ENFOQUE DEL RIESGO EN LA GESTIÓN DE LA IA

La ciberseguridad es una preocupación global que afecta a instituciones públicas, empresas y personas. A medida que más servicios se vuelven digitales y se crean sistemas críticos, es crucial protegerlos debido a los siguientes cambios en la sociedad digital:

- Crecimiento exponencial de la cantidad datos (Big Data).
- Amenazas en permanente evolución: La inteligencia artificial puede adaptarse y aprender de nuevas amenazas para proporcionar una mejor protección.
- Automatización y respuesta rápida: La inteligencia artificial puede responder automáticamente a los ciberataques mucho más rápido que un humano, tomando medidas como aislar dispositivos comprometidos o bloquear accesos sospechosos.

Con la creciente integración de la inteligencia artificial en la ciberseguridad, surgen preocupaciones éticas y regulatorias. ¿Quién es responsable si la IA toma decisiones incorrectas? ¿Cómo aseguramos que la IA sea justa y no discriminativa? Estas son preguntas importantes que subrayan la necesidad de abordar la IA desde una perspectiva técnica, ética y social.

En cuanto a las amenazas de la IA, la Agencia Europea de Ciberseguridad (ENISA) señala diversos actores potencialmente amenazantes, como ciberdelincuentes, personas con información privilegiada, estados-nación o agentes patrocinados por Estados, y competidores. Las amenazas de la IA incluyen:

- Actividades maliciosas: acciones intencionadas para dañar sistemas, infraestructura y redes TIC con el fin de robar, alterar o destruir un objetivo específico.
- Escucha, interceptación y secuestro: acciones destinadas a escuchar, interrumpir o tomar el control de comunicaciones de terceros sin permiso.
- Ataques físicos: acciones dirigidas a dañar, exponer, alterar, inhabilitar, robar o acceder no autorizado a activos físicos, como infraestructuras, hardware o conexiones.
- Daños no intencionados: acciones accidentales que causan destrucción, daño o lesiones a personas o bienes, lo que puede resultar en fallos o pérdida de utilidad.
- Fallos y malfuncionamientos: funcionamiento insuficiente o defectuoso de activos (tanto hardware como software).
- Interrupciones: interrupciones inesperadas del servicio o disminución de la calidad por debajo de lo requerido.
- Desastres: accidentes repentinos o catástrofes naturales que causan daños significativos o pérdidas de vidas humanas.
- Acciones legales: acciones legales contra terceros (ya sean contratantes o no) con el propósito de prohibir ciertas actividades o compensar pérdidas según la legislación aplicable.

Los Libros Blancos de la Comisión Europea son documentos que contienen propuestas de acciones de la Unión Europea (UE) en un campo específico. Recientemente han escrito uno sobre IA, en el que promueven su adopción pero también la gestión de sus riesgos.

Los objetivos de la Unión Europea (UE) en relación con la IA son:

- Asegurar que los sistemas de IA en el mercado de la UE sean seguros y cumplan con la legislación de derechos fundamentales y valores de la UE.
- Proporcionar claridad legal para fomentar la inversión e innovación en IA.
- Mejorar la gobernanza y la aplicación efectiva de la legislación de derechos fundamentales y requisitos de seguridad en sistemas de IA.
- Facilitar un mercado único para un uso legal, seguro y confiable de aplicaciones de IA, evitando divisiones en el mercado.

El Parlamento Europeo y el Consejo proponen regulaciones para abordar las preocupaciones asociadas con la IA, como opacidad, complejidad y autonomía. La IA debe proteger derechos fundamentales, como la dignidad humana, la privacidad, la no discriminación y la igualdad. La ley impone que se tomen las restricciones mínimas necesarias para prevenir riesgos graves y violaciones de derechos fundamentales.

