos laterales, como el robo del modelo o la extracción de información a partir de sus respuestas. Los ataques de extracción pueden facilitar ataques de **transferencia**, en los que, una vez conocido el modelo, es posible aplicar otros tipos de ataques (evasión, inversión) al sistema de IA de alto riesgo atacado. Al igual que otros ataques, estos ocurren en la fase de **inferencia** con el **sistema** de IA en **producción**, pero el sistema debe ser diseñado y entrenado para resistirlos en la mayor medida posible.
- Ataques de canal lateral: Este tipo de ataque busca obtener información del sistema de IA a través de características físicas observables durante su funcionamiento, como patrones de consumo energético, tiempos de respuesta o emisiones electromagnéticas. A diferencia de los ataques directos sobre el modelo o los datos, estos ataques aprovechan vulnerabilidades en la implementación física o el

hardware donde se ejecuta el sistema. Los ataques de canal lateral pueden ocurrir tanto en fase de **inferencia** con el sistema en **producción** como durante el entrenamiento, y pueden facilitar otros tipos de ataques al revelar información sobre la arquitectura o los parámetros del modelo. Es fundamental implementar contramedidas específicas a nivel de hardware y software para minimizar las fugas de información a través de estos canales.

- Ataques a la cadena de suministro: Estos ataques se centran en comprometer la integridad del sistema de IA a través de vulnerabilidades en su cadena de suministro, desde el desarrollo hasta el despliegue. El atacante puede introducir componentes maliciosos o manipulados en cualquier punto del ciclo de vida del sistema, incluyendo:
    - Componentes de software: La introducción de dependencias comprometidas, bibliotecas maliciosas o código vulnerable en el proceso de desarrollo puede crear puertas traseras o vulnerabilidades explotables una vez que el sistema está en producción. Esto es especialmente crítico en sistemas que dependen de múltiples componentes de código abierto o de terceros.
    - Infraestructura y hardware: La manipulación de los componentes físicos o la infraestructura de computación puede permitir ataques persistentes que son difíciles de detectar y mitigar.

  Estos ataques pueden afectar tanto a la fase de entrenamiento como a la de inferencia, y requieren una rigurosa verificación y validación de todos los elementos de la cadena de suministro.

- Ataques de Denegación de Servicio (DoS): Este tipo de ataque busca interrumpir o degradar el funcionamiento normal del sistema de IA, sobrecargando sus recursos computacionales o saturando sus capacidades de procesamiento. A diferencia de los DoS tradicionales que se centran en la infraestructura de red, los ataques DoS específicos para IA pueden explotar características únicas de estos sistemas:
    - Sobrecarga computacional: Consiste en generar consultas específicamente diseñadas para maximizar el consumo de recursos del modelo, por ejemplo, mediante entradas que requieren cálculos especialmente intensivos o que activan los peores casos de rendimiento del sistema. Estos ataques son particularmente efectivos en sistemas de IA que operan en tiempo real o tienen requisitos estrictos de latencia.
    - Ataques de agotamiento de otros recursos: Se centran en explotar limitaciones en los recursos empleados por el sistema inteligente para poder desarrollar la tarea encomendada. Por ejemplo, el sistema puede generar un número masivo de ficheros que comprometan el funcionamiento del sistema operativo. Al igual que otros ataques, ocurren principalmente en fase de **inferencia** con el sistema en **producción**, pero requieren consideraciones especiales durante el diseño y desarrollo para implementar mecanismos de protección efectivos.

El acceso de terceros no autorizados a los activos presentes en la cadena de valor de un sistema de inteligencia de alto riesgo, sin ser una vulnerabilidad directamente relacionada con la IA, si supone un amplio vector de ataque para cualquiera de los ataques anteriormente mencionados, con el riesgo de escalar los ataques de **caja negra** donde el atacante no tiene conocimiento del sistema a **caja blanca o gris** donde hay un conocimiento parcial o incluso total. En esta guía se presenta en un Anexo, la relación entre las políticas de acceso y los sistemas de inteligencia artificial de alto riesgo.

Otro concepto fundamental de ciberseguridad de la IA son los bucles de retroalimentación, particularmente importantes en sistemas de entrenamiento continuo. que continúan aprendiendo después de su despliegue. Estos bucles pueden resultar en la amplificación de sesgos o en un comportamiento no deseado del sistema si los resultados de salida influyen en los datos de entrada de futuras operaciones.

Se suele decir que *"una cadena es tan fuerte como su eslabón más débil"*, y esto es especialmente relevante en el contexto de la ciberseguridad. La **formación** es una herramienta muy potente dentro de una organización para establecer eslabones sólidos. Se aborda en la guía de manera resumida, en otro Anexo dedicado a ello, aspectos relativos a cómo deben las organizaciones (de proveedor y de responsable del despliegue) formar a sus equipos en el ámbito de la ciberseguridad orientada a inteligencia artificial. Esto permite tener una distribución a lo largo de la organización de los conceptos y riesgos inherentes en materia de ciberseguridad para los sistemas de inteligencia artificial de alto riesgo.

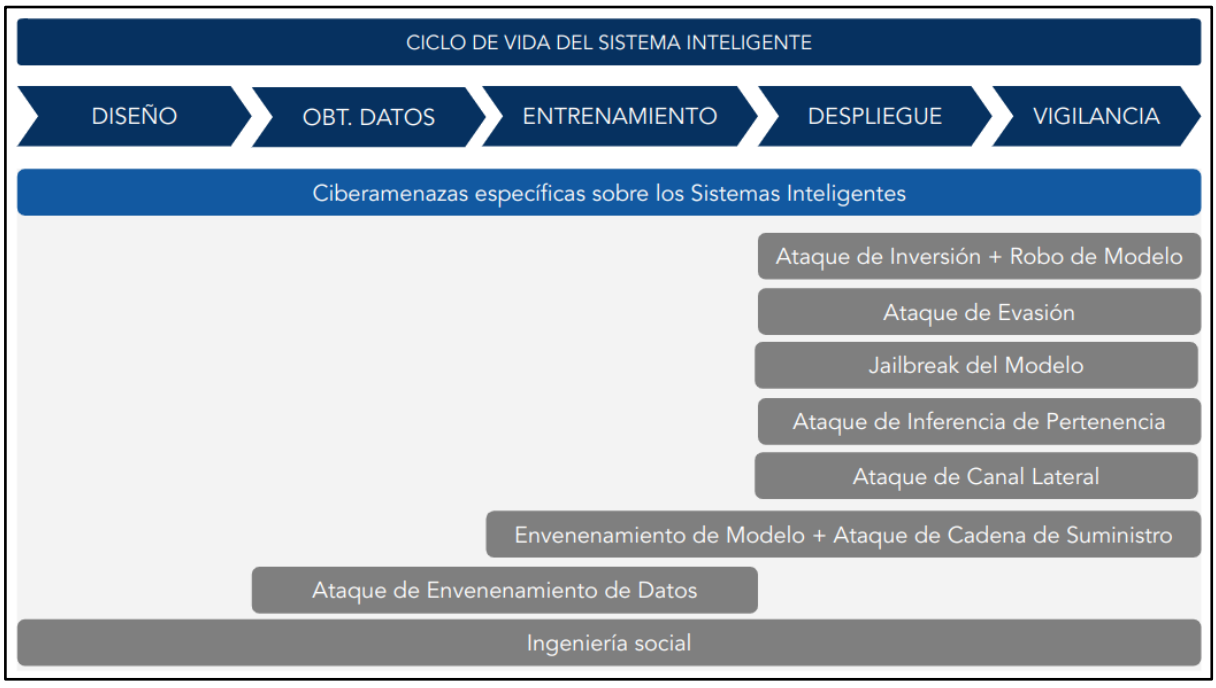

Imagen 1. Esquema de las principales amenazas dentro de la ciberseguridad para sistemas inteligentes relacionando su momento de aparición con el ciclo de vida del sistema inteligente.

3. Reglamento de Inteligencia Artificial

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

El Reglamento 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024 (Reglamento Europeo de Inteligencia Artificial), ya en su considerando (76) establece claramente sus objetivos, indicando que existe una ciberseguridad específica para la inteligencia artificial, estrechamente relacionada con el concepto ya existente, pero con particularidades propias. En este apartado se incluyen los artículos referentes a la generación de ciberseguridad del Reglamento Europeo de Inteligencia Artificial y se detalla en qué secciones de esta guía se abordan los diferentes elementos de dichos artículos.

3.1 Análisis previo y relación de los artículos

El artículo 15 establece los requerimientos que deben cumplirse en materia de tres aspectos fundamentales *"Precisión, solidez y ciberseguridad"*. Precisión y solidez son tratados de manera específica en sus guías.

En el artículo 15 se indica también que estos sistemas tendrán que estar protegidos frente a ataques de **envenenamiento de los datos, ataques de adversario** en todas sus variantes y deberán estar diseñados y desarrollados de tal manera que pueda ser resistentes a la explotación de fallos **por terceros**. Los ataques de terceros podrán ser realizados por usuarios, legítimos o no, los cuales pudieran emplear la capacidad del sistema de IA para fines ilícitos o ilegales que no se correspondan en ningún caso con la finalidad prevista del sistema. Por ejemplo, usuarios legítimos podrían utilizar el sistema para extraer consultas realizadas por usuarios previos.

En esta guía se va a realizar énfasis en las partes de dicho artículo que están orientadas específicamente a ciberseguridad en IA, que son del artículo 15 los puntos 1 y 5.

Artículo 15: Precisión, solidez y ciberseguridad, punto 1 Establece la necesidad de un adecuado nivel de ciberseguridad de manera uniforme durante todo su ciclo de vida.

Artículo 15: Precisión, solidez y ciberseguridad, punto 5 Incide en las medidas y componentes en materia de ciberseguridad específica de la IA sobre los que se desarrolla el contenido a lo largo de la guía.

3.2 Requisitos del 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 artículo 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	Nivel de ciberseguridad y consistencia	Apartado 4.1
15.5	Sistemas resistentes a alteraciones de uso	Apartado 4.2
15.5	Prevenir y controlar la manipulación del conjunto de datos	Apartado 4.3
15.5	Protección frente a ataques adversarios	Apartado 4.4
15.5	Protección frente a los defectos del sistema	Apartado 4.5

4. ¿Cómo abordar los requisitos?

4.1 Nivel de ciberseguridad y consistencia

En el Reglamento Europeo de Inteligencia Artificial, en el artículo 15 relativo a precisión, solidez y ciberseguridad, se indica lo siguiente:

AI Act

Art.15.1 – Precisión, solidez y ciberseguridad

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**.

A lo largo de este apartado vamos a indicar las medidas de alto nivel que se considera que permitirán alcanzar el objetivo descrito en el artículo, teniendo en cuenta que serán más desarrolladas en aspectos específicos en los apartados siguientes. El enfoque de este apartado es permitir al proveedor o al responsable del despliegue conocer el alcance de los controles de ciberseguridad para IA necesarios para cubrir las vulnerabilidades detalladas en la guía que le sean de aplicación. Es importante que se sea consciente que se presenta una información en la guía que en ningún caso pretende ser exhaustiva al respecto de las vulnerabilidades existentes, si no que estas se muestran como una lista de aquellas posibles, y que, dados los ámbitos indicados, (envenenamiento de datos, ataques adversarios o manipulación de defectos) estos se deberán ampliar acorde a las necesidades del sistema de IA para la finalidad prevista. Especialmente con el objetivo de cubrir los riesgos en materia de derechos y libertades.

En cada elemento del ciclo de vida de un sistema de IA de alto riesgo, se establece un nivel de ciberseguridad 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 funcionamiento (responsable del despliegue).

En este contexto se entiende "ciclo de vida del sistema de IA", como la duración para el sistema de IA, desde su concepción, diseño, implementación y puesta en marcha, hasta su retirada. Diferentes fases pueden aplicar a diferentes sistemas de IA dependiendo de su finalidad prevista y forma de comercialización, por lo que en términos generales consideramos estas fases:

- Concepción.

- Diseño.
- Implementación.
- Verificación y prueba.
- Puesta en marcha.
- Vigilancia poscomercialización.
- Retirada.

Este ciclo de vida aquí descrito aplica al sistema de IA, en exclusiva, y en ningún caso se puede considerar representativo de otros elementos, como por ejemplo del ciclo de vida de los datos (que se aborda en la guía de datos), o el ciclo de vida de otros sistemas de información o infraestructura utilizados para soportar el sistema de IA. Del mismo modo, la relación de la ciberseguridad como establecimiento de controles de seguridad está directamente relacionada con el sistema de gestión de riesgos (ver guía de gestión de riesgos), especialmente aquellos relacionados con los derechos y libertades de las personas físicas, daños a la salud, daños graves a la propiedad y/o el medio ambiente. La protección del sistema frente a la tipología de amenazas descritas en esta guía está directamente relacionada con la solidez del sistema de IA de alto riesgo (ver guía de solidez).

4.1.1 Medidas aplicables

Las medidas técnicas se traducirán en **controles de seguridad**, aplicados al inventario de activos para defenderse ante amenazas y prevenir los riesgos, de manera que así sean según el Reglamento Europeo de IA en su Art. 15:

AI Act

> Art.15.5 – Precisión, solidez y ciberseguridad
>
> [...] **adecuadas a las circunstancias y los riesgos pertinentes.**

Una vez definidas y puesta en marcha las medidas, estas deberán **garantizarse y mantenerse durante todo el ciclo de vida**, de manera que el nivel sea el adecuado, no solo a las amenazas y las medidas de seguridad necesarias en el momento de la puesta en marcha del sistema, si no adaptarse y adecuarse a nuevas amenazas que puedan aparecer, sin que se **degrade** el nivel de ciberseguridad.

Proveedor

El proveedor debe tomar las siguientes medidas organizativas para asegurar que el sistema de inteligencia artificial de alto riesgo tiene un nivel de ciberseguridad adecuado.

- Planificar de manera global el nivel de ciberseguridad aplicada a IA durante el diseño y desarrollo, siguiendo los puntos detallados en esta guía.
- Implicar al delegado de protección de datos (DPD), desde el diseño del sistema de IA y para la planificación de la ciberseguridad, como interlocutor dentro del grupo de trabajo establecido para desarrollar la planificación. Deberá, además, estar presente cuando se tomen decisiones específicas de ciberseguridad para comprobar su implicación con la protección de datos.
- Las instrucciones de uso del sistema de inteligencia artificial deberán ir acompañadas de las recomendaciones de alto nivel en medidas de ciberseguridad enfocada a IA, específicas del sistema de IA y su finalidad prevista, para tener en cuenta por el responsable del despliegue. Además, se recomienda que el sistema de IA, a través de la interfaz de interacción para el usuario final disponga de recomendaciones de uso en materia de ciberseguridad, de manera que estén accesibles para el usuario final del sistema. Es posible también incluir un mecanismo de interacción (tipo ventana de confirmación o similar) cada vez que se acceda al mismo que garanticen que las instrucciones en materia de ciberseguridad han sido leídas.
- Dentro del proceso de diseño se deben establecer responsables del seguimiento de las medidas de ciberseguridad aplicada al sistema de IA. Definir un cuadro de mando para el seguimiento de la operación. Se establecerán los indicadores de ciberseguridad para IA obligatorios a incluir en el cuadro de mando para el seguimiento de la operación del sistema de IA. Se establecerá la periodicidad mínima en la que se realizará una auditoría de seguridad del sistema de IA, independientemente de que sea interna o externa.
- Las medidas de ciberseguridad aplicables durante todo el ciclo de vida del sistema de inteligencia artificial, que se detallan en esta guía, serán todas ellas aplicables por parte del proveedor, si este proporciona al responsable del despliegue el sistema de IA como un MLSaaS (*Machine Learning as a Service*) u otras formas de comercialización y puesta en marcha que impliquen una instalación automática por el responsable del despliegue donde no haya configuración del sistema.
- Si el sistema de IA va a ser entregado al responsable del despliegue en un formato *on-premise* o *in-cloud*, gestionado por el responsable del despliegue, el proveedor debe proporcionar unas instrucciones adecuadas para realizar la protección del sistema especialmente durante el tiempo de inferencia del sistema de IA en producción. El proceso de instalación del sistema de IA debe contar con mecanismos que garanticen que la instalación tenga en cuenta las instrucciones de manera obligatoria, bien a través de procedimientos de script automáticos o semi automáticos, la obligación de disponer de las instrucciones abiertas para el proceso antes de continuar, la solicitud explicita de la lectura de estas o la referencia a la lectura especifica a conocer para cada paso del proceso.

