Departamento de
Ciencia, Innovación
y Tecnología

Documento de política

Código de buenas prácticas para la ciberseguridad de la IA

Publicado el 31 de enero de 2025

Contenido

This publication is available at https://www.gov.uk/government/publications/ai-cyber-security-code-of-practice/code-of-practice-for-the-cyber-security-of-ai

Introducción

El gobierno del Reino Unido está llevando a cabo una intervención en dos partes para abordar los riesgos de seguridad cibernética para la IA. Esto implica el desarrollo de un Código de Buenas Prácticas voluntario que se utilizará para ayudar a crear una norma mundial en el Instituto Europeo de Normas de Telecomunicaciones (ETSI) que establezca los requisitos de seguridad básicos. Creemos que se necesita un código centrado específicamente en la ciberseguridad de la IA porque la IA tiene claras diferencias con el software. Estos incluyen riesgos de seguridad por envenenamiento de datos, ofuscación de modelos, inyección indirecta de avisos y diferencias operativas asociadas con la administración de datos. En el Apéndice B del Marco de Gestión de Riesgos del Instituto Nacional de Normas y Tecnología (NIST) se pueden encontrar más ejemplos de los riesgos únicos que plantean los sistemas de IA.

El gobierno también está interviniendo en esta área porque el software debe ser seguro por diseño y las partes interesadas en la cadena de suministro de IA requieren claridad sobre qué requisitos de seguridad básicos deben implementar para proteger los sistemas de IA.

La intervención propuesta fue respaldada por el 80% de los encuestados en la convocatoria (https://www.gov.uk/government/calls-for-evidence/cyber-security-of-ai-a-call-for-views) del Departamento de Ciencia, Innovación y Tecnología (DSIT), que se celebró del 15 de mayo al 9 de agosto de 2024. El apoyo a cada principio del Código osciló entre el 83% y el 90%. Este documento también se basa en las Directrices del NCSC para el desarrollo seguro de la IA (https://www.ncsc.gov.uk/collection/guidelines-secure-ai-system-development), que se publicaron en noviembre de 2023 y fueron respaldadas por 19 socios internacionales. Como se establece en el enfoque modular de DSIT para los códigos de prácticas de ciberseguridad (https://www.gov.uk/government/collections/cyber-security-codes-of-practice), las partes interesadas en la IA deben considerar este documento como un apéndice al Código de Prácticas de Software.1

Alcance

El ámbito de aplicación de este Código de Buenas Prácticas voluntario se centra en los sistemas de IA. Esto incluye sistemas que incorporan redes neuronales profundas, como la IA generativa. En aras de la coherencia, hemos utilizado el término "sistemas de IA" en todo el documento al enmarcar el alcance de las disposiciones y la "seguridad de la IA", que se considera un subconjunto de la ciberseguridad. El Código no está diseñado para académicos que están creando y probando sistemas de IA solo con fines de investigación (sistemas de IA que no se van a implementar).

El Código establece los requisitos de ciberseguridad para el ciclo de vida de la IA. Reconocemos que no existe una visión coherente en los marcos internacionales sobre lo que conforma el ciclo de vida de la IA. Sin embargo, para ayudar a las partes interesadas, hemos separado los principios en cinco fases. Estos son el diseño seguro, el desarrollo seguro, la implementación segura, el mantenimiento seguro y el final de la vida útil seguro. También hemos señalizado las normas y publicaciones pertinentes al comienzo de cada principio para poner de relieve los vínculos entre los distintos documentos y el Código. Esta no es una lista exhaustiva.2

Guía de implementación

A raíz de los comentarios de la convocatoria de opiniones, hemos creado una guía de implementación para ayudar a las organizaciones a cumplir con los requisitos del Código voluntario (y de la futura norma). La guía se elaboró tras una revisión exhaustiva de los estándares y marcos de software e IA, así como de documentos publicados por otros gobiernos y reguladores. El gobierno del Reino Unido planea presentar el Código y la Guía de Implementación en ETSI para que la futura norma vaya acompañada de una guía. El gobierno actualizará el contenido del Código y la Guía para reflejar la futura norma y guía mundial del ETSI.

Audiencia

En esta sección se definen los grupos de partes interesadas que forman la cadena de suministro de IA. Se da una indicación para cada principio en el que las partes interesadas son las principales responsables de su aplicación. Es importante destacar que una misma entidad puede desempeñar múltiples funciones de partes interesadas en este Código voluntario, así como responsabilidades de diferentes regímenes normativos3.