Conceptos:

- Sistema de inteligencia artificial: un software que utiliza técnicas y estrategias para generar información (como contenido, predicciones, recomendaciones o decisiones) con base en objetivos definidos por personas.
- Proveedor: una persona, entidad o autoridad que desarrolla un sistema de IA o utiliza uno desarrollado con la intención de introducirlo en el mercado con su propio nombre o marca comercial, ya sea de forma gratuita o remunerada.
- Uso indebido razonablemente previsible: utilización de un sistema de IA de un modo que no corresponde a su finalidad prevista, que puede derivarse de un comportamiento humano o una interacción con otros sistemas razonablemente previsible.
- Comercialización: el suministro de un sistema de IA para su distribución o uso en el mercado de la Unión Europea como parte de una actividad comercial, ya sea mediante pago o de forma gratuita.
- Modificación sustancial: Cambios realizados en un sistema de IA después de su lanzamiento o introducción en el mercado.
- Datos de entrenamiento: datos usados para entrenar un sistema de IA.
- Datos de prueba: datos usados para evaluar un sistema de IA previamente entrenado y validado, con el fin de confirmar su funcionamiento antes de su lanzamiento o puesta en servicio.

La ciberseguridad es esencial para proteger los sistemas de inteligencia artificial (IA) de posibles ataques maliciosos que buscan aprovechar sus debilidades para alterar su funcionamiento o comprometer su seguridad. Los ciberataques pueden apuntar a partes específicas de la IA, como los datos de entrenamiento (por ejemplo, manipulación de datos) o

los modelos ya entrenados (por ejemplo, ataques adversarios), o explotar vulnerabilidades en los componentes digitales de la IA y su infraestructura TIC.
Los sistemas de IA considerados de alto riesgo (más abajo está explicado qué es) deben ser diseñados y desarrollados de manera que, según su propósito, mantengan un alto nivel de precisión, solidez y ciberseguridad a lo largo de toda su vida útil. Estos sistemas también deben ser capaces de resistir errores, fallos e incoherencias, especialmente cuando interactúan con personas u otros sistemas. Para lograr esta solidez, se pueden implementar soluciones técnicas redundantes, como copias de seguridad o planes para prevenir fallos.

Los problemas que pueden surgir en el contexto de la inteligencia artificial (IA) son diversos:

- Durante la recopilación de los datos:
    - Sobredimensionamiento: riesgo de recopilar más datos de los necesarios, invadiendo la privacidad.
    - Consentimiento: Recopilación de datos sin el conocimiento o consentimiento del usuario, lo que plantea problemas éticos y legales.
- Con el almacenamiento y uso de datos:
    - Seguridad de los datos: almacenar grandes conjuntos de datos hace a las organizaciones objetivos atractivos para ciberdelincuentes, lo que podría resultar en brechas de seguridad y exposición de información confidencial.
    - Perfilado: con suficientes datos, la IA puede usarse para perfilar a los individuos basándose en su comportamiento online, lo que puede llevar a decisiones sesgadas o a discriminación.
- Con la transparencia y toma de decisiones:
    - Decisiones de "caja negra": muchos modelos de IA no explican claramente cómo toman decisiones, lo que puede generar desconfianza y dificultades para verificar su justicia o adecuación.
    - Sesgo y justicia: si los datos utilizados para entrenar modelos de IA están sesgados, las decisiones que toma el modelo también lo estarán.
- Durante la vigilancia y supervisión:
    - Abuso potencial: las soluciones de ciberseguridad basadas en IA que monitorean redes y sistemas para detectar amenazas también pueden ser usadas para vigilar el comportamiento de los usuarios con propósitos maliciosos o invasivos.
- Con la rendición de cuentas y responsabilidad:
    - Falta de responsabilidad: Determinar la responsabilidad en caso de fallos o errores de un sistema basado en IA puede ser complicado, especialmente si no está claro cómo el sistema tomó una decisión particular.
- Con regulaciones y directrices éticas:
    - Necesidad de marcos regulatorios: Para garantizar que se aborden las preocupaciones éticas, es esencial contar con directrices y regulaciones claras que guíen el desarrollo y aplicación de soluciones de IA en ciberseguridad.