- Las actualizaciones del sistema de inteligencia artificial de alto riesgo deben de ser tratadas con todas las medidas aplicables descritas en esta guía.

Como complemento a estas medidas organizativas, el proveedor deberá alinear las siguientes medidas técnicas:

- En los procesos de instalación y/o configuración del sistema de Inteligencia Artificial y en el manual de instrucciones, el proveedor debe incluir información relativa a los riesgos de ciberseguridad específicos del sistema y como este se encuentra protegido.
- Durante el ciclo de vida del sistema de IA de alto riesgo se deben utilizar herramientas que permitan automatizar las pruebas de seguridad. El uso de estas herramientas, desde el inicio del desarrollo del sistema, permite realizar un diseño orientado a seguridad, concibiendo las pruebas de ataque al sistema en paralelo a su desarrollo y ciclo de vida. Existen, tanto en el mercado como variantes de código abierto, una gran variedad de herramientas que cumplen con lo indicado.
- Las actualizaciones del sistema de IA deben desarrollarse aplicando las medidas técnicas descritas en los sucesivos apartados para que el nivel de ciberseguridad aplicable a IA sea consistente, continuo y no se degrade con la aplicación de estas.

Ejemplo - Asistencia al trabajo

Para abordar la ejecución de las medidas de ciberseguridad en IA, dentro del equipo de análisis y diseño del sistema de IA de alto riesgo, se establecen las responsabilidades y las personas que realizarán las acciones. Se han comprobado los perfiles presentes en la organización y se ha valorado la posibilidad de ampliar su personal para cubrir los esfuerzos necesarios para la ciberseguridad enfocada a IA del sistema. El proveedor ha concluido que, con sus actuales recursos de personal asociados a la gestión de riesgos, desarrollo y análisis del sistema de IA, así como el equipo de implantación, pueden llevar a cabo los aspectos necesarios.

El DPD se ha incluido en el equipo de trabajo desde la fase de diseño, asegurando su participación en todas las decisiones de ciberseguridad. También se ha desarrollado un plan integral de gestión de riesgos que contempla específicamente las amenazas relacionadas con sistemas de IA y se ha actualizado el proceso de control de seguridad de entrega de software para incluir pruebas específicas contra estas amenazas (este plan se revisa y actualiza trimestralmente).

Se han implementado herramientas automatizadas de análisis de vulnerabilidades que se ejecutan en cada nueva *release* del sistema y cuando se incorporan nuevos conjuntos de datos de entrenamiento. Estas herramientas incluyen escáneres de seguridad especializados en IA y sistemas de vigilancia continua del modelo. Se ha establecido un programa de

auditorías con revisiones internas trimestrales y una auditoría externa anual obligatoria, documentando los hallazgos y acciones correctivas en el plan de gestión de riesgos.

Tras definir todos los puntos anteriores, el proveedor ha programado una formación para todo el personal implicado, que permite consolidar los conocimientos y el proceso de ciberseguridad en IA dentro de la organización, utilizando como base esta guía y la aproximación indicada en el apartado 7.2 de la presente guía. Se ha incluido en el plan de seguridad del proveedor la formación específica en ciberseguridad para sistemas IA, con actualizaciones semestrales del programa formativo para incorporar nuevas amenazas y contramedidas.

Responsable del despliegue

Las medidas organizativas en materia de ciberseguridad aplicables al responsable del despliegue dependen del nivel de implicación de este en el ciclo de vida del sistema de Inteligencia Artificial de alto riesgo.

- Deberá distribuir la información de instalación y configuración en su organización.
- Si el sistema se encontrase en sus instalaciones, o fuese gestionado por este (on-premise o in-cloud gestionada por el responsable del despliegue), debe establecer, **dentro** de su **política general** de **ciberseguridad** todos los **activos** asociados al sistema de Inteligencia Artificial que pudieran serle de aplicación. Estos podrán ser datos, modelo, procesos, entornos/herramientas y artefactos. Además de la identificación de los actores en su organización sobre cada uno de estos activos. Así sobre el conjunto de activos a las condiciones de uso que haga del sistema de IA, acorde a su finalidad prevista, se aplicarán las consideraciones indicadas en esta guía, especialmente aquellas centradas en fase de **inferencia** con el sistema de IA **en producción**.

Como medida técnica es importante disponer de un repositorio centralizado donde gestionar la información, accesible para aquellos actores identificados en el desarrollo del sistema de inteligencia artificial.

Ejemplo – Sistema automático de concesión de ayudas

Los responsables del despliegue del sistema de IA puedes ser la Administración General del Estado y otras Administraciones Públicas, como Comunidades Autónomas o entidades locales. En cualquiera de los casos, pueden optar por una instalación *on-premise* del sistema. El *proveedor* ha proporcionado, por un lado, un **manual de instrucciones** y por otro, de la **configuración del sistema** (que se realizará on-premise).

En ambos documentos se recoge específicamente la información de ciberseguridad relacionada con ambos procesos. Sobre ella se realizan dos acciones:

- El personal de la Administración Pública encargado de gestionar la respuesta del sistema de IA ha recibido el manual de instrucciones y una formación específica en ciberseguridad para IA acorde a su alcance.
- El personal de administración de sistemas y tecnologías IT relacionado ha recibido el manual de **configuración del sistema** y los aspectos centrados en ciberseguridad. Particularmente en las Entidades Locales, se considera actualizar los conocimientos del personal en el área de ciberseguridad de IA con una formación adaptada a sus conocimientos técnicos. Este personal específicamente recibe también esta guía.

4.2 Sistemas resistentes a alteraciones de uso

Como establece el Reglamento Europeo de IA en el apartado 5 de su artículo 15 relativo a precisión, solidez y ciberseguridad:

AI Act

Art.15.5 – Precisión, solidez y ciberseguridad

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**.

Las medidas descritas en el apartado anterior van acompañadas de la realización de inventarios de los **activos** del sistema de inteligencia artificial durante su ciclo de vida, y los **actores** que interactúan con dichos activos. Gracias a estos inventarios se cubren dos aspectos indicados por el Reglamento Europeo de IA tal y como se ha indicado al principio del punto 5 del artículo 15:

- Establecer la base para proteger el sistema de los intentos de terceros no autorizados de alterar el uso y funcionamiento del sistema de inteligencia artificial. Al disponer de inventarios de activos y actores se pueden aplicar políticas de acceso (ver Anexo I).

- Poder dimensionar sobre los activos identificados soluciones técnicas que garanticen la ciberseguridad, tal y como se describe, identificar las vulnerabilidades sobre estos y sus controles (ver en detalle del subapartado 4.3 Vulnerabilidades y controles de seguridad para datos al subapartado 4.5 Ataques a los defectos del sistema de IA).

4.2.1 Medidas aplicables

Proveedor

Las medidas organizativas del proveedor relacionadas con los inventarios de activos y actores son:

- Se deben inventariar todos los **actores** implicados en el proceso de diseño y desarrollo del sistema de IA. Las acciones de terceros no autorizados no solo se refieren a terceros externos si no a terceros internos con funciones inadecuadas.
- Una vez definidos los **actores**, establecer el nivel de alcance y permisos que cada uno de ellos puede tener y definirlos adecuadamente. Ver 7.1 Anexo I: Políticas de acceso. Una aproximación a sistemas de inteligencia artificial de alto riesgo.
- La documentación del sistema de IA debe reflejar claramente los roles implicados en la utilización de la herramienta, y los permisos que cada uno de ellos deberá tener. Las instrucciones o pasos de instalación deben reflejar con detalle la relación entre los actores y los elementos y acciones de la instalación.
- Planificación y realización del **inventario de activos** en el desarrollo del sistema de Inteligencia Artificial, considerando especialmente las **herramientas, datos, procesos y modelo.**
- Sobre el inventario de activos, se identificarán las **vulnerabilidades** asociadas en esos activos. Esa identificación se realiza en apartados sucesivos de la guía.
- Las instrucciones del sistema de inteligencia artificial deben incluir una **lista de las amenazas/vulnerabilidades** inherentes al sistema durante su utilización.
- Desarrollar un modelo de amenazas riguroso para comprender todos los posibles vectores de ataque y con la consideración del riesgo para la vulneración de derechos y libertades de las personas físicas, daños a la salud, daños graves a la propiedad y/o el medio ambiente alineado completamente con el análisis de riesgos realizado (ver guía del sistema de riesgos).
- El modelo de amenaza establecido, para cada identificación de actores/activos/vulnerabilidades se, recomienda que mantenga un esquema RACI (*Responsible, Accountable, Consulted, Informed).* En los apartados siguientes (del apartado 4.3 al apartado 4.5), se definirán las vulnerabilidades de manera que el proveedor del sistema pueda elaborar un modelo de amenaza riguroso de aplicación a su sistema de inteligencia artificial y establecer los controles técnicos necesarios, añadiendo las características de una matriz RACI tal y como se ha propuesto.

En este proceso de inventariado de los **actores** implicados y **activos del sistema**, y aunque su ámbito más específico está indicado dentro de la guía de datos, se debe implicar al

delegado de protección de datos DPD o su equipo. Así se podrá establecer un inventario los más completo posible sobre el que se pueda establecer el modelo de amenaza.

Las medidas técnicas que el proveedor debe realizar respecto a estos inventarios son meramente de soporte a las organizativas:

- El inventario de activos tendrá un soporte informático adecuado a las dimensiones y capacidades de la organización. Podrá abarcar desde un archivo ofimático o una base de datos simple donde se almacenen los datos de los activos de IA hasta herramientas y soluciones de mercado ya existentes para la gestión de inventarios de activos de IT. Así mismo para dimensionar adecuadamente el sistema, se tendrán en cuenta el nivel de accesos concurrentes y la estacionalidad (aquellos momentos en los que el volumen de accesos puede ser más acusado) para garantizar la disponibilidad del sistema y evitar bloqueos por exceso de intentos de conexión.
- Disponer de los medios técnicos necesarios para que las políticas de accesos establecidas puedan ejecutarse (Ver 7.1 Anexo I: Políticas de acceso. Una aproximación a sistemas de inteligencia artificial de alto riesgo).
- Sistema documental y de gestión de activos indicado, deberá estar centralizado con los inventarios actualizados y disponibles.

Ejemplo – Sistema automático de concesión de ayudas

El proveedor del sistema de IA realiza las acciones de inventario de activos y actores:

- Activos: datos de entrenamiento, prueba y validación; los datos son especialmente sensibles y se utilizaran **datos completamente anonimizados** proporcionados por la Administración General del Estado, estos representan datos históricos de concesión/denegación de ayudas de los últimos 10 años. Modelo seleccionado. Librerías y software que interacciona directamente con el modelo de IA. Herramienta para el control de versiones y entorno de desarrollo. Documentación interna de desarrollo y comercial. Documentación externa.
- Actores: Científicos de datos, Expertos en ML e IA, desarrolladores de software, jefes de proyecto y equipo de sistemas. También ha identificado dentro de su equipo comercial, a aquellos que se encargarán de promocionar el producto.

El proveedor como pyme de nueva creación (*start-up)*, debe especificar los roles múltiples que parte de su personal realiza sobre los activos, construyendo una matriz de relación de los activos y actores. Así, para cada activo, se ha identificado el actor relacionado, indicando si se dispone de acceso. Esta matriz se ha realizado utilizando una herramienta ofimática convencional para registrar los datos y se ha compartido en la organización a través de la Wiki interna colaborativa.

Con esta matriz se ha localizado una solución técnica de mercado que permita implementar las políticas de acceso.

Responsable del despliegue

Para el ámbito de actores y activos, el responsable del despliegue del sistema de inteligencia artificial deberá simplemente conocer aquello que directamente le aplica.

Organizativamente deberá considerar:

- Integrar la documentación proporcionada por el proveedor sobre los roles de los diferentes actores con su organigrama interno. Definir y establecer las funciones y las personas que los cubrirán. Establecer mecanismos de substitución de estas personas cuando estén solos a cargo de alguna función, de manera que bajas inesperadas o la a la marcha de un empleado mantenga el sistema y los procesos sin discontinuidades. Siempre que la organización y el tamaño de la empresa lo permitan, los diferentes roles, acorde a lo definido por el proveedor, se asignarán a personas diferentes.
- Si el formato de uso y comercialización del sistema de IA contemplase que el responsable del despliegue dispusiese del **sistema de IA como un activo** de sus propias instalaciones, tendrá que considerar el sistema de IA como un activo que se debe proteger en materia de ciberseguridad.

No hay medidas técnicas relevantes en este ámbito para el responsable del despliegue más allá de las necesarias para la implantación correcta de las políticas de seguridad (ver Anexo I).

4.3 Vulnerabilidades y controles de seguridad para datos

La identificación de vulnerabilidades asociadas a los datos de entrenamiento y el establecimiento de los controles de seguridad necesarios para evitar su alteración o manipulación permiten proteger al sistema de inteligencia artificial de ataques de envenenamiento[1].

El Reglamento Europeo de IA, pone especial foco en las vulnerabilidades asociadas a datos cuando establece en su artículo 15.5 relativo a precisión solidez y ciberseguridad:

AI Act

Art.15.5 – Precisión, solidez y ciberseguridad

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»)

El objetivo del atacante es alterar el entrenamiento del sistema de inteligencia artificial con la intención de controlar sus resultados o comprometer su funcionamiento. Los cambios no intencionados en los datos de entrenamiento se abordan en la guía de datos. En el manual de uso del sistema, el proveedor debe enumerar las vulnerabilidades asociadas a los datos. Se debe identificar, para cada vulnerabilidad, el control de seguridad aplicado. Todas las vulnerabilidades y los controles asociados a los datos se presentan principalmente en la fase de diseño y entrenamiento del sistema, donde el riesgo de introducción de datos maliciosos es muy elevado y el impacto muy alto.