Todas las partes interesadas incluidas en la siguiente tabla deben tener en cuenta que cuando los datos utilizados para un sistema de IA son personales (incluidos los datos seudonimizados), pueden tener obligaciones de protección de datos y deberán consultar la guía de protección de datos del Reino Unido ofrecida por la ICO.4 Además, los líderes sénior de una organización también tienen la responsabilidad de ayudar a proteger a su personal e infraestructura, como se señala en el Código de Prácticas de Gobernanza Cibernética de (https://www.gov.uk/government/calls-for-evidence/cyber-governance-code-of-practice-call-for-views) DSIT. Algunas disposiciones del Código relativas a los

desarrolladores son menos aplicables a los sistemas de IA que utilizan modelos de código abierto. Recomendamos a los desarrolladores que revisen la Guía de implementación para confirmar qué requisitos se especifican para los diferentes tipos de sistemas de IA.

Interesado	Definiciones
Desarrolladores	Esto abarca cualquier tipo de empresa u organización de cualquier sector, así como a las personas que son responsables de crear o adaptar un modelo y/o sistema de IA. Esto se aplica a todas las tecnologías de IA, incluidos los modelos propietarios y de código abierto. Para contextualizar, una empresa u organización que crea un modelo de IA y que también es responsable de integrar/desplegar ese modelo/sistema en su organización se definiría en este Código voluntario como un desarrollador y un operador del sistema.
Operadores del sistema	Esto incluye cualquier tipo de empresa u organización de cualquier sector que tenga la responsabilidad de incorporar/desplegar un modelo y un sistema de IA en su infraestructura. Esto se aplica a todas las tecnologías de IA, incluidos los modelos propietarios y de código abierto. Este término también incluye a aquellas empresas que proporcionan un servicio contractual a las organizaciones para integrar/desplegar un modelo y sistema de IA con fines empresariales.
Custodios de datos	Esto incluye cualquier tipo de empresa, organización o individuo que controle los permisos de datos y la integridad de los datos que se utilizan para que funcione cualquier modelo o sistema de IA. Este grupo de partes interesadas también incluye a las entidades que establecen las políticas sobre cómo se utilizan y gestionan los datos para un modelo y/o sistema de IA. En el contexto de un sistema de IA, podría haber varios custodios de datos involucrados porque algunos datos utilizados para crear un modelo podrían provenir de la organización que está implementando/integrando el sistema en su infraestructura y otros datos podrían provenir de bases de datos públicas y otras fuentes.
Usuarios finales	Esto abarca a cualquier empleado de una organización o empresa y a los consumidores del Reino Unido que utilicen un modelo y sistema de IA para cualquier propósito, incluso para respaldar su trabajo y sus actividades cotidianas. Esto se aplica a todas las tecnologías de IA y a los modelos propietarios y de

Interesado	Definiciones
	código abierto. Este grupo de partes interesadas se ha creado porque el Código voluntario ha depositado expectativas en los desarrolladores, operadores de sistemas y custodios de datos para ayudar a informar y proteger a los usuarios finales.
Entidades afectadas	Abarca a todas las personas y tecnologías, como aplicaciones y sistemas autónomos, que no se ven directamente afectadas por los sistemas de IA o las decisiones basadas en los resultados de los sistemas de IA. Estas personas no necesariamente interactúan con el sistema o la aplicación implementados.

La siguiente tabla ofrece ejemplos de casos comunes que involucran a diferentes tipos de organizaciones que son relevantes para este Código de Prácticas voluntario, así como para el Código de Prácticas voluntario de Resiliencia de Software.

Grupos de partes interesadas	Orientación
Proveedores de software que también ofrecen servicios de IA a clientes/usuarios finales	Es probable que estas organizaciones sean desarrolladores y, por lo tanto, están dentro del alcance de este Código y del Código de Prácticas de Resiliencia del Software.
Proveedores de software que utilizan IA en su propia infraestructura que ha sido creada por un proveedor externo	Es probable que estas organizaciones sean operadores de sistemas y, por lo tanto, están dentro del alcance de las partes relevantes del Código y del Código de Prácticas de Resiliencia del Software.
Proveedores de software que crean IA internamente y la implementan dentro de su infraestructura	Es probable que estas organizaciones sean desarrolladores y operadores de sistemas y, por lo tanto, están dentro del alcance de este Código y del Código de Prácticas de Resiliencia del Software.
Proveedores de software que solo utilizan IA (componentes) de terceros para su uso interno	Es probable que estas organizaciones sean operadores de sistemas y, por lo tanto, están dentro del alcance de las partes relevantes del Código y del Código de Prácticas de Resiliencia del Software.