Las prácticas de IA se pueden dividir en 4 grupos según el riesgo que supongan:

1. Riesgo inaceptable (prácticas prohibidas):

    a. Sistemas que usen técnicas subliminales que puedan alterar el comportamiento de una persona de un modo que provoque o pueda provocar perjuicios a esa persona u otra.
    b. Sistemas que se aprovechen de alguna de las vulnerabilidades de un grupo específico (edad, discapacidad)
2. Alto riesgo: este grupo abarca el uso de IA en situaciones críticas, como:
    a. infraestructuras críticas que podrían poner en riesgo la vida y salud de los ciudadanos (ej: transporte)
    b. formación educativa y profesional, que puede determinar el acceso a la educación y el curso profesional de la vida de una persona (ej: calificación de exámenes)
    c. componentes de seguridad de los productos (ej: cirugía asistida por robots)
    d. empleo, gestión de trabajadores y acceso al autoempleo (ej: clasificación de CVs)
    e. servicios públicos y privados esenciales (ej: reparto de becas)
    f. aplicación de la ley que pueda interferir con los derechos fundamentales de las personas (ej: evaluación de confiabilidad de las pruebas)
    g. gestión de la migración, el asilo y el control de fronteras (ej: verificación de validez de documentos de viaje)
    h. administración de la justicia y procesos democráticos (ej:aplicación de una ley)
3. Riesgo limitado: se refiere a sistemas de IA que tienen requisitos específicos de transparencia. Por ejemplo, cuando los usuarios interactúan con chatbots, deben ser conscientes de que están hablando con una máquina para tomar decisiones informadas.
4. Riesgo mínimo o nulo: aplicaciones como videojuegos con IA o filtros de spam. La mayoría de los sistemas de IA que se utilizan actualmente en la UE entran en esta categoría.

Para reducir los riesgos de los sistemas de inteligencia artificial (IA) de alto riesgo, se propone un enfoque iterativo y continuo que evalúe los riesgos en diferentes etapas durante todo su ciclo de vida:

1. Identificación y análisis de riesgos conocidos y previsibles vinculados a cada sistema de IA de alto riesgo.
2. Estimación y evaluación de los riesgos que podrían surgir cuando dicho sistema de IA se utilice conforme a su finalidad prevista.
3. Evaluación de otros riesgos que podrían surgir a partir del análisis de los datos recogidos con el sistema de seguimiento posterior a la comercialización.

Aunque el objetivo es que muchos riesgos se reduzcan mediante un diseño y desarrollo adecuados, este análisis también permitirá implantar medidas de mitigación en caso de que algún riesgo no pueda eliminarse.

El resultado de estos análisis quedará descrito en los archivos de registro.

Las pruebas de los sistemas de IA de alto riesgo deberán realizarse durante el proceso de desarrollo y, en todo caso, antes de su introducción en el mercado o puesta en servicio.

Cuando se implante el sistema de gestión de riesgos, se deberá prestar especial atención a la probabilidad de que menores accedan al sistema de IA que se trate o se vean afectados por él. Los conjuntos de datos de entrenamiento, validación y prueba deberán ser pertinentes y representativos, carecer de errores y estar completos. También, tendrán propiedades estadísticas adecuadas. Todos estos conjuntos de datos se someterán a prácticas adecuadas de gobernanza y gestión de datos. De esta forma, nos aseguramos de:

- elegir un diseño adecuado,
- realizar operaciones de tratamiento con los datos en caso de que sea necesario,
- detectar deficiencias en los datos para poder subsanarlas,
- detectar posibles sesgos
- y evaluar la disponibilidad, cantidad y adecuación de los datos.

Cuando sea necesario para abordar sesgos en sistemas de IA de alto riesgo, los proveedores pueden manejar categorías especiales de datos personales (como origen étnico o datos de salud) con salvaguardias adecuadas para proteger los derechos y libertades individuales. Esto incluye la aplicación de medidas de seguridad y protección de la privacidad, como la anonimización o el cifrado (cuando la anonimización pueda afectar significativamente al objetivo perseguido).