4.3.1 Medidas aplicables

Proveedor

El proveedor debe identificar las vulnerabilidades aplicables a los conjuntos de datos de entrenamiento. Se debe definir la lista, en base a la finalidad prevista del Sistema de IA. A continuación, se presenta una lista de vulnerabilidades asociadas a la manipulación de datos de entrenamiento, validación y prueba:

- Modelo fácil de envenenar

[1] Las palabras y descripciones destacadas se corresponden con términos que son desarrollados en el glosario

- Datos insuficientes para incrementar la resistencia al envenenamiento.
- Gestión de derechos de acceso a los datos inadecuada
- Gestión de los datos inadecuada
- Fuentes de datos no controlados
- Falta de control para detectar envenenamiento en el sistema
- No detectar muestras envenenadas en el conjunto de entrenamiento.
- Uso de fuente de etiquetado de datos incorrecto

Queda fuera del ámbito de esta guía, pero es importante que el proveedor tenga en cuenta, dentro del ciclo de vida propio de los datos, la eliminación segura de estos, una vez que deja de estar en producción o el sistema entra en un evolutivo. En la guía de datos, se trata de manera especial este aspecto.

En la tabla siguiente, se indican algunas de las vulnerabilidades más relevantes y los controles de seguridad aplicables para mitigar su explotación, en el momento de creación de esta guía. Dado el carácter dinámico y la rápida evolución de las amenazas en ciberseguridad, es recomendable estar siempre al tanto de la información técnica más actualizada y de referencia. Entre los recursos recomendados se encuentran el OWASP Machine Learning Security Top Ten, el OWASP Top 10: LLM & Generative AI Security Risks y el OWASP AI Top Ten, que proporcionan análisis y medidas actualizadas frente a riesgos emergentes en sistemas de inteligencia artificial.

Vulnerabilidad	Controles de seguridad
Modelo fácil de envenenar	En fase de diseño, escoger o definir un modelo más resiliente. **Utilizar técnicas de bagging, boosting, o el algoritmo TRIM** durante el entrenamiento para fortalecer la robustez del modelo. Para mitigar el efecto de la aleatoriedad y asegurar la reproducibilidad, utilizar semillas iniciales definidas cuando sea necesario.ISO27001/2 - NIST 800-53: Realizar auditorías periódicas y regulares sobre la gestión datos y realizar planes de acción para corregir las deficiencias.
Datos insuficientes para incrementar la resistencia al	Ampliar el conjunto de datos utilizando técnicas de aumento de datos, para sets demasiado pequeños o insuficientes para la finalidad prevista

Vulnerabilidad	Controles de seguridad
envenenamiento.	
Gestión de derechos de acceso a los datos inadecuada	Políticas de acceso adecuadas, ver Anexos 5.1 Anexo I: Políticas de acceso. Una aproximación a sistemas de inteligencia artificial de alto riesgo.
Fuentes de datos no controlados	Todas las fuentes de datos deben mantenerse bajo control de versiones, para disponer de una traza de los cambios en los datos y los actores que los han realizado.
Fuentes de datos no controlados (continuación)	ISO27001/2 - NIST 800-53: Se proponen diversos mecanismos. Clasificar los datos adecuadamente utilizando herramientas estadísticas, considerando la finalidad prevista y revisar adecuadamente la clasificación; los datos deben estar siempre cifrados en tránsito, utilizar cifrados con garantía de autenticidad, cifrando los datos utilizando cifrados acordes a los estándares vigentes; almacenar siempre cifrado en reposo las fuentes de los datos, a nivel de disco duro y de cifrado de archivo individual, igualmente usando estándares vigentes. Utilizar herramientas de prevención de pérdida de datos (DLP) en aquellos datos sensibles y relevantes.
Falta de control para detectar envenenamiento en el sistema	**Aplicar la técnica STRIP**, que implica perturbar los datos de manera controlada y observar la variancia entre los datos perturbados y no perturbados para identificar posibles envenenamientos.

Vulnerabilidad	Controles de seguridad
No detectar muestras envenenadas en el conjunto de entrenamiento	Utilizar técnicas de saneado de datos para detectar y eliminar los outlier: Z-score, Local Outlier Factor, Bosques de Aislamiento (Isolation Forest). Antes de eliminar cualquier dato, realizar un diagnóstico exhaustivo para **evitar la eliminación de datos válidos** en intervalos extremos: 1. Verificar previamente la hipótesis sobre la distribución de la variable y, si la distribución no es normal, evitar el uso de Z-score. 2. Diagnosticar la naturaleza del outlier para decidir si se debe eliminar o conservar Para detectar cambios en las muestras, certificar la distribución de los datos de entrenamiento por marginales (distribución estadística empírica de cada variable de la base de datos, ya sea numérica o no), y multivariada (por relaciones bivariantes o más complejas) periódicamente verificar que se mantiene estable. Ante cambios importantes en la distribución de una de las variables o de la relación entre varias, profundizar para identificar los cambios acontecidos en los datos Analizar el impacto de los sets de datos en la exactitud del modelo, en procesos de re-entrenamiento o aprendizaje continuo: RONI (reject on negative impact) o tRONI (público-aware RONI)
Uso de fuente de etiquetado de datos incorrecto	ISO27001/2 - NIST 800-53: Gestión de todas las interconexiones con sistemas externos de los que se reciban datos y revisión. Establecer un proceso formal para realizar una revisión de todas las interconexiones y regular y vigilar su utilización. Usar técnicas de protección frente a intercambio de etiquetas (Label Flipping), basadas en la identificación por k-NN (k-nearest neighbours) en el data-set. Esta técnica puede utilizarse cuando los conjuntos de datos etiquetados no sean confiables, pero no exista otra alternativa en los mismos.

Tabla 1 Ataques Envenenamiento: Vulnerabilidades y controles de seguridad

Todos los controles y vulnerabilidades asociadas a datos se focalizan principalmente a la fase de **diseño y entrenamiento del sistema**, donde el riesgo de introducción de ataque por envenenamiento es muy elevado y el impacto muy alto.

En el **manual de uso del sistema** de IA, el proveedor debe enumerar las vulnerabilidades asociadas a los datos de entrenamiento, identificando para cada vulnerabilidad el **control de seguridad** aplicado. En el caso de que el responsable del despliegue del sistema de IA de alto riesgo deba realizar parte del entrenamiento antes de su puesta en marcha, se deberá detallar la manera de aplicar los controles de seguridad.

Ejemplo – Asistencia al trabajo

El sistema va a trabajar con datos biométricos obtenidos de los empleados para su entrenamiento, prueba y validación, de tal manera que sea capaz de identificar adecuadamente a los empleados desde las diversas ubicaciones de las cámaras (o sensores biométricos) para registrar su asistencia al trabajo.

En el análisis de riesgos se ha identificado la posibilidad de que un usuario actúe de manera malintencionada para alterar las etiquetas de los datos biométricos. Esto podría permitirle asociar sus propios datos con los de otros empleados, comprometiendo así la detección del sistema.

Dado que el principal origen de esta amenaza es el envenenamiento de datos, durante el entrenamiento se utilizan técnicas de perturbación STRIP, específicamente destinadas para detectar disparadores de envenenamiento (o troyanos) en el conjunto de datos.

Adicionalmente, el equipo de diseño decide proteger al sistema de IA contra un posible ataque por Label Flipping (ver Glosario) en el que el atacante haya cambiado malintencionadamente etiquetas de imágenes de diferentes empleados.

Responsable del despliegue

La principal acción organizativa que debe realizar el responsable del despliegue para abordar la protección de sistemas de inteligencia artificial frente a ataques de envenenamiento es realizar una lectura y compresión del manual de instrucciones. Como norma general, el responsable del despliegue no participa de la fase de diseño, desarrollo y entrenamiento del sistema.

No obstante, si el formato de uso y comercialización del sistema de IA contemplase que el responsable del despliegue utilizase el sistema de IA realizando un **entrenamiento final** del mismo, o en los sistemas de entrenamiento continuo, estando este previamente preentrenado parcialmente por el proveedor, le serían de aplicación todas las medidas de ciberseguridad aplicada a IA del proveedor en materia de los datos. En este escenario, el responsable del despliegue **pasaría a tener responsabilidades de proveedor** (organizativas y técnicas) en todos los aspectos descritos en el anterior apartado.

4.4 Protección frente a ataques adversarios

Se ha separado en esta guía los ataques adversarios de tipo envenenamiento de datos, de aquellos relacionados con el propio modelo. Además, el sistema está más expuesto durante la fase de **inferencia**, una vez se encuentra **en producción** (tras la puesta en marcha) a este tipo de ataques. A diferencia de la exposición relativa a datos.

El Reglamento Europeo de IA en el artículo 15, relativo a precisión solidez y ciberseguridad, párrafo 4 establece:

Art.15.5 – Precisión, solidez y ciberseguridad

Medidas para prevenir, detectar, combatir, resolver y controlar **los ataques** que traten de manipular *[...]* **la información de entrada** diseñada para hacer que **el modelo de IA cometa un error («ejemplos adversarios» o «evasión de modelos») [...]**

Para ello se deben establecer los **controles de seguridad** necesarios para evitar que la introducción de determinados datos de entrada haga cometer un error al modelo o permita extraer información de este. Se trata de ataques adversarios de tipo oráculo, extracción, inversión o ataques de evasión en el sistema de inteligencia artificial de alto riesgo.

4.4.1 Medidas aplicables

Proveedor

El proveedor debe identificar las vulnerabilidades relacionadas con la manipulación de los datos de entrada, las cuales pueden hacer que el modelo cometa errores o permitir la inferencia de los datos usados en el entrenamiento a través de consultas al modelo. Se debe evaluar y priorizar la lista de vulnerabilidades en función de la finalidad prevista del sistema de IA. A continuación, se presenta una lista de vulnerabilidades:

- Transferencia de ataques adversarios.
- Gestión inadecuada de los derechos de acceso al modelo
- Falta de consideración de los posibles ataques a los que el sistema de IA puede estar expuesto.
- Ausencia de un proceso de seguridad que mantenga el nivel de protección de los componentes del sistema de IA.
- Mecanismo de débil de control de accesos a los componentes del sistema de IA.
- Falta de detección de entradas de datos anómalos, aumentando la exposición a los ataques de pertenencia para inferir datos o ataques de inversión.
- No considerar en el diseño e implementación del modelo la posibilidad de ataques de evasión.
- No considerar en el diseño e implementación del modelo la posibilidad de ataques adversarios, por ejemplo: **clasificación errónea, clasificación errónea origen/objetivo, clasificación errónea aleatoria y reducción de confianza.**
- Uso de modelos ampliamente conocidos en el sistema de IA, lo que permite la réplica total o parcial y, por tanto, facilita la transferencia de ataques adversarios.
- Entrada de datos totalmente controlada por el atacante (sin preprocesamiento o por vulneración del mecanismo de entrada), lo que permite crear pares de datos de

entrada y salida. Exposición a ataques de inversión. Tanto ataques de inferencia de pertenencia (*Membership Inference Attack*), que permitan saber si unos datos de entrada han formado parte o no del set de datos de entrenamiento, como ataques de inferencia de propiedades (*Property Inference*) que busca conocer los datos usados en el modelo. Este riesgo afecta a la privacidad de los datos personales utilizados en el entrenamiento y, por tanto, debe tenerse en cuenta como riesgo de ciberseguridad, según establece la norma ISO7IEC 27001. Ambos ataques pueden preceder o ir dirigidos a conseguir la **reconstrucción** del conjunto completo de datos de entrenamiento.

- Exceso de información disponible en el modelo. Exceso de información sobre el modelo en sus salidas.
- Falta de consideración de la posibilidad de ataques de extracción.
- Robo o compromiso del modelo.

En el manual de uso del sistema, el proveedor debe enumerar las vulnerabilidades asociadas a la manipulación de los datos en los que el sistema se encuentra expuesto. Aquellas medidas que han sido puestas en marcha durante el entrenamiento del modelo deben de ser indicadas. Si el sistema no se ofrece como un MLSaaS, o por su naturaleza pudiera encontrase físicamente en las instalaciones del responsable del despliegue, el manual de instrucciones y su documentación tendrán que hacer referencia a las vulnerabilidades de aplicación y a la forma de aplicar aquellos controles de seguridad que puedan ser de su responsabilidad.

Las medidas técnicas aplicables en este caso son los controles de seguridad que impiden o mitigan la explotación de las vulnerabilidades. Si algunas de las medidas **deben ser realizadas en el proceso de instalación y utilización** del sistema por parte del responsable del despliegue, se debe indicar en el manual de instrucciones.

A continuación, se muestra una lista detallada de vulnerabilidades:

Vulnerabilidades	Controles de seguridad
Transferencia de ataques adversarios	En la medida de lo posible, se debe evitar utilizar modelos en el sistema de IA que sean de amplio uso y conocimiento, ya que estos pueden contener vulnerabilidades conocidas que faciliten ataques de transferencia. También es recomendable evitar partir de modelos de base o frameworks de trabajo en versiones ya comprometidas, que permitan identificar el ataque y explotar esta vulnerabilidad. Todas las familias de modelos y sistemas de IA tienen vulnerabilidades conocidas o aspectos en los que son más

Vulnerabilidades	Controles de seguridad
	susceptibles a ciertos tipos de ataques adversarios. Por ello, es fundamental conocer, identificar y proteger el sistema de IA desde la fase de diseño, implementación, verificación y validación hasta su fase de puesta en marcha y funcionamiento. Además, para reducir el riesgo asociado al uso de modelos ampliamente conocidos: - Se recomienda aplicar técnicas de ofuscación o enmascaramiento que dificulten la identificación y explotación del modelo por parte de un atacante. - Evaluar de manera proactiva la robustez del sistema frente a ataques transferidos mediante pruebas específicas y metodologías de evaluación continua.
Gestión inadecuada de los derechos de acceso al modelo	Políticas de acceso, ver apartado 7 Anexos 7.1 Anexo I: Políticas de acceso. Una aproximación a sistemas de inteligencia artificial de alto riesgo.
Falta de consideración de los posibles ataques a los que el sistema de IA puede estar expuesto	Integrar las especificidades de seguridad para sistemas de IA en la estrategia de sensibilización y asegurarse que todas las partes implicadas lo reciben (ver Anexo II: Formación en ciberseguridad y sistemas de inteligencia artificial de alto riesgo). Seguimiento del estado del arte de los ataques adversarios: https://csrc.nist.gov/publications/detail/nistir/8269/draft
Ausencia de un proceso de seguridad que mantenga el nivel de protección de los componentes del sistema de IA	ISO27001/2 - NIST 800-53: Realizar auditorías periódicas y regulares sobre el modelo en base a las medidas detalladas en esta guía, y realizar planes de acción tras cada caso. Realizar una aproximación equipo azul /equipo rojo; el equipo azul desarrolla el modelo y el equipo rojo lo pone a prueba con diversos formatos de ataque de tipo adversario al mismo (caja blanca, gris, negra).
Falta de detección de entradas de datos anómalos	Implementar herramientas específicas para detectar si un valor de entrada puede ser un ejemplo adversario. Una estrategia efectiva es añadir al sistema subredes de detección de adversarios, que funcionan de manera

Vulnerabilidades	Controles de seguridad
	separada al modelo principal. Estas subredes se entrenan específicamente para identificar patrones asociados a ataques adversarios, analizando características particulares de las entradas sospechosas y alertando sobre posibles intentos de manipulación. Además, se deben implementar métricas y umbrales de confianza en las clasificaciones y en la distribución de los datos de entrada. Configurar niveles de confianza permite generar alertas cuando los resultados del modelo muestran incertidumbre o valores alejados del comportamiento esperado, indicando posibles ataques o anomalías en las entradas. Del mismo modo, establecer umbrales sobre distribuciones estadísticas ayuda a identificar desviaciones en los datos, como outliers, que podrían estar relacionados con intentos de ataques adversarios. Otras medidas complementarias incluyen: - Mecanismos de detección de elementos OoD (Out of Domain data): Identificar datos de entrada situados en las fronteras de los dominios de clasificación o fuera del espacio de solución previsto por el modelo. Para sistemas no basados en clasificación, localizar y analizar outliers, teniendo en cuenta la relación de estos con otras variables relevantes del sistema. - Control de orígenes de llamadas al sistema: Monitorizar y detectar solicitudes sospechosas que intenten explorar o manipular el espacio de solución del modelo. Los accesos provenientes de estos orígenes deben ser bloqueados temporal o permanentemente, o ralentizados mediante el aumento de los tiempos de respuesta. - Niveles de confianza y mecanismos de alerta: Establecer un mecanismo que asocie cada resultado del modelo a un nivel de confianza. Los valores bajos de confianza deben activar alertas automáticas para advertir de entradas potencialmente anómalas.