Grupos de partes interesadas	Orientación
Organización que crea un sistema de IA para uso interno	Es probable que estas organizaciones sean Desarrolladores y, por lo tanto, están dentro del alcance de este Código.
Organización que solo utiliza componentes de IA de terceros	Es probable que estas organizaciones sean operadores de sistemas y, por lo tanto, estén dentro del alcance de las partes relevantes del Código.
Proveedores de IA	Las organizaciones que ofrecen o venden modelos y componentes, pero que no desempeñan un papel en el desarrollo o la implementación de los mismos, no es probable que estén dentro del ámbito de aplicación de este Código. Es probable que estas organizaciones estén dentro del ámbito de aplicación del Código de Prácticas de Software y del Código de Gobernanza Cibernética.

Terminología

Hemos utilizado la terminología "deberá" y "debería" para cada disposición del Código voluntario para alinearse con la redacción utilizada por las organizaciones de desarrollo de estándares. En el cuadro que figura a continuación se exponen las definiciones de estas palabras en el contexto del carácter voluntario del presente Código de Buenas Prácticas. En el Anexo A figura un glosario.

Término	Definición
Deber	Indica un requisito para el Código voluntario
Deber	Indica una recomendación para el Código voluntario
Mayo	Indica dónde algo es posible, por ejemplo, que una organización o individuo es capaz de hacer algo

Estructura del Código de Buenas Prácticas voluntario

Principio 1: Concienciar sobre las amenazas y los riesgos de seguridad de la IA

Principio 2: Diseñe su sistema de IA para la seguridad, así como para la funcionalidad y el rendimiento

Principio 3: Evalúe las amenazas y gestione los riesgos para su sistema de IA

Principio 4: Permitir la responsabilidad humana de los sistemas de IA

Principio 5: Identifique, rastree y proteja sus activos

Principio 6: Proteja su infraestructura

Principio 7: Asegure su cadena de suministro

Principio 8: Documente sus datos, modelos e indicaciones

Principio 9: Realizar pruebas y evaluaciones apropiadas

Principio 10: Comunicación y procesos asociados con los usuarios finales y las entidades afectadas

Principio 11: Mantener actualizaciones de seguridad, parches y mitigaciones periódicas

Principio 12: Supervise el comportamiento de su sistema

Principio 13: Garantizar la eliminación adecuada de datos y modelos

Principios del Código de Buenas Prácticas

Diseño seguro

Principio 1: Concienciar sobre las amenazas y los riesgos de seguridad de la IA

Se aplica principalmente a: operadores de sistemas, desarrolladores y custodios de datos

[NIST 2022, NIST 2023, ASD 2023, WEF 2024, OWASP 2024, MITRE 2024, Google 2023, ESLA 2023, Cisco 2022, Deloitte 2023, Microsoft 2022].

1.1. El programa de formación en ciberseguridad de las organizaciones incluirá contenidos de seguridad de la IA que se revisarán y actualizarán periódicamente, por ejemplo, en caso de que surjan nuevas amenazas importantes para la seguridad relacionadas con la IA.

1.1.1 La formación en seguridad de la IA se adaptará a las funciones y responsabilidades específicas de los miembros del personal.

1.2. Como parte de un programa más amplio de formación del personal de una organización, exigirán a todo el personal que esté al tanto de las últimas amenazas y vulnerabilidades de seguridad relacionadas con la IA. Cuando esté disponible, este conocimiento incluirá las mitigaciones propuestas.

1.2.1. Estas actualizaciones deben comunicarse a través de múltiples canales, como boletines de seguridad, boletines informativos o plataformas internas de intercambio de conocimientos. Esto garantizará una amplia difusión y comprensión entre el personal.

1.2.2 Las organizaciones proporcionarán a los desarrolladores formación en codificación segura y técnicas de diseño de sistemas específicas para el desarrollo de la IA, centrándose en la prevención y mitigación de las vulnerabilidades de seguridad en los algoritmos, modelos y software asociado de la IA.