Principios del RGPD (Reglamento General de Protección de Datos):

- Licitud, transparencia y lealtad: los datos se deben tratar de manera lícita, leal y transparente para el interesado.
- Finalidad: los datos sólo deben recopilarse y utilizarse para propósitos específicos y legítimos, y no pueden ser procesados de manera incompatible con esos fines.
- Minimización de datos: solo se deben recopilar los datos necesarios y pertinentes para el propósito que se persigue.
- Exactitud: deben tomarse medidas para asegurar que los datos sean precisos y estén actualizados. Si los datos son inexactos, deben corregirse o eliminarse.
- Limitación del plazo de conservación: los datos sólo deben conservarse durante el tiempo necesario para cumplir con los fines para los que se recopilaron y, después, deben ser eliminados o anonimizados.
- Seguridad: se deben tomar medidas para garantizar la seguridad de los datos, incluyendo la integridad, disponibilidad y confidencialidad.
- Responsabilidad activa o demostrada: Las organizaciones deben demostrar constantemente que están cumpliendo con sus obligaciones de protección de datos y que están tomando medidas adecuadas para cuidar la información personal que manejan.

Todo sistema de IA de alto riesgo debe tener una documentación técnica, que se prepara antes de su salida al mercado y se va manteniendo actualizada. Esta documentación debe demostrar que el sistema de IA de alto riesgo cumple los requisitos mencionados anteriormente y proporcionar toda la información necesaria para evaluar si los cumple. Debe contener los siguientes elementos:

1. Descripción general, con:
    a. Finalidad prevista, responsables de desarrollo, fecha y versión del sistema

b. Versiones de software y microprogramas pertinentes
c. Cómo interactúa el sistema de IA con soportes físicos o software externos
d. Descripción del soporte físico en el que se prevé que opere el sistema de IA
e. Instrucciones de uso para el usuario
f. Descripción de todas las formas en las que el sistema de IA se introducirá en el mercado.

2. Descripción detallada, con:
    a. Los métodos y las medidas adoptadas para el desarrollo del sistema de IA, incluido, el recurso a sistemas o herramientas previamente entrenados facilitados por terceros y cómo se han utilizado, integrado o modificado por parte del proveedor.
    b. Las especificaciones de diseño del sistema y la justificación de su elección.
    c. La descripción de la arquitectura del sistema que detalle cómo se incorporan los componentes del software, y cómo se integran en el procesamiento general
    d. Cuando proceda, los requisitos sobre datos en forma de fichas técnicas que describan las metodologías y técnicas de entrenamiento.
    e. La evaluación de las medidas de vigilancia humana necesarias
    f. Los procedimientos de validación y prueba utilizados, incluida la información acerca de los datos de validación y prueba empleados y sus características principales; los parámetros utilizados para medir la precisión, la solidez, la ciberseguridad, etc.
    g. Una descripción de todo cambio introducido en el sistema a lo largo de su ciclo de vida.
    h. Una lista de las normas armonizadas, aplicadas total o parcialmente, cuyas referencias se hayan publicado en el Diario Oficial de la Unión Europea; cuando no se hayan aplicado normas armonizadas.
    i. Una copia de la declaración UE de conformidad, con:
        i. Nombre y tipo del sistema de IA
        ii. Nombre y dirección del proveedor
        iii. Afirmación de que la declaración de la UE de conformidad se emite bajo la exclusiva responsabilidad del proveedor
        iv. Afirmación de que el sistema de IA es conforme con el RGPD y cualquier otra legislación de la UE.
        v. Referencias a todas las normas utilizadas o cualquier otra especificación común respecto a las cuales se declara la conformidad.
        vi. Nombre y número de identificación del organismo notificado, descripción del procedimiento de evaluación de la conformidad llevado a cabo e identificación del certificado emitido.
        vii. Lugar y fecha de emisión de la declaración, nombre y cargo de la persona que la firma, indicación de en nombre o por cuenta de quien lo hace y firma.