Vulnerabilidades	Controles de seguridad
	En la interfaz del sistema, se debe incluir: - Registros de entradas anormales: Un sistema de trazabilidad que permita registrar y analizar las entradas sospechosas. - Alertas visuales y avisos al usuario: Facilitar que los usuarios sean advertidos de posibles entradas anómalas en tiempo real. Para facilitar la implementación, se pueden integrar herramientas open-source o comerciales especializadas en la detección de entradas anómalas y ejemplos adversarios. Se listan a continuación algunos ejemplos de herramientas open-source: • Adversarial Robustness Toolbox (ART) de IBM: Un framework open-source que proporciona métodos para evaluar y mejorar la robustez de los modelos frente a ejemplos adversarios. • Foolbox: Herramienta open-source para pruebas de robustez y generación de ejemplos adversarios en modelos de aprendizaje automático. • TensorFlow Privacy y PyTorch-based libraries: Soluciones que ayudan a aplicar técnicas de detección y mitigación, como análisis de confianza y monitoreo de outliers.
No considerar en el diseño e implementación del modelo la posibilidad de ataques de evasión	Diseñar un modelo resiliente contra ataques de evasión mediante aleatorización de entradas, **regularización del gradiente de entrada y destilación defensiva** (defensive distillation) Utilizar recursos de referencia para el estado del arte de los ataques de adversarios de referencia de las que existen diferentes formatos, muchas de ellas de código abierto. Estos conjuntos de herramientas proporcionan conjuntos de técnicas adecuadas para verificar el modelo adecuadamente frente a ataques adversarios.

Vulnerabilidades	Controles de seguridad
No considerar en el diseño e implementación del modelo la posibilidad de ataques de evasión	MITRE ATLAS (https://atlas.mitre.org/) es una extensa base de conocimiento con información actualizada sobre las amenazas y casos de estudio basado en observaciones de mundo real y de sistemas de inteligencia artificial operativos.
Clasificación errónea y clasificación errónea aleatoria	Los atacantes generan una muestra que no está en la clase de entrada del clasificador objetivo pero que es clasificada por el modelo como esa clase de entrada particular. En la variante aleatoria el resultado puede ser cualquiera diferente de la clasificación correcta. Utilización del framework Highly Confident Near Neighbor (HCNN). Las explicaciones de decisiones individuales basadas en las atribuciones realizadas por el sistema de IA (Attribution-driven Causal Analysis).
Clasificación errónea origen/objetivo	El objetivo del ataque es obtener un resultado específico (o clasificación) para una entrada dada. Para mitigarlo: Eliminación de características: Feature Denoising Una técnica que mitiga especialmente este tipo de ataques es incluir en el set de entrenamiento ejemplos adversarios conocidos de manera que el sistema se entrene con ejemplos adversarios construidos para tal fin. Compresión o exprimido de las características (feature squeezing).
Reducción de confianza	Este tipo de ataque está destinado a reducir la confianza del sistema en una clasificación correcta, para que de esta manera se influya en la percepción que actúa en el proceso de supervisión. Puede mitigarse mediante el control de las entradas, con la disminución de entradas admitidas al sistema en un formato específico o procedentes de una dirección IP, origen o demarcación sospechosa. Estas entradas podrán bloquearse en caso de riesgo.

Vulnerabilidades	Controles de seguridad
Uso de modelos ampliamente conocidos en el sistema de IA facilitando la transferencia de ataques adversarios	Utilizar modelos con menos riesgo de transferencia. Durante el proceso de elección del modelo, conocer y analizar los riesgos de transferencia o, en caso general, su popularidad y uso generalizado, para seleccionar aquel que tenga un riesgo menor de transferencia.
Uso de modelos ampliamente conocidos en el sistema de IA facilitando la transferencia de ataques adversarios	No reutilizar modelos directamente desde internet, sin su comprobación y validación. En la medida de lo posible y equilibrando la finalidad prevista y protección, seleccionar modelos en los que la exposición de ataques de adversarios, siendo conocido, tenga controles de seguridad establecidos.
Datos de entrada total o parcialmente controlados por el atacante	Aplicar modificaciones a los datos de entrada, realizando operaciones de preprocesado: en la medida que no comprometa la exactitud del sistema, realizar una normalización de los datos, eliminación de ruido de los datos. La eliminación de valores fuera de los rangos esperados puede mitigar el control de los datos de entrada por parte del atacante. Los sistemas NLP, pueden ser muy vulnerables frente a la introducción de modificaciones en los datos de entrada. La generación de textos adversarios durante el entrenamiento utilizando, por ejemplo, CAT-Gen. Mantener un histórico de datos de entrada que permita detectar datos de entrada extremadamente cercanos, y levantar una alarma en el sistema por niveles (registro, notificación y bloqueo). Si los datos se introducen de manera directa al modelo por parte del posible atacante, esta introducción de datos debe estar controlada por políticas de acceso adecuadas (ver Anexos 7.1 Anexo I: Políticas de acceso. Una aproximación a sistemas de inteligencia artificial de alto riesgo). El acceso a introducir datos directamente al modelo no debe ser

Vulnerabilidades	Controles de seguridad
	público salvo que sea estrictamente necesario, y en ese caso, deberán controlarse como se ha indicado.
Vulnerabilidad a ataques de inferencia de pertenencia (Membership Attacks)	La investigación y artículos publicados en este ámbito indican que la utilización de técnicas de privacidad diferencial supone una mitigación efectiva a estos tipos de ataques. También es posible crear conjuntos de datos privados, diferencialmente derivados de datos desprotegidos. Para ello se pueden utilizar herramientas que permitan estas acciones. La utilización de Neural dropout o Model stacking puede incrementar la resiliencia del sistema de IA frente a este ataque.
Ataques de Inferencia de Propiedades (Property Inference)	Establecer un límite por origen (IP, usuario, dispositivo) y disponer de mecanismos de alarma y bloqueo de orígenes. Todas las entradas de los datos al modelo accesibles deben tener un mecanismo de validación.
No considerar la existencia de posibilidad de ataques de extracción	Al igual que los ataques de transferencia, se debe controlar la información que retorna el sistema, así como los orígenes con *demasiadas* consultas, pues es un vector de ataque para los ataques de extracción. Existen herramientas que, dentro de sus capacidades, permiten generar defensas modificando la salida del modelo haciendo redondeos o añadiendo ruido estadístico, para dificultar así al atacante el conocimiento del razonamiento del modelo. Se recomienda la utilización de estas herramientas para este tipo de control de seguridad.
Robo del modelo	Control de versiones del modelo, para evitar que se pueda modificar de manera malintencionada el modelo en cualquier punto del ciclo de vida desde la concepción y el diseño, entrenamiento y validación como en producción.

Vulnerabilidades	Controles de seguridad
	Como se ha indicado en el caso de ataques de transferencia, es conveniente que la exposición del modelo (incluso, la publicación de su código fuente) no suponga una vulnerabilidad en sí misma. Esto se consigue aplicando los criterios de protección frente a otras vulnerabilidades a lo largo del sistema, de tal manera que incluso en un nivel de exposición del modelo no suponga una vulnerabilidad.

Tabla 2 Ataques adversarios al modelo: vulnerabilidades y controles de seguridad

Ejemplo – Sistema automático de concesión de ayudas

Durante el desarrollo de del sistema de IA, el proveedor ha establecido un equipo de desarrollo del sistema y otro destinado a ponerlo a prueba, un proceso basado en equipo azul/equipo rojo, de tal manera que cada iteración del modelo (equipo azul) es puesta a prueba (equipo rojo).

Para ello el equipo rojo ha diseñado, herramientas específicas que permiten automatizar las pruebas de seguridad, una serie de pruebas al sistema de IA que son verificadas cada vez que el equipo azul realiza un cambio en el modelo. Estas pruebas automatizadas replican los ataques a vulnerabilidades conocidas y se deben actualizar continuamente cuando se vayan conociendo nuevas vulnerabilidades que potencialmente podrían ser explotadas.

Se ha puesto el foco también en el control de datos de entrada, dado que son los solicitantes los que introducen los datos en el sistema. Se dispone en la recepción de los datos de un mecanismo de control OoD (Out of Domain data). Se ha seleccionado este procedimiento debido a que los datos introducidos se encuentran dentro de secuencias de valores estructurados y tabulados, con valores numéricos o series discretas y permiten establecer controles sobre entradas sospechosas.

Dado que el sistema se ha entrenado con datos proporcionados por la Administración General del Estado, sobre históricos de datos anonimizados los últimos 10 años, se ha identificado la vulnerabilidad de ataques de inferencia de pertenencia como un riesgo relevante a pesar de la anonimización. Para proteger al sistema se ha recurrido a técnicas de privacidad diferencial. El conjunto de datos inicial ha sido ampliado utilizando herramientas que permiten aumentar los datos de inyectando ruido en los mismos. Se seleccionan estas herramientas, por ser adecuadas cuando se trata de evitar el riesgo de ataques de inversión.

Considerando la vulnerabilidad clasificación errónea origen/objetivo, que permitiría conseguir que un solicitante pudiera obtener la ayuda con un patrón malicioso de entrada, se ha aplicado la técnica Feature Denoising a los datos de entrada. El modelo además ha sido entrenado con ejemplos adversarios.

Responsable del despliegue

En este ámbito de ataques el **modelo de comercialización** del sistema de inteligencia artificial de alto riesgo **determina** el alcance de las acciones relativas a ciberseguridad que debe realizar el responsable del despliegue.

En cualquier caso, y según el modelo de comercialización del sistema de IA estas estarán **claramente** definidas en el **manual de instrucciones**. Organizativamente el responsable del despliegue deberá conocer y analizar las vulnerabilidades y los controles que puedan ser de su responsabilidad y asignar recursos humanos y técnicos para cubrirlas.

4.5 Ataques a los defectos del sistema de IA

En el párrafo 5, artículo 15 relativo a precisión solidez y ciberseguridad, del Reglamento Europeo de IA, se pone de manifiesto que los defectos del modelo deben ser tenidos en cuenta a la hora de considerar:

AI Act

Art.15.5 – Precisión, solidez y ciberseguridad

… prevenir, detectar, combatir, resolver y controlar **los ataques** que traten de manipular *[…]* **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**.

Los ataques dirigidos a descubrir y explotar defectos del sistema de IA aprovechan vulnerabilidades presentes en dos niveles: los **defectos propios del modelo** utilizado, y, los **defectos derivados de su integración en el entorno** software que lo rodea y como ha sido **configurado**.

- Defectos propios del modelo: Son vulnerabilidades intrínsecas al diseño, entrenamiento o arquitectura del modelo de IA. Pueden surgir debido a errores en el diseño del modelo, datos de entrenamiento incompletos o sesgados, o configuraciones inapropiadas de los hiperparámetros. Estos defectos permiten a un atacante manipular las predicciones del modelo, generar resultados controlados o forzar errores sistemáticos.
- Defectos en la integración del modelo: Surgen cuando el modelo se incorpora al ecosistema software o hardware en el que opera, debido a configuraciones incorrectas, dependencias inseguras, o fallos en la implementación y despliegue. Estos defectos pueden ser explotados para alterar la funcionalidad del sistema, provocar fugas de información o facilitar otros tipos de ataques, como los de evasión y extracción

Estos defectos pueden detectarse y explotarse mediante ataques de oráculo, que consisten en interactuar con el sistema para inferir su funcionamiento interno y descubrir sus vulnerabilidades. Los ataques de oráculo, a su vez, pueden servir como punto de partida para variantes más específicas, como los ataques adversarios (evasión, extracción o inversión) ya descritos en esta guía.

4.5.1 Medidas aplicables

Proveedor

En lo referente a la explotación de defectos de modelo, el proveedor es responsable de identificar y mitigar las vulnerabilidades que el modelo pueda tener, y en su integración con el ecosistema tecnológico. Este proceso debe abarcar todo el ciclo de vida del sistema, desde su diseño y desarrollo hasta su implementación y operación, asegurando una revisión continua de posibles defectos. El conjunto de medidas aplicadas no solo fortalece la seguridad, sino que también cumple con los requisitos del artículo 15 del Reglamento.

A continuación, presentamos una lista de las vulnerabilidades que pueden presentarse en este ámbito.

- Defectos intrínsecos del modelo.
- Existencia de sesgos en el modelo o en los datos.
- Compromiso del framework utilizado para el modelo del sistema de IA. Se entiende este como el conjunto de librerías, elementos o artefactos software definidos en conjunto y mantenidos por un tercero, destinados a la implementación de sistemas de IA.
- Existencia de puertas traseras en el modelo.
- Exposición y acceso al código del modelo.
- Vulnerabilidades desconocidas en relación con potenciales defectos del modelo.

Las vulnerabilidades específicas y asociadas a los ataques de adversarios son propias del modelo, pero dada su naturaleza, se han tratado con detalle en los apartados anteriores.

En la siguiente tabla podemos ver los controles de seguridad aplicables a las anteriores vulnerabilidades.

Vulnerabilidades	Controles de seguridad
Defectos intrínsecos del modelo	Cuando se seleccione un modelo para el sistema de IA, el proveedor debe conocer e investigar sus defectos intrínsecos tanto aquellos específicos dentro de la familia a la que pertenece el modelo (CNN, Randomforest, KNN, etc.) como aquellos aplicables a la finalidad prevista.