Principio 2: Diseñe su sistema de IA para la seguridad, así como para la funcionalidad y el rendimiento

Se aplica principalmente a: Operadores y desarrolladores de sistemas

[OWASP 2024, MITRE 2024, WEF 2024, ENISA 2023, NCSC 2023, BSI1 2023, Cisco 2022, Microsoft 2022, G7 2023, HHS 2021, OpenAI2 2024, ASD 2023, ICO 2020].

2.1 Como parte de la decisión de crear un sistema de IA, un Operador del Sistema y/o un Desarrollador deberán llevar a cabo una evaluación exhaustiva que incluya la determinación y documentación de los requisitos comerciales y/o el problema que buscan abordar, junto con los riesgos de seguridad de IA asociados y las estrategias de mitigación.5

2.1.1 Cuando el Custodio de Datos forme parte de una organización de Desarrolladores, se le incluirá en las discusiones internas a la hora de determinar los requisitos y las necesidades de datos de un sistema de IA.

2.2: Los desarrolladores y operadores de sistemas se asegurarán de que los sistemas de IA estén diseñados e implementados para resistir ataques de IA adversarios, entradas inesperadas y fallos del sistema de IA.

2.3 Para respaldar el proceso de preparación de datos, auditoría de seguridad y respuesta a incidentes para un sistema de IA, los Desarrolladores documentarán y crearán una pista de auditoría en relación con el sistema de IA. Esto incluirá el funcionamiento y la gestión del ciclo de vida de los modelos, conjuntos de datos y avisos incorporados al sistema.

2.4 Si un Desarrollador u Operador del Sistema utiliza un componente externo, deberá llevar a cabo una evaluación de riesgos de seguridad de la IA y un proceso de diligencia debida en línea con sus procesos de desarrollo de software existentes, que evalúe los riesgos específicos de la IA.6

2.5 Los Custodios de Datos se asegurarán de que el uso previsto del sistema sea adecuado a la sensibilidad de los datos con los que se entrenó, así como a los controles destinados a garantizar la seguridad de los datos.

2.5.1 Las organizaciones deben asegurarse de que se aliente a los empleados a informar e identificar de forma proactiva cualquier riesgo potencial de seguridad en los sistemas de IA y garantizar que se implementen las salvaguardias adecuadas.

2.6 Cuando el sistema de IA vaya a interactuar con otros sistemas o fuentes de datos (ya sean internas o externas), los desarrolladores y operadores de sistemas se asegurarán de que los permisos concedidos al sistema de IA en otros sistemas solo se proporcionen según sea necesario para la funcionalidad y se evalúen los riesgos.

2.7 Si un Desarrollador u Operador de Sistema decide trabajar con un proveedor externo, deberá llevar a cabo una evaluación de diligencia debida y asegurarse de que el proveedor se adhiere a este Código de Prácticas.

Principio 3: Evalúe las amenazas y gestione los riesgos para su sistema de IA

Se aplica principalmente a: desarrolladores y operadores de sistemas

[OWASP 2024, WEF 2024, Nvidia 2023, ENISA 2023, Google 2023, G7 2023, NCSC 2023, Deloitte 2023], MITRE, OWASP, NIST Taxonomía de riesgos, ISO 27001]

3.1 Los desarrolladores y operadores de sistemas analizarán las amenazas y gestionarán los riesgos de seguridad para sus sistemas. El modelado de

amenazas debe incluir revisiones y actualizaciones periódicas y abordar los ataques específicos de la IA, como el envenenamiento de datos, la inversión del modelo y la inferencia de pertenencia.

3.1.1 El proceso de modelización de amenazas y gestión de riesgos se llevará a cabo para abordar cualquier riesgo de seguridad que surja cuando se implemente o actualice una nueva opción de configuración o ajuste en cualquier etapa del ciclo de vida de la IA.

3.1.2 Los desarrolladores gestionarán los riesgos de seguridad asociados a los modelos de IA que proporcionen funcionalidades superfluas, en los que una mayor funcionalidad conduzca a un mayor riesgo. Por ejemplo, cuando se utiliza un modelo multimodal pero solo se utiliza una modalidad única para el funcionamiento del sistema.

3.1.3 Los gestores de sistemas aplicarán controles a los riesgos identificados a través del análisis sobre la base de una serie de consideraciones, incluido el coste de implementación, de acuerdo con su tolerancia al riesgo institucional.