Los sistemas de IA de alto riesgo se deben desarrollar de un modo que garanticen un nivel de transparencia suficiente para que los usuarios interpreten y usen correctamente la información

que generan. Deben estar acompañados de instrucciones de uso que sean claras, completas y accesibles para los usuarios, incluyendo datos de contacto del proveedor, información sobre la supervisión humana, las capacidades del sistema de IA y su vida útil prevista.

Con respecto a la vigilancia humana, los sistemas de IA de alto riesgo se deben desarrollar de modo que puedan ser vigilados de manera efectiva por personas. Estas personas:

- Deben entender por completo las capacidades y limitaciones del sistema de IA y controlar adecuadamente su funcionamiento.
- Deben ser conscientes de la posible tendencia a confiar automáticamente o en exceso en la información de salida generada por un sistema de IA de alto riesgo.
- Deben interpretar correctamente la información de salida del sistema de IA de alto riesgo, teniendo en cuenta las características particulares del sistema
- Deben tener la capacidad de intervenir en el funcionamiento del sistema o detenerlo si es necesario a través de un botón específicamente designado para este propósito.

Obligaciones de los proveedores de sistemas de IA de alto riesgo:

- Elaborar la documentación técnica.
- Asegurarse de que los sistemas de IA de alto riesgo sean sometidos a un procedimiento de evaluación de la conformidad antes de su introducción al mercado.
- Obligaciones de registro
- Estar preparados para demostrar, si una autoridad nacional competente lo solicita, que sus sistemas de IA de alto riesgo cumplen los requisitos establecidos.
- Establecer un sistema de control de calidad que garantice el cumplimiento del RGPD, que se documente de manera sistemática y ordenada, y que debe incluir al menos los siguientes aspectos:
    - Una estrategia para el cumplimiento reglamentario, incluido el cumplimiento de los procedimientos de evaluación de la conformidad.
    - Las técnicas, los procedimientos y las actuaciones sistemáticas que se utilizarán en el diseño y el control y la verificación del diseño del sistema.
    - Los procedimientos de examen, prueba y validación que se llevarán a cabo antes, durante y después del desarrollo del sistema de IA de alto riesgo.
    - Las especificaciones técnicas, incluidas las normas, que se aplicarán y, cuando las normas armonizadas pertinentes no se apliquen en su totalidad, los medios que se utilizarán para velar por el sistema de IA.
    - Las técnicas, los procedimientos y las actuaciones sistemáticas que se utilizarán en el diseño y el control y la verificación del diseño del sistema
    - Sistema de gestión de riesgos.

Los organismos notificados (que son los que realizan la evaluación de conformidad y certificación de los productos en relación con las normativas de la UE) tienen las siguientes responsabilidades y requisitos:

- Verificar la conformidad de los sistemas de IA de alto riesgo siguiendo los procedimientos de evaluación de la conformidad establecidos.

- Cumplir con requisitos organizativos, de gestión de calidad, recursos y procesos necesarios para realizar sus funciones.
- La estructura y funcionamiento de estos organismos deben inspirar confianza en la ejecución de sus actividades de evaluación de la conformidad
- Deben ser independientes de los proveedores de sistemas de IA de alto riesgo en los que realizan actividades de evaluación.
- Deben operar de manera independiente y objetiva, garantizando la imparcialidad en sus actividades.
- Deben tener procedimientos que se adapten a factores como el tamaño de las empresas, el sector en el que operan y la complejidad de los sistemas de IA que evalúan.
- Deben tener un seguro de responsabilidad adecuado para cubrir sus actividades de evaluación de conformidad.

Además, antes de que un sistema de IA de alto riesgo se comercialice, estos organismos notificados emiten un certificado que es válido por un período de tiempo determinado, generalmente no superior a 5 años. Si se descubre que un sistema de IA ya no cumple con los requisitos establecidos, el certificado puede ser suspendido, retirado o se pueden imponer restricciones, a menos que el proveedor tome medidas correctivas adecuadas en un plazo específico para garantizar el cumplimiento de los requisitos.