Vulnerabilidades	Controles de seguridad
	Estos defectos deberán estar inventariados en el modelo de amenaza y documentada la medida aplicada para mitigar su explotación como vulnerabilidades. La información recogida dependerá del modelo utilizado y sus defectos conocidos.
Existencia de sesgos en el modelo, o en los datos	Si bien la existencia de sesgos, puede ser un defecto explotable por el atacante, no es el objetivo de esta guía, y se describen en otras guías con más detalle (guía de datos y precisión).
Compromiso del framework del modelo	Los modelos se encuentran dentro de frameworks (marcos de trabajo) utilizados para el desarrollo y la implementación de sistemas de IA. Para mitigar el riesgo de compromiso, es fundamental que los frameworks utilizados estén actualizados y que se conozcan y monitoricen las vulnerabilidades publicadas, como las registradas en CVE (Common Vulnerabilities and Exposures) o CWE (Common Weakness Enumeration). Además, se recomienda integrar procesos automatizados de revisión y análisis de vulnerabilidades en el ciclo de desarrollo del sistema, utilizando herramientas especializadas como: • OWASP Dependency-check: herramienta open-source que permite detectar vulnerabilidades conocidas en dependencias y librerías utilizadas en el sistema. • Dependency-Track: solución que facilita el análisis continuo de componentes y la gestión del riesgo asociado a dependencias de terceros. Estas herramientas permiten identificar y alertar de manera proactiva sobre dependencias vulnerables, automatizando la monitorización y facilitando la actualización o sustitución de componentes inseguros. De esta forma, se refuerza la seguridad del framework y se reduce significativamente el riesgo de explotación.
Existencia de puertas traseras en el modelo	Para mitigar el riesgo de puertas traseras en los modelos de IA, se deben realizar pruebas periódicas y análisis de seguridad utilizando técnicas específicas: • Análisis periódicos de código tipo **SAST** (Static Application Security Testing). Para **identificar** en las **librerías** y herramientas utilizadas vulnerabilidades que pudieran ser un vector de ataque para cualquiera de las

Vulnerabilidades	Controles de seguridad
	vulnerabilidades descritas en esta guía. Además, permitirá **identificar** un **modelo** vulnerable o un **framework** de IA desactualizado o comprometido. • Análisis periódicos de código tipo **DAST** (Dynamic Application Security Testing). El enfoque de un análisis tipo DAST centrado en inteligencia artificial, debe tener especial consideración los escenarios de acceso al modelo, control de los datos de entrada y el riesgo de escalado de privilegios en las políticas de seguridad que permitan la manipulación o la intervención en los datos de entrada. • Realización de **prueba de penetración** orientado a Inteligencia Artificial y Machine Learning. En este caso, las pruebas deberán estar orientadas a realizar ataques de tipo adversario: evasión, inversión y extracción. Para facilitar la ejecución de estas pruebas, se recomienda el uso de frameworks específicos para seguridad en IA, como: o MLSecTools: herramienta diseñada para realizar pruebas de penetración y evaluar la robustez de modelos de Machine Learning frente a ataques adversarios. o Adversarial Robustness Toolbox (ART): framework open-source de IBM que proporciona un conjunto de herramientas para realizar pruebas de robustez y ataques adversaries en modelos de IA. La integración de estas herramientas en el proceso de desarrollo y evaluación garantiza una detección más eficaz de vulnerabilidades relacionadas con puertas traseras y otros vectores de ataque, mejorando la resiliencia del modelo frente a manipulaciones maliciosas.
Exposición y acceso al código del modelo	La protección del código del modelo con políticas de acceso es fundamental para evitar manipulaciones errónea o malintencionadas, y las medidas de seguridad aplicables en este ámbito son las mismas que deben implementarse en cualquier sistema informático. Estos riesgos no son exclusivos de sistemas

Vulnerabilidades	Controles de seguridad
	de IA, pero adquieren especial relevancia dada la sensibilidad y complejidad de los modelos utilizados. Las políticas de acceso deben restringir el acceso al código únicamente al personal autorizado y cualificado. Esto incluye la implementación de controles de identidad y autenticación robustos, uso de MFA, certificados, etc. (autenticación multifactor), y el seguimiento de todas las interacciones mediante registro detallados (logs). Los equipos del personal que participa en el desarrollo del sistema de IA deben estar adecuadamente protegidos frente a exfiltraciones, por ejemplo, no permitiendo utilizar los medios extraíbles (USB, disco externo…), o acceso a redes públicas. En aquellos escenarios de trabajo con el repositorio central, el canal de comunicación securizado y cifrado (como VPN, etc.). El tránsito se tiene que realizar cifrado, para evitar la posibilidad de una manipulación del modelo en tránsito a través de ataques Man-In-The-Middle. Estas medidas aseguran la integridad del código y deben aplicarse de manera coherente tanto a sistemas tradicionales como a aquellos que implementan inteligencia artificial, garantizando la seguridad del entorno completo en el que opera el sistema.
Vulnerabilidades desconocidas en relación con potenciales defectos del modelo	Análisis continuado de los indicadores de ciberseguridad del cuadro de mando para el seguimiento de la operación del sistema de IA. Análisis de las auditorías del sistema de IA realizadas.

Tabla 3 Vulnerabilidades y controles de seguridad para defectos intrínsecos

Ejemplo – Asistencia al trabajo

En el análisis de riesgos realizado se ha identificado que el riesgo de poder falsear una identificación para el sistema de asistencia al trabajo, haciendo así sería posible que se generasen registros de asistencia que realmente no se han producido. Las vulnerabilidades de **clasificación errónea y clasificación errónea aleatoria** se han considerado defectos intrínsecos al modelo y relacionadas directamente con este riesgo.

Un atacante podría utilizar patrones adversarios, por ejemplo, en una camiseta, que pudiera causar una identificación errónea tanto de otro usuario como no localizar al usuario en su base de datos (ver por ejemplo al respecto https://arxiv.org/pdf/2211.07383.pdf).

Como control de seguridad al desarrollar el modelo, para mitigar esta vulnerabilidad, el equipo de diseño utiliza el framework Highly Confident Near Neighbor (HCNN) en combinación con un Attribution-driven Causal Analysis.

Además, el framework del modelo utilizado y todas las librerías de ML asociadas siguen escaneos periódicos tipo **SAST** para localizar vulnerabilidades en el *perímetro del sistema de IA* que sirviesen de puerta de entrada.
En el entorno productivo, se somete de manera periódica a un análisis de tipo **DAST** centrado en una escalada de privilegios y acceso sobre el modelo. En este entorno y de manera periódica también, se realizan test de **penetración** orientados a IA, realizando ataques adversarios al sistema.

Para los defectos que pueden aparecer en el sistema de inteligencia artificial, como resultado de la configuración, integración o implementación se pueden identificar las siguientes vulnerabilidades:

- El sistema de inteligencia artificial permite extraer información privada.
- Demasiada información proporcionada en el modelo en la información de salida.
- Demasiada información disponible de forma pública sobre el modelo.
- No se han considerado los ataques a los que el sistema de inteligencia de IA puede estar expuesto.

- No existe un proceso de seguridad que mantenga el nivel de seguridad de los componentes del sistema de IA.
- Uso de componentes vulnerables.
- Desajuste entre los entornos de desarrollo/pruebas/producción.
- No se han definido indicadores del correcto funcionamiento del sistema, haciendo complejo identificar el compromiso del sistema.
- Malas prácticas por falta de concienciación en ciberseguridad.
- Contratos con terceras partes que no tienen un nivel de seguridad adecuado.
- Vulnerabilidades desconocidas como resultado de la configuración, integración o implementación.

Vulnerabilidades	Controles de seguridad
Gestión de derechos de acceso inadecuada	Unos accesos inadecuados o no controlados pueden desembocar en la introducción de defectos por personal no cualificado.
El sistema de inteligencia artificial permite extraer información privada	El sistema de IA dispone de un mecanismo de privacidad diferencial; el diseño del modelo se ha realizado utilizando PATE (Private Aggregation of Teacher Ensembles). Las técnicas utilizadas para proteger frente ataques de inversión son aplicables en la misma medida como controles de seguridad para esta vulnerabilidad.
Uso de componentes vulnerables	En este caso ISO27001/2 y NIST 800-53 recomiendan: • Disponer de un inventario de los componentes utilizados y realizar un seguimiento constante de los mismos y sus versiones para estar al tanto de las vulnerabilidades encontradas en ellos utilizando las bases de datos de referencia para ellas (de tipo CVE, https://cve.mitre.org/ o de tipo CWE, https://cwe.mitre.org/). Se pueden utilizar herramientas automáticas, como OWASP Dependency-check, para comprobar los componentes utilizados y realizar las comprobaciones en las bases de datos de referencia.

Vulnerabilidades	Controles de seguridad
	• Gestionar aquellos componentes obsoletos y planificar su eliminación de la cadena de producción. **Sin ser de aplicación específica para IA,** realizar análisis frecuentes de vulnerabilidades en todos los sistemas que rodean al sistema de IA.
No se han definido indicadores del correcto funcionamiento del sistema, haciendo complejo identificar el compromiso del sistema.	El sistema de IA debe disponer de una serie de indicadores definidos y asociados a su naturaleza y finalidad prevista. Estos indicadores deben presentarse en la interfaz del sistema de IA, mediante un dahsboard o panel de información accesible, donde el responsable del despliegue pueda visualizar y conocer en tiempo real el estado de desempeño del sistema y detectar la posible existencia de fallos o posible situación de compromiso. Para cumplir su propósito, estos indicadores deben incluir umbrales claramente definidos que permitan alertar sobre anomalías en el comportamiento del sistema. Un umbral fuera de rango puede ser indicativo de un fallo, un intento de ataque o una degradación del rendimiento, facilitando la respuesta temprana y mitigación de riesgo. Los indicadores deben referirse, al menos, a las siguientes métricas: • Métricas de precisión y solidez del sistema de IA, para evaluar su desempeño en condiciones esperadas y adversas. • Indicadores distintivos de las bases de datos de entrenamiento utilizadas, tales como la distribución de los datos o la detección de sesgos, para orientar el correcto uso del sistema y alertar de posibles desviaciones. La combinación de indicadores con umbrales definidos garantiza que el sistema pueda monitorearse proactivamente, facilitando la identificación de problemas en tiempo real y mejorando la resiliencia frente a situaciones inesperadas.

Vulnerabilidades	Controles de seguridad
Desajuste entre los entornos de desarrollo / pruebas / producción	Los sistemas entrenados en entornos de simulación pueden cometer errores en el mundo real debido a desajustes entre entornos de entrenamiento y de producción, o los entornos de simulación y producción. Ambos entornos deben estar definidos por descriptores de contenedores cuando sea posible o dispositivos hardware idénticos en los mismos entornos (IoT, edge computing). En los escenarios de comercialización distintos al MLSaaS, las instrucciones deben ser claras (incluso con ejemplos de los contenedores software si es necesario) y las especificaciones hardware estarán basadas en estándares, para el caso de IoT o en los que se realice edge-computing.
Contratos con terceras partes que no tienen un nivel de seguridad adecuado	En ocasiones, una tercera parte puede desarrollar un elemento o sección del sistema de IA. Si el proveedor delega en una tercera parte la implementación completa del sistema, deberán tenerse en cuenta y cumplirse todas las medidas de seguridad descritas en esta guía. Para el caso de un desarrollo parcial o específico del sistema de IA (como modelos, preprocesado de datos o infraestructura de soporte), es fundamental integrar requisitos de seguridad en los contratos con terceros. Las recomendaciones de ISO 27001/2 o NIST 800-53 incluyen integrar cláusulas específicas que establezcan controles de seguridad claros y verificables. Ejemplos de estas cláusulas son: • Requisitos de auditorías periódicas: Obligar a la realización de auditorías de seguridad del desarrollo y los procesos del proveedor, con acceso a los resultados para el cliente. • Certificaciones de seguridad: Exigir que la tercera parte acredite el cumplimiento de estándares como ISO 27001, ISO 27017 (para seguridad en la nube) o certificaciones equivalentes. • Revisión y gestión de vulnerabilidades: Incorporar la obligación de realizar pruebas de seguridad periódicas (SAST/DAST), con la remediación de vulnerabilidades críticas en plazos específicos.

Vulnerabilidades	Controles de seguridad
	• Confidencialidad y protección de la información: Incluir políticas estrictas de protección de datos y propiedad intelectual en el desarrollo del sistema. • Planes de contingencia y respuesta a incidentes: Establecer la obligación de contar con un plan de acción en caso de ciberincidentes que puedan afectar al sistema de IA. Además, se recomienda monitorizar y revisar continuamente el cumplimiento de estos requisitos mediante indicadores de desempeño y revisiones periódicas de los servicios proporcionados.
Vulnerabilidades desconocidas como resultado de la configuración, integración o implementación.	Análisis continuado de los indicadores de ciberseguridad del cuadro de mando para el seguimiento de la operación del sistema de IA. Análisis de las auditorías del sistema de IA realizadas. Se recomienda implementar herramientas de monitoreo en tiempo real basadas en IA que detecten patrones anómalos en el comportamiento del sistema. Adicionalmente, las auditorías periódicas deben incluir simulaciones de ataques adversarios para descubrir posibles defectos no documentados

Tabla 4 Vulnerabilidades y controles de seguridad para defectos de integración/configuración

Ejemplo – Sistema automático de concesión de ayudas

Este sistema se despliega en las instalaciones propias de la administración pública que hace uso de este, especialmente en las Comunidades Autónomas y Entidades Locales. En ese proceso se ha identificado que puede existir un **desajuste entre los entornos de desarrollo, pruebas y producción**. Estos desajustes incluyen diferencias en las configuraciones del hardware, las versiones de las bibliotecas o frameworks utilizados, y las dependencias específicas del sistema.

Para mitigar este riesgo, y facilitar el trabajo del equipo de administración de sistemas y mantenimiento en la infraestructura de las CCAA y de las Entidades Locales, se encapsula el sistema de IA como un conjunto de servicios (entradas de datos, modelo, sistema de almacenamiento, etc.) a través de un sistema de mercado basado en descripción de contenedores (por ejemplo, Docker), de tal manera que, con el mismo mecanismo de descripción, se puede replicar el entorno en cualquier infraestructura, con las mismas condiciones y artefactos software.

Ejemplo - Asistencia al trabajo

En la construcción del sistema de IA se ha incluido una interfaz de usuario que permite visualizar el comportamiento del sistema de forma agregada y desglosada por secciones y departamentos. Este panel utiliza métricas avanzadas como Highly Confident Near Neighbor (HCNN) en combinación con un Attribution-driven Causal Analysis para identificar comportamientos anómalos relacionados con los mecanismos de control de adversarios.

En este panel se visualiza una información de histórico de dicha relación con datos de los últimos días, semana y mes. Se muestra gráficamente el progreso de dichos valores de métricas. En este panel, se puede filtrar por secciones de las instalaciones y por departamentos ofreciendo los mismos datos y visualización que el agregado.

Así el sistema tiene un panel donde se puede detectar el **comportamiento anómalo general, por demarcación e incluso por zona**. El panel dispone de un mecanismo de configuración para permitir definir niveles de alarma y notificación. Esta solución facilita la detección y respuesta rápida ante anomalías, ayudando a mantener la robustez del sistema según lo requerido por el Artículo 15 del Reglamento Europeo de IA.

Estas vulnerabilidades, especialmente las correspondientes a la integración y configuración del sistema, están en la **frontera** entre ciberseguridad para sistemas de inteligencia artificial de alto riesgo y la ciberseguridad del ecosistema tecnológico que le acompaña. Se ha considerado su inclusión, para disponer de un perímetro adecuado que permita una continuidad en la ciberseguridad (más allá del ámbito de esta guía). Así mismo, se han proporcionado referencias que permitan a la organización trabajar tener una perspectiva más completa en ciberseguridad de manera esta sea continúa e integral.