3.2 Cuando se identifiquen amenazas a la seguridad de la IA que no puedan ser resueltas por los desarrolladores, se comunicará a los operadores del sistema para que puedan modelar sus sistemas mediante amenazas. Los Operadores del Sistema comunicarán esta información a los Usuarios Finales, para que tengan conocimiento de estas amenazas. Esta comunicación debe incluir descripciones detalladas de los riesgos, los impactos potenciales y las acciones recomendadas para abordar o monitorear estas amenazas.

3.3 Cuando una entidad externa sea responsable de los riesgos de seguridad de la IA identificados dentro de la infraestructura de una organización, los operadores del sistema deben obtener la garantía de que estas partes son capaces de hacer frente a dichos riesgos.

3.4 Los desarrolladores y operadores de sistemas deben monitorear y revisar continuamente su infraestructura de sistemas de acuerdo con el apetito de riesgo. Es importante reconocer que seguirá existiendo un mayor nivel de riesgo en los sistemas de IA a pesar de la aplicación de controles para mitigarlos.

Principio 4: Permitir la responsabilidad humana de los sistemas de IA

Se aplica principalmente a: desarrolladores y operadores de sistemas

[OWASP 2024, MITRE 2024, BSI1 2023, Microsoft 2022]

4.1 Al diseñar un sistema de IA, los desarrolladores y/o operadores de sistemas deben incorporar y mantener capacidades para permitir la supervisión humana.7

4.2 Los desarrolladores deben diseñar sistemas que faciliten a los humanos la evaluación de los resultados de los que son responsables en dicho sistema (por ejemplo, asegurándose de que los resultados de los modelos sean explicables o interpretables).

4.3 Cuando la supervisión humana sea un control de riesgos, los Desarrolladores y/o los Operadores del Sistema deberán diseñar, desarrollar, verificar y mantener medidas técnicas para reducir el riesgo a través de dicha supervisión.

4.4 Los desarrolladores deben verificar que los controles de seguridad especificados por el Custodio de Datos se hayan integrado en el sistema.

4.5 Los desarrolladores y operadores de sistemas deben informar a los usuarios finales de los casos de uso prohibidos del sistema de IA.

Desarrollo seguro

Principio 5: Identifique, rastree y proteja sus activos

Se aplica principalmente a: desarrolladores, operadores de sistemas y custodios de datos

[OWASP 2024, Nvidia 2023, NCSC 2023, BSI1 2023, Cisco 2022, Deloitte 2023, Amazon 2023, G7 2023, ICO 2020]

5.1 Los desarrolladores, custodios de datos y operadores de sistemas mantendrán un inventario exhaustivo de sus activos (incluidas sus interdependencias/conectividad).

5.2 Como parte de prácticas de seguridad de software más amplias, los desarrolladores, custodios de datos y operadores de sistemas deberán disponer de procesos y herramientas para rastrear, autenticar, gestionar el control de versiones y proteger sus activos debido a la creciente complejidad de los activos específicos de la IA.

5.3 Los gestores de sistemas desarrollarán y adaptarán sus planes de recuperación en caso de catástrofe para tener en cuenta ataques específicos dirigidos a sistemas de IA.

5.3.1 Los operadores del sistema deben asegurarse de que se pueda restablecer un buen estado conocido.

5.4 Los desarrolladores, los operadores de sistemas, los custodios de datos y los usuarios finales protegerán los datos sensibles, como los datos de formación o de prueba, contra el acceso no autorizado.

5.4.1 Los desarrolladores, los custodios de datos y los operadores del sistema aplicarán comprobaciones y saneamiento a los datos y entradas al diseñar el modelo en función de su acceso a dichos datos y entradas y del lugar donde se almacenan dichos datos y entradas. Esto se repetirá cuando las revisiones del modelo se realicen en respuesta a los comentarios de los usuarios o al aprendizaje continuo.

5.4.2 En los casos en que los datos de entrenamiento o las ponderaciones del modelo puedan ser confidenciales, los Desarrolladores establecerán protecciones proporcionadas.

Principio 6: Proteja su infraestructura

Se aplica principalmente a: desarrolladores y operadores de sistemas

[OWASP 2024, MITRE 2024, WEF 2024, NCSC 2023, Microsoft 2022, ICO 2020]

6.1 Los desarrolladores y operadores de sistemas evaluarán los marcos de control de acceso de su organización e identificarán las medidas adecuadas para proteger las API, los modelos, los datos y las tuberías de formación y procesamiento.