Aspectos de carácter exógeno a la Tecnología IA:

- Normas jurídicas: conjunto de regulaciones legales que son parte del marco legal aplicable a la tecnología de IA
- Principios éticos: conjunto de directrices de comportamiento ético generalmente aceptadas por la comunidad de usuarios de productos y servicios de IA.
- Estándares: conjunto de normas técnicas, a menudo establecidas por organizaciones internacionales, que buscan normalizar la construcción y uso de productos, procesos o servicios relacionados con la IA.

Aspectos de carácter endógeno a la Tecnología IA:

- Sesgo: cualidad no deseada en los resultados de la IA, a menudo causada por un diseño incorrecto o la elección inadecuada de datos de entrenamiento, lo que puede llevar a resultados parciales o sesgados.
- Error: fallos no deliberados en el diseño o desarrollo de la tecnología de IA que pueden resultar en un funcionamiento incorrecto o insatisfactorio de los sistemas.

(Esto viene tal cual en las diapositivas pero no tengo ni idea de lo que es)
Control Ex-ante: medidas de control previas al desarrollo o implementación de la Tecnología, Producto o Servicio IA. Seguridad desde el diseño. Actividades esenciales comprendidas:

- Concepción.
- Planificación.
- Desarrollo.
- Pruebas.

Control Ex-post: medidas de control posteriores al desarrollo o implementación de la Tecnología, Producto o Servicio IA. Seguridad en la operación. Actividades esenciales comprendidas:

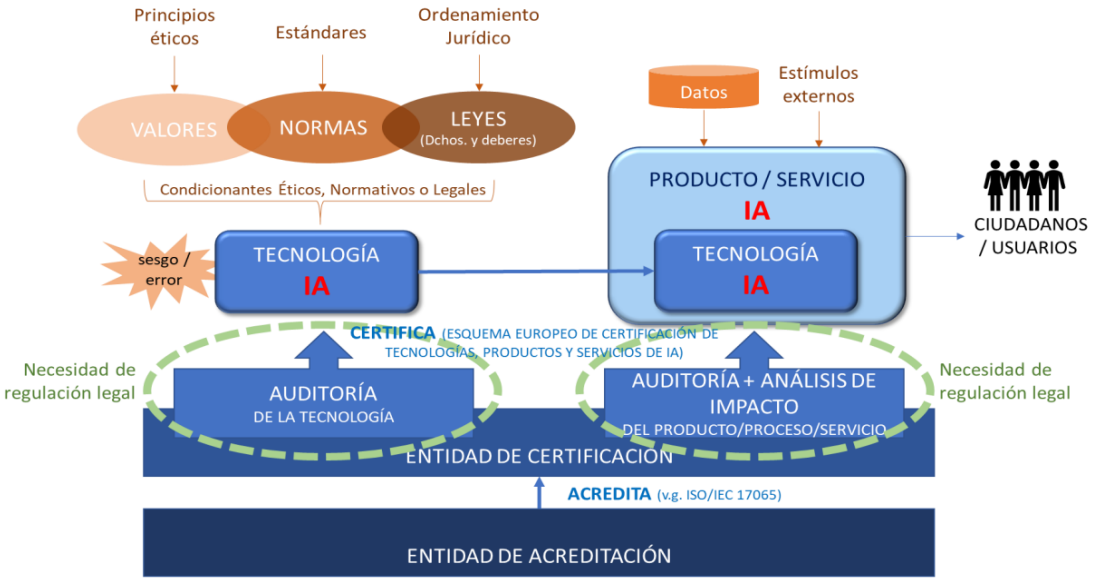
- Integración.
- Adquisición.
- Despliegue.
- Explotación.
- Publicación.
- Conservación.
- Acceso.
- Interconexión.

Verificación de la conformidad de la solución con:

- El ordenamiento jurídico aplicable.
- Principios éticos.
- Estándares técnicos.
- Análisis de impacto en la sociedad (comunidad destinataria de los productos/servicios).

Satisfacción de las exigencias de:

- Transparencia e inteligibilidad de los sistemas.
- Posibilidad de acceso y verificación.
- Explicabilidad, trazabilidad y rendición de cuentas (responsabilidad).



El modelo propuesto de Certificación de la IA en Europa