Si el modelo va a utilizarse por el usuario on-premise (en la nube o en sus instalaciones), las **instrucciones de instalación y configuración han de ser guiadas**, claras y paso a paso, indicando la funcionalidad y relación de cada parámetro y los errores de configuración. El proceso de cambio de parámetros deberá preguntar por la confirmación de estos y esperar a la confirmación del usuario previamente identificado (por los mecanismos disponibles para la naturaleza del sistema).

Responsable del despliegue

En este apartado las medidas de organizativas para el responsable del despliegue implican una lectura y compresión del manual elaborado por el proveedor, en lo relativo tanto a los defectos intrínsecos del sistema como a los mecanismos de configuración que le sean de aplicación de IA dentro del alcance de la finalidad prevista para este. Deberá conocer la documentación relativa a los indicadores del correcto funcionamiento del sistema.

Si el formato de uso y comercialización del sistema de IA contemplase que el responsable del despliegue dispusiese del sistema de IA como **un activo de sus propias instalaciones** (ya sea on-premise, o en cualquier infraestructura o instalación gestionada por el responsable del despliegue) deberá aplicar las indicaciones proporcionadas por el proveedor en materia de ciberseguridad para AI que aparezcan en el manual.

El responsable del despliegue deberá disponer de los medios técnicos que le sean de aplicación para que los mecanismos (alertas, notificaciones) o las interfaces (web, desktop, terminales) de las que disponga que sean accesibles para el personal que deba interactuar con el sistema de IA de alto riesgo.

5. Documentación técnica

Esta guía de ciberseguridad se acompaña de una específica de documentación técnica que establece cómo debe documentarse un sistema de IA de alto riesgo, en relación con las indicaciones proporcionadas por el Reglamento Europeo de IA. Siendo la citada guía el documento de referencia para la cumplimentación de la documentación técnica, en este apartado vamos a indicar de manera agregada aquellos puntos que se deben tener en cuenta para elaborar la documentación técnica sobre la ciberseguridad para IA del sistema de alto riesgo, y que han sido mencionados a lo largo de esta guía. Para una referencia completa, consultar la guía de documentación técnica.

En el Anexo IV del Reglamento de Inteligencia Artificial se establece el contenido mínimo que debe incluir la documentación técnica. En el ámbito de la ciberseguridad, el punto 1.h hace referencia a las instrucciones de uso, en las cuales deben detallarse los principales riesgos y controles de ciberseguridad del sistema, con el fin de informar al responsable de su despliegue. Asimismo, el punto 2.h dispone que deberán documentarse las medidas de ciberseguridad aplicadas.

A continuación, indicamos una referencia inicial de cómo abordar estos requisitos.

Manual de instrucciones

Las medidas de ciberseguridad puestas en marcha para este sistema de IA, que puedan ser de aplicación al responsable del despliegue, especialmente aquellas que estén implicadas en los procesos de instalación del sistema en el caso en el que se realicen, para evitar la aparición de ataques que exploten defectos de configuración (ver apartado 4.5).

También el manual de instrucciones debe contener información sobre los riesgos de ataques adversarios que puede sufrir el sistema, así como los controles de seguridad puestos en marcha para mitigar su efecto (ver apartado 4.4). De igual modo, como se ha explicado en los ejemplos de caso de uso, las interfaces presentes en el sistema, que muestren indicadores de correcto funcionamiento, deben estar detallados en las instrucciones, con claras indicaciones de su interpretación fácilmente entendibles por personal no especializado.

Parámetros utilizados para medir la ciberseguridad

La documentación técnica del sistema debe contener los parámetros utilizados para medir, monitorizar y actualizar la ciberseguridad del sistema. Estos parámetros deben estar alineados con las estrategias de mitigación de riesgos y vinculados a las vulnerabilidades identificadas. Al menos se debe documentar:

- El plan de ciberseguridad para IA establecido, incluyendo su alance y objetivos.
- Los activos y actores identificados en el análisis de ciberseguridad.

- Las vulnerabilidades identificadas y los controles aplicados sobre los datos para los ataques de envenenamiento.
- Las vulnerabilidades identificadas y los controles aplicados para la protección frente a ataques adversarios.

Las vulnerabilidades y controles asociados a los defectos del modelo, incluyendo las medidas de prevención y detección implementadas.

Además, esta documentación debe cumplir con los procedimientos e indicaciones proporcionados en esta guía, especificando las técnicas utilizadas para evaluar y mitigar riesgos, su relación directa con los riesgos mitigados y las vulnerabilidades abordadas y un enfoque claro para su mantenimiento y actualización en el ciclo de vida del sistema.

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: Políticas de acceso. Una aproximación a sistemas de inteligencia artificial de alto riesgo

Podemos decir que la primera y autentica línea de defensa respecto a muchas de las medidas específicas de IA, se basa en controlar de manera adecuada el acceso a todos los activos durante el ciclo de vida.

Las políticas de acceso, aun no tratándose de un tema específico de inteligencia artificial, se encuentran lo suficientemente cerca de todos los activos y procesos a lo largo del ciclo de vida para dedicarle un enfoque que permita trazar a la organización una continuidad que asegure la protección.

7.1.1 Para el proveedor

Las políticas de acceso, entendidas como las reglas de acceso a los activos, así como las reglas de acción sobre estos son muy importantes a la hora de aplicar medidas específicas de ciberseguridad a los sistemas de IA de alto riesgo. Estas políticas establecen **qué** puede hacerse y **quién** puede hacerlo sobre los activos del sistema de IA y procesos del ciclo de vida de este. Los activos protegidos son cruciales en la ciberseguridad del sistema. Dejarlos expuestos es un riesgo inasumible para cualquier sistema de IA.

Se recomienda que las políticas de acceso se sustenten sobre el paradigma RBAC (Sistema de acceso basado en roles), que permita asignar roles a los responsables del despliegue y acciones permitidas a dichos roles.

El enfoque de las políticas de acceso debe ser de **mínimos privilegios** y se debe establecer un mecanismo de autentificación robusto, rotación de contraseñas y/o acceso mediante certificados en función del sistema. Se recomienda la inclusión de un sistema de **factor múltiple de autenticación**, que evite la suplantación, por un compromiso de una contraseña. Las políticas de **revocación de accesos** deben formar parte de los mecanismos de salida de la compañía para evitar la existencia de accesos para personas que ya no están en la organización.

Se recomienda que sea compatible con soluciones de gestión de identidad de mercado, extendidas en las organizaciones, como puede ser Active Directory o equivalente. Dicha integración estará **documentada** en el **manual de instrucciones** para facilitar el proceso a los usuarios del sistema.

¿A quién aplica dentro de la organización?

Las políticas de acceso al sistema de IA aplican a todo el personal que tenga interacción con el sistema a cualquier nivel desde su desarrollo, implementación y comercialización. Es muy importante trasladar a cada uno de los grupos de actores indicados los mínimos privilegios posibles.

Estos grupos podrán ser:

- Actores sobre datos: perfiles que actúen sobre los datos que alimentan el aprendizaje del sistema de IA (Analistas de datos, Científicos de datos, Expertos en IA, etc.)
- Actores sobre la configuración y operación del sistema de IA (Expertos en AI, Administradores del sistema de IA, Analistas de sistemas, Expertos en ML, etc.)
- Desarrolladores de integración y/o interacción con el sistema de IA (Incluido MLOps, en su caso).
- Actores de administración de sistemas que hospedan y operan el sistema de IA (Administradores del sistema de IA, Analistas de sistemas, etc.)
- Actores sobre el diseño, implementación y mantenimiento del software que materializa el sistema de IA en sí mismo (Incluido en su caso el uso de firmware o similares) (Desarrolladores, Analistas, Expertos en IA, etc.).
- Actores sobre la arquitectura, administración y mantenimiento del sistema de IA en su ciclo de vida, tanto durante el desarrollo como una vez puesto en marcha (Expertos en AI, Administradores del sistema de IA, Expertos en ML, etc.).
- El personal general de la organización aplicando la política de mínimos privilegios, no deberá tener acceso al sistema de IA, al desarrollo del modelo o a los datos. En aquellos casos fuera de los actores identificados en los puntos anteriores, el acceso debe analizarse y registrar la necesidad, si esta es temporal deberá revocarse adecuadamente.

Véase también ISO/IEC 22989:2022 Information technology — Artificial Intelligence — Artificial intelligence concepts and terminology

¿Qué aspectos debe cubrir?

Las políticas de acceso establecidas sobre cualquiera de los activos descritos en este apartado deben revisarse periódicamente estando actualizados los siguientes puntos:

- Personas y roles asignados, actualización y cambios.
- Roles definidos y acciones.
- Proceso de revocación y eliminación de permisos.
- Permisos sin asignación a un usuario.
- Sistema de registro de accesos y monitorización de uso.

Las acciones sobre **datos de entrenamiento** deben estar protegido por las políticas de seguridad, considerando al menos:

- Incorporación de nuevos datos.

- Eliminación de datos existentes.
- Edición de datos existentes.
- Utilización de los datos.

La segmentación de las acciones sobre datos es fundamental para tener un control y protección sobre los riesgos del **envenenamiento de datos**.

El **modelo** como parte fundamental del sistema de inteligencia artificial debe ser protegido por las políticas de seguridad. Se deben controlar al menos, las siguientes acciones:

- Acceso (visualización) al código del modelo.
- Edición del código del modelo.
- Utilización de modelo para entrenamiento.
- Proceso de entrenamiento.
- Validación y despliegue (liberación) del modelo a producción.
- Acceso, edición y supresión de cualquier documentación técnica y de diseño relativa al modelo y descriptiva de la naturaleza del sistema y sus componentes.

La segmentación de las acciones sobre el propio modelo permite proteger a este de manipulaciones inadvertidas del mismo, tanto intencionadas realizadas por atacantes malintencionados, como erróneas por personas no cualificadas (sin el permiso establecido) introduciendo errores en el modelo. Evita la exfiltración de este para ataques de transferencia y controla el proceso de entrenamiento para la protección contra **ataques adversarios.**

7.1.2 Para el responsable del despliegue

Los responsables del despliegue también deben tener establecida y definida unas políticas de acceso al sistema de IA. Los accesos a los sistemas de Inteligencia Artificial de alto riesgo por parte del personal del responsable del despliegue jamás deberán ser anónimos y sin acceso controlado.

Adicionalmente, y de acuerdo con el modelo de comercialización del sistema de inteligencia artificial, serán necesarias establecer otras políticas de acceso.

El enfoque de las políticas de acceso debe ser de mínimos privilegios y se debe establecer un mecanismo de rotación de contraseñas robusto.

¿A quién aplica dentro de la organización?

El personal de la organización usuaria que tenga acceso al sistema y a su interacción. Si debido al modelo comercialización, es esta organización la que administra el sistema (ya sea in-cloud, on-premise o cualquier variante) deberá incluirse en las políticas de acceso todo el personal relacionado.

Los accesos, su actividad y las personas autorizadas deberán estar registrados y auditados periódicamente, especialmente las revocaciones de permisos, que deberán surtir efecto inmediato.

El sistema de IA deberá integrarse, en aquellos escenarios de comercialización, donde sea necesario, con el mecanismo de identidad de la organización usuaria para el acceso al sistema de IA, siguiendo las indicaciones y recomendaciones del **manual de instrucciones** proporcionado por el proveedor.

¿Qué aspectos debe cubrir?

El principal activo protegido por políticas de seguridad para el caso del responsable del despliegue es **el propio sistema de IA**. El acceso al sistema, a su panel de control, interfaz o equivalente, deberá estar protegido siguiendo las instrucciones del proveedor, e integrado con los mecanismos de identificación propios de la organización del responsable del despliegue en aquellos casos en los que aplique.

Es **responsabilidad del responsable del despliegue** gestionar este acceso y disponer de políticas adecuadas.

7.2 Anexo II: Formación en ciberseguridad y sistemas de inteligencia artificial de alto riesgo

Como se ha comentado en la introducción, la formación permite fortalecer la cadena de seguridad que protege el sistema de inteligencia artificial de alto riesgo. De nuevo este anexo cubre un aspecto que no es central para la ciberseguridad de la AI, pero que se encuentra en una región de frontera y que es **necesario** cubrir.

7.2.1 Para el proveedor

La ciberseguridad para IA aplica al proveedor durante todo el ciclo de vida del sistema de IA y a lo largo de todos los activos (datos, modelo etc.) de dicho sistema. Para que las políticas de ciberseguridad sean efectivas, estas deben tener una comunicación a todo el personal implicado. La formación en ciberseguridad para sistemas de IA tiene un objetivo doble dependiendo del alcance de la ciberseguridad en la organización del proveedor:

- Si el proveedor **ya cuenta** con un proceso de ciberseguridad, este tendrá como uno de sus pilares la formación. Esta formación **debe verse ampliada** para cubrir los tópicos aquí descritos y alcanzar a las personas consideradas.
- Si el proveedor **no cuenta** con un proceso ciberseguridad específicamente diseñado, **deberá organizar** unos planes de formación que cubran los aspectos descritos en este apartado.

La formación proveedor- responsable del despliegue debe estar claramente definida. En aquellos modelos de comercialización y/o puesta en servicio del sistema de inteligencia

artificial que requieran participación del personal del responsable del despliegue, más allá de la propia utilización del sistema, los aspectos básicos a cubrir por el plan de formación en ciberseguridad deben estar descritos en el manual de instrucciones.

¿A quién aplica dentro de la organización?

Los receptores de la formación en ciberseguridad podrán ser:

- Actores sobre datos: perfiles que actúen sobre los datos que alimentan el aprendizaje del sistema de IA (Analistas de datos, Científicos de datos, Expertos en IA, etc.)
- Actores sobre la configuración y operación del sistema de IA (Expertos en AI, Administradores del sistema de IA, Analistas de sistemas, Expertos en ML, etc.)
- Desarrolladores de integración y/o interacción con el sistema de IA (Incluido MLOps, en su caso).
- Actores de administración de sistemas que hospedan y operan el sistema de IA (Administradores del sistema de IA, Analistas de sistemas, etc.)
- Actores sobre el diseño, implementación y mantenimiento del software que materializa el sistema de IA en sí mismo (Incluido en su caso el uso de firmware o similares) (Desarrolladores, Analistas, Expertos en IA, etc.).
- Actores sobre la arquitectura, administración y mantenimiento del sistema de IA en su ciclo de vida, tanto durante el desarrollo como una vez puesto en marcha (Expertos en AI, Administradores del sistema de IA, Expertos en ML, etc.).
- El personal general de la organización aplicando la política de mínimos privilegios, no deberá tener acceso al sistema de IA, al desarrollo del modelo o a los datos. En aquellos casos fuera de los actores identificados en los puntos anteriores, el acceso debe analizarse y registrar la necesidad, si esta es temporal deberá revocarse adecuadamente.

Véase también ISO/IEC 22989:2022 Information technology — Artificial intelligence — Artificial intelligence concepts and terminology

¿Qué aspectos debe cubrir?

Los aspectos aquí descritos deben tratarse en diferente profundidad, según el personal receptor de los mismos y su vinculación con el sistema de inteligencia artificial. Este alcance podrá ir desde una sensibilización para todo el personal a una mayor complejidad para aquellos directamente relacionados con el sistema de IA.

Esta formación debe cubrir, al menos, los siguientes tópicos de ciberseguridad aplicada a sistemas de inteligencia artificial:

- ¿Por qué es importante la ciberseguridad en IA? Ciberseguridad durante el ciclo de vida del sistema de IA.
- La ciberseguridad de los datos de entrenamiento: riesgos de envenenamiento de los datos de entrenamiento.

- Protección del sistema de IA. Riesgos de filtración de información y código del sistema de IA o manipulación.
- ¿Qué son los ataques adversarios? Definición. Diferentes tipos. Ataques adversarios para el sistema de Inteligencia Artificial en consideración.
- Explotación de defectos del modelo. Definición de ataques por transferencia. Necesidad de utilización de modelos/frameworks seguros. ¿Qué son los defectos intrínsecos del modelo? Configuración e integración segura. Evitar errores en el modelo por el uso de malas prácticas.
- La importancia del control de acceso basado en roles (RBAC) para acceso tanto a los datos como al modelo y robustez del sistema de autenticación, incluyendo refuerzos como autenticación de doble factor.

La formación deberá actualizarse en relación con los cambios del modelo y el paradigma de ciberseguridad en inteligencia artificial, que es un área cambiante y en continua evolución.

Podrá requerir que la recepción de la formación por determinados perfiles (científicos de datos, expertos en ML/AI) requiera la realización de pruebas y/o certificaciones.

Para el responsable del despliegue del sistema de IA se debe desarrollar una **indicación del material formativo,** explicando el nivel de ciberseguridad requerido para la utilización del sistema de Inteligencia artificial: FAQs, documentación explicativa, descripciones paso a paso de la operación y puntos críticos. Todo ello adaptado al nivel de comercialización del sistema de IA (SaaS, On-premise, in-cloud etc.)

7.2.2 Para el responsable del despliegue

La formación en ciberseguridad de aplicación al responsable del despliegue dependerá del enfoque que tenga el modelo de comercialización del sistema de inteligencia artificial considerado. En aquellos escenarios en los que el sistema se instale en infraestructura del responsable del despliegue (modelos on-premise, in-house) el responsable del despliegue deberá de adecuar sus planes de formación en materia de ciberseguridad para cubrir los aspectos indicados en el manual de instrucciones del sistema.

¿A quién aplica dentro de la organización?

El personal de la organización usuaria del sistema de inteligencia artificial, que vaya a realizar la interacción directa con este. Dependiendo del enfoque de comercialización:

- Personal de administración de sistemas.
- Personal responsable de la configuración del sistema de IA.
- Personal responsable de la supervisión del sistema de IA.

¿Qué aspectos debe cubrir?

El responsable del despliegue deberá utilizar el material generado por el proveedor en materia de ciberseguridad aplicada a su sistema de IA para orientar y cubrir los conocimientos y/o perfiles necesarios.

Independientemente del formato de comercialización en que reciba el sistema, todo su personal en interacción directa con el sistema de IA deberá de recibir formación en materia de sensibilización, vulnerabilidades de los sistemas de IA, envenenamiento, ataques adversario…

El personal que interactúa con el sistema de IA y que es encargado de trabajar con este debe recibir una formación específica en ataques adversarios para el sistema de IA y su finalidad prevista.

Acorde a los formatos de comercialización, si la configuración o la administración del sistema de IA se realiza por personal del responsable del despliegue (administración de sistemas propia, por ejemplo) la formación deberá incluir:

- Explotación de defectos del modelo. Definición de ataques por transferencia. ¿Qué son los defectos intrínsecos del modelo? Configuración e integración segura.

La formación deberá ser periódica y actualizarse acorde a los cambios y actualizaciones que proporcione el proveedor.

7.3 Glosario

Término	Definición
Ataque Adversario	Los ataques que se realizan contra un sistema de inteligencia artificial, se consideran ataques adversarios. En relación con el nivel de conocimiento que el atacante los ataques adversarios puede ser: - Caja Blanca: Acceso la arquitectura del modelo, datos de entrenamiento, parámetros e hiper parámetros. - Caja Negra: entradas y salidas. - Caja gris: una mezcla de ambas. En relación con el modo de operación los ataques pueden ser (consultar en detalle las definiciones en el glosario): - Envenenamiento. - Evasión, que en ocasiones se consideran directamente de manera genérica los ataques adversarios, aunque realmente se trata de un tipo específico de los mismos. - Extracción. - Inversión.

Término	Definición
Ataque Evasión	Un tipo de ataque que se producen durante la fase de inferencia con el sistema de IA en producción. El atacante no tiene acceso a los datos de entrenamiento, solo se comunica con el sistema vía interfaz (física, visual o tipo API). En general este tipo de ataques busca que una predicción del modelo sea errónea.
Ataque Extracción	Tipo de ataque durante la fase de inferencia con el sistema de IA en producción. Sin acceso a los datos de entrenamiento, solo se comunica con el sistema vía interfaz (física, visual o tipo API). El atacante mediante el acceso al modelo proporciona inputs a éste, y registras los outputs, para inferir una replica total o parcial del modelo a través de sus respuestas.
Ataque Inversión	Es un tipo de ataque que se produce durante la fase inferencia con el sistema de IA en producción. En estos ataques el usuario no tiene acceso directo a los datos, pero si puede tener acceso al modelo a través de su interfaz (ya sea esta física, visual o tipo API). Los ataques de inversión como objetivo la obtención de conocimiento de los datos del modelo. Como se ha indicado en la guía pueden ser: - Membership Inference Attack para conocer si una muestra pertenece a los datos de entrenamiento. - Property Inference Attack, que busca conocer los datos utilizados por el modelo durante en el entrenamiento para realizar las predicciones. - Existe un tercer tipo que es reconstrucción, que es más complicado de ejecutar, porque su intención es reconstruir todos los datos de entrada.
Attribution-Driven Causal Analysis	Este concepto se relaciona con que existe una conexión entre la resistencia a las perturbaciones adversariales y la explicación basada en la atribución de las decisiones individuales generadas por los modelos de aprendizaje automático. Las entradas adversariales no son robustas en el espacio de atribución, es decir, el enmascaramiento de unos pocos rasgos con alta atribución lleva a cambiar la indecisión del modelo de aprendizaje automático en los ejemplos adversariales. En cambio, las entradas naturales son robustas en el espacio de atribución. El concepto ha sido desarrollado en un artículo, que puede ser encontrado aquí https://arxiv.org/abs/1903.05821
Bagging	Utilización de modelos entrenados en paralelo para constituir el sistema de inteligencia artificial. El objetivo es aprovechar la independencia entre los modelos entrenados de manera separada sobre diferentes subconjuntos del conjunto original de datos de entrenamiento (con la

Término	Definición
	misma finalidad prevista para el sistema) para promediar su salida. Este tipo de técnica es un mecanismo de protección frente a envenenamiento de datos. También es una técnica eficiente frente a los ataques de evasión, al no depender de un único modelo que pueda ser vulnerable al ataque considerado. Para clasificaciones, los diferentes modelos entrenados realizan una *votación* de la pertenencia a una clase, para aquellas operaciones que no realizan una clasificación (regresiones, estimaciones etc.) se realiza una media aritmética. En ocasiones aparece también como *Bootstrap Aggregation*.
Boosting	El refuerzo es una técnica de conjunto que busca cambiar los datos de entrenamiento y ajustar el peso de las observaciones en función de la clasificación anterior. A diferencia del enfoque bagging, el boosting implica la dependencia de modelos débiles. Los aprendices débiles tienen en cuenta los resultados del modelo débil anterior y ajustan los pesos de los puntos de datos, convirtiéndolo en un modelo fuerte. El Boosting cambia el peso asociado a una observación que fue clasificada incorrectamente tratando de aumentar el peso asociado a ella. El refuerzo tiende a disminuir el error de sesgo, pero a veces puede llevar a un sobreajuste del conjunto de datos de entrenamiento.
CAT-Gen	Al desarrollar modelos de NLP podremos realizar diferentes tareas de manera automática como traducir textos, resumirlos, etc. El problema de este tipo de modelos es que son muy poco robustos frente a pequeñas modificaciones en los datos de entrada. Sin embargo, en la utilización de esta técnica demuestra como al introducir el aprendizaje adversario, en este tipo de modelos, la robustez frente a cambios en los datos de entrada mejora notablemente Este método puede generar textos adversarios más diversos y fluidos. La utilización de estos textos adversarios permite mejorar los modelos mediante el entrenamiento adversario. El desarrollo de la técnica está detallado por los autores se puede encontrar en https://arxiv.org/abs/2010.02338
CVE	Common Vulnerabilities and Exposures (CVE) es una lista de vulnerabilidades y exposiciones de seguridad de la información divulgadas públicamente. Existe una lista accesible en https://www.cve.org/
CWE	Common Weakness Enumeration (CWE) es un sistema de categorías para las debilidades y vulnerabilidades del software. Se puede consultar en https://cwe.mitre.org/

Término	Definición
Defensive Distillation	La destilación defensiva es una técnica de entrenamiento adversarial que añade flexibilidad al proceso de clasificación de un algoritmo para que el modelo sea menos susceptible de ser explotado. En el entrenamiento de destilación, un modelo se entrena para predecir las probabilidades de salida de otro modelo que fue entrenado en un estándar de referencia anterior para enfatizar la precisión. El primer modelo se entrena con etiquetas "duras" para lograr la máxima precisión, por ejemplo, exigiendo un umbral de probabilidad del 100%. Si un atacante aprende qué características y parámetros busca el sistema, puede enviar unos datos de entrada sólo con las características relevantes, lo que genera una coincidencia falsa positiva. El primer modelo proporciona entonces etiquetas "blandas" con una probabilidad del 95%. Esta incertidumbre se utiliza para entrenar al segundo modelo para que actúe como un filtro adicional. Como ahora hay un elemento de aleatoriedad para conseguir una coincidencia perfecta, el segundo algoritmo o "destilado" es mucho más robusto y puede detectar más fácilmente los intentos de suplantación.
DLP (herramientas)	Son herramientas destinadas a evitar la pérdida de datos. Su relación con la ciberseguridad para inteligencia artificial viene dictada por la necesidad de control, estabilidad y seguimiento de los conjuntos de datos de entrenamiento. De esta manera el control que establecen estas herramientas permite añadir un control de seguridad sobre los ataques de envenenamiento de datos. Existen varias herramientas de mercado y soluciones opensource.
Envenenamiento (poisoning)	El envenenamiento es un tipo de ataque, en la fase de diseño y entrenamiento, que trata de manipular el conjunto de datos de entrenamiento, con el objetivo de controlar las predicciones del sistema de inteligencia artificial de tal manera que el modelo transforme, en la fase de inferencia una vez desplegado en producción, entradas maliciosas en resultados correctos.
Feature Denoising	Este concepto se basa en que las perturbaciones adversariales realizadas en imágenes provocan ruido en las características construidas por las redes neuronales. Se trata de una arquitectura que incremente la resistencia a ataques adversarios realizando una eliminación de ruido en las características. Esta arquitectura ha sido propuesta en un artículo que puede ser encontrado en https://arxiv.org/abs/1812.03411
Feature Squeezing	Esta técnica puede utilizarse para proteger los modelos mediante la detección de ejemplos adversos. En la reducción (*squeezing*) de características disminuye el espacio de búsqueda disponible para un adversario al fusionar muestras que corresponden a muchos vectores

Término	Definición
	de características diferentes en el espacio original en una sola muestra. Al comparar la predicción de un modelo DNN sobre la entrada original con la de la entrada comprimida, esta técnica detecta los ejemplos adversarios con gran precisión y pocos falsos positivos. Si los ejemplos originales y comprimidos producen resultados sustancialmente diferentes del modelo, es probable que la entrada sea adversa. Los autores de esta técnica han publicado sus resultados en https://evademl.org/squeezing/
GPAI (general purpose AI)	De acuerdo con Reglamento Europeo de Inteligencia Artificial, en el **Artículo 3 Definiciones (1b):** *"un sistema de IA que -independientemente de cómo se comercialice o se ponga en servicio, incluso como software de código abierto- está destinado por el proveedor a realizar funciones de aplicación general, como el reconocimiento de imágenes y del habla, la generación de audio y vídeo, la detección de patrones, la respuesta a preguntas, la traducción y otras; un sistema de IA de propósito general puede utilizarse en una pluralidad de contextos e integrarse en una pluralidad de otros sistemas de IA".* Dado que la utilización de este tipo de sistemas en el desarrollo de un sistema de inteligencia artificial, puede ser posible, cabe aclarar aquí, y añadido a la definición indicada un extracto del considerando 12c siguiente: *"[...] los proveedores de sistemas de inteligencia artificial de propósito general, independientemente de que puedan ser utilizados como sistemas de inteligencia artificial de alto riesgo como tales por otros proveedores o como componentes de sistemas de inteligencia artificial de alto riesgo, deben cooperar, según proceda, con los proveedores finales de los respectivos sistemas de inteligencia artificial de alto riesgo para permitir su cumplimiento de las obligaciones pertinentes en virtud del presente Reglamento y con las autoridades competentes establecidas en virtud del mismo.".*
Highly Confident Near Neighbor (HCNN)	Se trata de un framework de trabajo en redes neuronales que combina la información de confianza y la búsqueda del vecino más cercano, para reforzar la robustez adversarial de un modelo base. Esto puede ayudar a distinguir entre las predicciones correctas y erróneas del modelo en una vecindad de un punto muestreado de la distribución de entrenamiento utilizada. La técnica está desarrollada en un artículo al que se puede acceder en el siguiente enlace https://arxiv.org/abs/1711.08001
Input gradient	Consiste en imponer restricciones al gradiente de entrada, o regularización. La técnica de regularización en general consiste en

Término	Definición
regularisation	restringir las estimaciones de los coeficientes a cero. Esta técnica permite evitar un *overfitting* sobre el conjunto de entrenamiento, y fundamentalmente es una técnica utilizada para proporcionar resistencia a los ataques adversarios, o a la mitigación de estos. La regularización en el gradiente de entrada evita que este sea demasiado grande, lo que hace que las redes neuronales sean vulnerables a los ataques adversarios.
Isolation Forest	Este método, se basa en el algoritmo del árbol de decisión. Aísla los valores atípicos seleccionando aleatoriamente una característica del conjunto de características dado y, a continuación, seleccionando aleatoriamente un valor de división entre los valores máximo y mínimo de esa característica. Esta partición aleatoria de las características producirá recorridos más cortos en los árboles para los puntos de datos anómalos, distinguiéndolos así del resto de los datos. Este algoritmo se basa en el principio de que las anomalías son observaciones que son pocas y diferentes, lo que debería facilitar su identificación. Isolation Forest utiliza un conjunto de árboles de aislamiento para los puntos de datos dados para aislar las anomalías. De esta manera tanto los outliers presentes en el set de datos, como aquellos que puedan haber sido introducidos para generar un *envenenamiento* pueden ser detectados y mitigado el impacto
Label Flipping	Se trata de un tipo de ataque que, para datos capturados de fuentes no seguras, y en los que no es posible obtener los datos de otra manera, existe el riesgo de que un atacante haya podido alterar el etiquetado de los datos en una parte del conjunto de datos de entrenamiento. Aplicado a sistemas de alto riesgo, es importante considerar que debería tenerse en cuenta este concepto, incluso como ocurrido de manera *natural o no malintencionada* pero que el cualquier caso pudiera afectar al rendimiento del sistema. En ese escenario se deben aplicar técnicas para detectar el Re etiquetado de datos para mitigar este efecto. Tal y como se describe en, https://www.researchgate.net/publication/323550082_Label_Sanitization_against_Label_Flipping_Poisoning_Attacks, el procedimiento utiliza k-NN para asignar la etiqueta a cada instancia del conjunto de entrenamiento. El objetivo es reforzar la homogeneidad de la etiqueta entre las instancias que están cerca, especialmente en las regiones que están lejos del límite de decisión
Local Outlier Factor	Es un método que permite tanto detectar outlier *naturales*, es decir que puedan estar presentes de manera natural en una muestra (dada la medición, errores de almacenamiento etc.) como aquellos que pueden haberse insertado como intención de un ataque de envenenamiento. Se

Término	Definición
	trata un método de detección de anomalías no supervisado que calcula la desviación de la densidad local de un punto de datos dado con respecto a sus vecinos. Considera como valores atípicos las muestras que tienen una densidad sustancialmente inferior a la de sus vecinos.
MITRE ATLAS (herramientas/fuentes)	Es una fuente de información de estado del arte de técnicas de ataque adversario, basada en sistemas que se encuentran en producción. Es una fuente de información muy relevante, que debe ser consultada tanto en fase de diseño, como desarrollo. Durante la fase de inferencia, el sistema podrá requerir actualizaciones en relación con casos de uso encontrados en otros sistemas que permitan ataques de transferencia. Toda la información es accesible a través de la dirección https://atlas.mitre.org/
Model stacking	Stacking es un algoritmo de aprendizaje automático conjunto que combina de la mejor manera posible las predicciones de múltiples modelos de aprendizaje automático de buen rendimiento. La ventaja de esta técnica es que puede aprovechar las capacidades de una serie de modelos de buen rendimiento en una tarea de clasificación o regresión y hacer predicciones que tienen mejor rendimiento que cualquier modelo individual del conjunto. El siguiente enlace proporciona un ejemplo de cómo aplicar la técnica usando Python https://machinelearningmastery.com/stacking-ensemble-for-deep-learning-neural-networks/
Neural Dropout	El dropout es un método de regularización que aproxima el entrenamiento de un gran número de redes neuronales con diferentes arquitecturas en paralelo. Durante el entrenamiento, un cierto número de salidas de las capas se ignoran o se "descartan" aleatoriamente. Esto tiene el efecto de hacer que la capa parezca y sea tratada como una capa con un número diferente de nodos y conectividad a la capa anterior. En efecto, cada actualización de una capa durante el entrenamiento se realiza con una "vista" diferente de la capa configurada. Esta técnica se implementa por capa en una red neuronal. Puede utilizarse con la mayoría de los tipos de capas, como las capas densas totalmente conectadas, las capas convolucionales y las capas recurrentes, como la capa de red de memoria a corto plazo. (ver https://arxiv.org/pdf/1806.01246v2.pdf)
OoD Data (Out of	Se trata de datos de entrada, recibidos por el sistema en fase de inferencia una vez puesto el sistema de IA en producción, que no se encuentran dentro del dominio establecido de entrenamiento y

Término	Definición
Domain Data)	relacionado con la finalidad prevista del sistema. Se trata de un indicador de posible ataque adversarial, y pertenece a las técnicas de control de los parámetros de entrada. En el caso de datos estructurados y tabulados, o datos numéricos, series discretas etc., es relativamente sencillo detectar aquellos valores que no se encuentran dentro del dominio y restringir o prohibir su entrada al sistema de inteligencia artificial. En lo referente a sistemas basados en procesamiento del lenguaje natural, es especialmente importante detectar aquellos elementos que no son parte del dominio (ver https://arxiv.org/pdf/1909.03862.pdf).
Privacidad Diferencial	La privacidad diferencial busca proteger los valores de los datos personales mediante la incorporación de "ruido" estadístico al proceso de análisis. Los cálculos necesarios para formar ruido son complejos, pero el principio es bastante intuitivo: el ruido garantiza que las agregaciones de datos permanezcan estadísticamente coherentes con los valores de datos reales, lo que permite una variación aleatoria, pero dificulta trabajar con los valores individuales de los datos agregados. Además, el ruido es diferente para cada análisis, por lo que los resultados son no deterministas; en otras palabras, dos análisis que realizan la misma agregación pueden generar resultados ligeramente diferentes. https://becominghuman.ai/what-is-differential-privacy-1fd7bf507049 https://learn.microsoft.com/es-es/training/modules/explore-differential-privacy/2-understand-differential-privacy
RONI	La premisa de este tipo de mitigación en los ataques se centra en realizar una *simulación* del efecto de la inclusión de los datos de entrada de entrenamiento, **antes de que estos** se incorporen al proceso. Para ello el dato se rechaza en el caso de que su impacto sobre el modelo sea negativo (Reject On Negative Impact) RONI. Este tipo de defensas es especialmente interesante si el sistema tiene una fase de aprendizaje paralela a la inferencia con el sistema en producción, o si la fuente de datos de entrenamiento es muy extensa y difícil de controlar su estado (por ejemplo, textos de email) y por tanto evitar que sean candidatos para envenenar el modelo. Establece un umbral observando el impacto negativo medio de cada instancia en el conjunto de entrenamiento y marca una instancia cuando su impacto en el rendimiento supera el umbral. Se puede encontrar un trabajo detallando el proceso en https://people.eecs.berkeley.edu/~tygar/papers/SML/misleading.learners.pdf

Término	Definición
STRIP	La técnica STRIP está destinada a poder mitigar los ataques de envenenamiento de datos que generan, o permiten la creación de una *puerta trasera* en una red neuronal profunda. Generalmente estos ataques disponen de un *disparador* en los datos que permite activar la respuesta del modelo con los que se ha envenenado éste y generar una clasificación errónea. STRIP es la abreviatura de **STRong Intentional Perturbation,** esta técnica para detectar estas respuestas y está enfocada en sistemas de visión artificial. La técnica ha sido desarrollada en el siguiente paper que puede encontrase aquí https://arxiv.org/abs/1902.06531
Transferencia de ataques adversarios	Esta vulnerabilidad representa la exposición del sistema de inteligencia artificial a recibir ataques adversarios para otros sistemas, dada la similitud que pueda tener con estos estos. Este tipo de ataque se puede realizar una vez que se ha identificado en mayor o menor medida el sistema atacado y puedan utilizarse ataques conocidos para el tipo o familia de modelo.
TRIM	Se trata de una técnica que, especialmente para aplicable al caso de regresiones. El método TRIM estima los parámetros de regresión de forma iterativa, utilizando una función de pérdida recortada (trimmed loss function) para eliminar los puntos con grandes residuos. Tras unas pocas iteraciones TRIM es capaz de aislar la mayoría de los puntos de envenenamiento y aprender un modelo de regresión robusto. Un mayor detalle se puede encontrar en https://www.researchgate.net/figure/Several-iterations-of-the-TRIM-algorithm-Initial-poisoned-data-is-in-blue-in-top-left_fig2_324166859
tRONI	Si el ataque al que estamos expuestos es está dirigido, las instancias de datos envenenadas puede que no causen una caída significativa en el rendimiento. Esta variante aprovecha el conocimiento previo sobre una clasificación errónea en tiempo de prueba para determinar que instancias de entrenamiento podrían haberla causado. Mientras que RONI estima el impacto negativo de una instancia en su conjunto, tRONI considera su efecto exclusivamente en la clasificación del objetivo. Los procedimientos aplicables para la realización de esta técnica se pueden encontrar en https://arxiv.org/pdf/1803.06975.pdf
Z-score	Es una técnica de normalización de los datos, o aplicado en la terminología de Inteligencia Artificial, se trata de un escalado de las características. En esta técnica, los valores se normalizan en función de la media y la desviación estándar de los datos. La esencia de esta técnica es la transformación de los datos mediante la conversión de los

Término	Definición
	valores a una escala común en la que la media es igual a cero y la desviación estándar es uno.

8. Referencias, estándares y normas

Para la elaboración de la guía de ciberseguridad se ha utilizado diversas fuentes consultadas. Se recomienda especialmente a los proveedores y responsable del despliegue la lectura de dos informes elaborados por la Agencia Europea para la Ciberseguridad, ENISA (European Union Agency for Cibersecurity) por sus siglas en inglés.

Dicha agencia europea, elabora documentos relevantes en el ámbito de la ciberseguridad y sin duda será un actor destacado en los aspectos de ciberseguridad en IA relacionados con el Reglamento Europeo

8.1 Estándares

Los siguientes estándares han sido consultados para reflejar en algunos casos los aspectos *intermedios* entre aspectos de seguridad genérica y aquellos específicos de IA. Se ha indicado en los puntos de la guía aquellas recomendaciones que podrían ser de aplicación.

- ISO/IEC 27001 Information technology – Security techniques – Information security management systems – Requirements
- NIST 800-53 Security and Privacy Controls for Information Systems and Organizations
- NISTIR 8269, A Taxonomy and Terminology of Adversarial Machine Learning.

8.2 ENISA

8.2.1 Artificial Intelligence cybersecurity challenges

Publicado en diciembre de 2020 este documento realiza un mapeo del ecosistema de ciberseguridad en IA y su entorno de amenazas. Ha sido realizada por el grupo de trabajo en ciberseguridad para la inteligencia artificial.

Puede consultarse completa en:

- https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges

8.2.2 Securing Machine Learning

Publicado en diciembre de 2021, en este documento se ha realizado revisando de manera sistemática la literatura existente sobre machine learning. El documento hace presentación especial de un análisis de amenazas para los sistemas de machine learning, proporcionando, como se ha indicado en esta guía, una relación con los controles de seguridad.

Puede consultarse completo en:

- https://www.enisa.europa.eu/publications/securing-machine-learning-algorithms

8.2.3 Cybersecurity of AI and standardization

En este documento, publicado en marzo de 2023, se presentan los estándares (existentes, en desarrollo y planificados) relacionados con la ciberseguridad en inteligencia artificial. Presenta sus alcances, y en sus conclusiones ofrece aquellos posibles huecos presentes en la estandarización planificada. Más allá del marco del sandbox de IA, el seguimiento de este tipo de esfuerzos de estandarización de ciberseguridad para IA es muy importante para el proveedor y responsable del despliegue de sistemas de inteligencia artificial de alto riesgo.

Puede consultarse completo en:

- https://www.enisa.europa.eu/publications/cybersecurity-of-ai-and-standardisation

8.3 Otras referencias

- https://www.querypie.com/blog/machine-learning-data-poisoning-and-how-to-prevent-it/
- https://arxiv.org/abs/2009.07008
- https://elie.net/blog/ai/attacks-against-machine-learning-an-overview/#toc-10
- https://arxiv.org/pdf/1706.03691.pdf

Bagging or weight bagging

- https://www.simplilearn.com/tutorials/machine-learning-tutorial/bagging-in-machine-learning
- https://www.projectpro.io/article/bagging-vs-boosting-in-machine-learning/579

Trim Algorithm

- https://www.researchgate.net/figure/Several-iterations-of-the-TRIM-algorithm-Initial-poisoned-data-is-in-blue-in-top-left_fig2_324166859

- https://secml.github.io/class11/

Aumento de datos

- https://arxiv.org/pdf/2010.01862.pdf
- https://research.aimultiple.com/data-augmentation-techniques/
- https://research.aimultiple.com/data-augmentation/

Data sanitization

- https://towardsdatascience.com/5-outlier-detection-methods-that-every-data-enthusiast-must-know-f917bf439210
- https://towardsdatascience.com/poisoning-attacks-on-machine-learning-1ff247c254db

STRIP

- https://arxiv.org/abs/1902.06531

IMPROPER LABELING

- https://www.researchgate.net/publication/323550082_Label_Sanitization_against_Label_Flipping_Poisoning_Attacks

Adversarial Examples:

- Estudio de ataques adversariales en medical machine learning: Adversarial attacks on medical machine learning – MIT Media Lab

Protección vulnerabilidades oráculo

- Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures https://dl.acm.org/doi/pdf/10.1145/2810103.2813677
- Membership Inference Attacks against Machine Learning Models https://arxiv.org/pdf/1610.05820.pdf.
- Secure, Robust and transparent application of AI https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/KI/Secure_robust_and_transparent_application_of_AI.pdf
- A Taxonomy and Terminology of Adversarial Machine Learning https://csrc.nist.gov/publications/detail/nistir/8269/draft
- SoK: Security and privacy in machine learning https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8406613

Protección vulnerabilidades evasión

- Adversarial attack and defence in reinforcement learning-from AI security view https://cybersecurity.springeropen.com/track/pdf/10.1186/s42400-019-0027-x.pdf
- On detecting Adversarial perturbation https://arxiv.org/pdf/1702.04267.pdf

- Modelo más resiliente:
- Algorithm stability on adversarial training https://proceedings.neurips.cc/paper/2021/file/df1f1d20ee86704251795841e6a9405a-Paper.pdf
- Input gradient regularisation https://towardsdatascience.com/the-many-uses-of-input-gradient-regularization-e2af244e6950
- Defensive distillation: https://deepai.org/machine-learning-glossary-and-terms/defensive-distillation
- Generating Adversarial Examples with Adversarial Networks: https://arxiv.org/pdf/1801.02610.pdf
- Adversarial Attacks and Defenses in Deep Learning, https://www.sciencedirect.com/science/article/pii/S209580991930503X
- Distillation as a defense to adversarial perturbations against DNN https://arxiv.org/pdf/1511.04508.pdf
- Towards Deep Learning Models Resistant to Adversarial Attacks https://arxiv.org/pdf/1706.06083.pdf
- Improving the Adversarial Robustness and Interpretability of Deep Neural Network by Regularizing their input Gradients https://arxiv.org/pdf/1711.09404.pdf

Differential privacy

- Modelo PATE (paper https://arxiv.org/abs/1610.05755, explicación http://www.cleverhans.io/privacy/2018/04/29/privacy-and-machine-learning.html)
- Modelo PATE (https://blog.openmined.org/build-pate-differential-privacy-in-pytorch/)
- Security and privacy in ML https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8406613
- Scalable Private Learning with Pate https://arxiv.org/pdf/1802.08908.pdf
- Sustracción de datos confidenciales (membership inference attacks)
- Artículo referenciado https://arxiv.org/pdf/1806.01246v2.pdf.
- También https://www.cs.cmu.edu/~mfredrik/papers/fjr2015ccs.pdf.
- Usar neural dropout y model stacking ayuda a mitigar el riesgo
- Neural Dropout https://towardsdatascience.com/dropout-in-neural-networks-47a162d621d9
- Model stacking https://towardsdatascience.com/simple-model-stacking-explained-and-automated-1b54e4357916

CAT-Gen

- https://arxiv.org/abs/2010.02338
- Reinforcing Adversarial Robustness using Model Confidence Induced by Adversarial Training - Xi Wu, Uyeong Jang, Jiefeng Chen, Lingjiao Chen, Somesh Jha

- Attribution-driven Causal Analysis for Detection of Adversarial Examples, Susmit Jha, Sunny Raj, Steven Fernandes, Sumit Kumar Jha, Somesh Jha, Gunjan Verma, Brian Jalaian, Ananthram Swami

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