6.2 Si un Desarrollador ofrece una API a clientes o colaboradores externos, deberá aplicar controles que mitiguen los ataques al sistema de IA a través de la API. Por ejemplo, poner límites a la velocidad de acceso al modelo para limitar la capacidad de un atacante de aplicar ingeniería inversa o abrumar las defensas para envenenar rápidamente un modelo.

6.3 Los desarrolladores también deberán crear entornos dedicados para las actividades de desarrollo y ajuste de modelos. Los entornos dedicados estarán respaldados por controles técnicos que garanticen la separación y el principio de mínimo privilegio. En el contexto de la IA, esto es especialmente necesario porque los datos de entrenamiento solo estarán presentes en los entornos de formación y desarrollo en los que estos datos de entrenamiento no se basen en datos disponibles públicamente.

6.4 Los desarrolladores y operadores de sistemas implementarán y publicarán una política de divulgación de vulnerabilidades clara y accesible.

6.5 Los desarrolladores y operadores de sistemas crearán, probarán y mantendrán un plan de gestión de incidentes del sistema de IA y un plan de recuperación del sistema de IA.

6.6 Los desarrolladores y operadores de sistemas deben asegurarse de que, cuando utilicen operadores de servicios en la nube para ayudar a ofrecer la capacidad, sus acuerdos contractuales respalden el cumplimiento de los requisitos anteriores.

Principio 7: Asegure su cadena de suministro

Se aplica principalmente a: desarrolladores, operadores de sistemas y custodios de datos

[Lista de materiales de software (SBOM), CISA, (https://www.cisa.gov/sbom0) OWASP 2024, NCSC 2023, Microsoft 2022, ASD 2023]

7.1 Los desarrolladores y operadores de sistemas seguirán procesos seguros de la cadena de suministro de software para el desarrollo de su modelo de IA y sistema.

7.2 Los Operadores de Sistemas que opten por utilizar o adaptar modelos o componentes que no estén bien documentados o asegurados deberán justificar su decisión de utilizar dichos modelos o componentes mediante documentación (por ejemplo, si no había otro proveedor para dicho componente).

7.2.1 En este caso, los Desarrolladores y Operadores de Sistemas deberán disponer de controles mitigantes y realizar una evaluación de riesgos vinculada a dichos modelos o componentes.

7.2.2 Los Operadores del Sistema compartirán esta documentación con los Usuarios Finales de forma accesible.

7.3 Los desarrolladores y operadores de sistemas deberán volver a realizar evaluaciones sobre los modelos publicados que tengan la intención de utilizar.

7.4 Los operadores de sistemas comunicarán su intención de actualizar los modelos a los usuarios finales de forma accesible antes de que se actualicen los modelos.

Principio 8: Documente sus datos, modelos e indicaciones

Se aplica principalmente a: Desarrolladores

[OWASP 2024, WEF 2024, NCSC 2023, Cisco 2022, Microsoft 2022, ICO 2020]

8.1 Los desarrolladores documentarán y mantendrán una pista de auditoría clara del diseño de su sistema y de los planes de mantenimiento posteriores a la implementación. Los desarrolladores deben poner la documentación a disposición de los operadores del sistema y los custodios de datos posteriores.

8.1.1 Los desarrolladores deben asegurarse de que el documento incluya información relevante para la seguridad, como las fuentes de los datos de entrenamiento (incluidos los datos de ajuste y los comentarios humanos u otros comentarios operativos), el alcance y las limitaciones previstos, las

medidas de protección, el tiempo de retención, la frecuencia de revisión sugerida y los posibles modos de error.

8.1.2 Los desarrolladores liberarán hashes criptográficos para los componentes del modelo que se pondrán a disposición de otras partes interesadas para permitirles verificar la autenticidad de los componentes.

8.2 Cuando los datos de entrenamiento se han obtenido de fuentes disponibles públicamente, existe el riesgo de que estos datos puedan haber sido envenenados. Dado que es probable que se descubran datos envenenados después del entrenamiento (si es que se producen), los desarrolladores documentarán cómo obtuvieron los datos de entrenamiento públicos, de dónde provienen y cómo se utilizan esos datos en el modelo.

8.2.1. La documentación de los datos de entrenamiento debe incluir, como mínimo, la fuente de los datos, como la URL de la página raspada, y la fecha/hora en que se obtuvieron los datos. Esto permitirá a los desarrolladores identificar si un ataque de envenenamiento de datos reportado estaba en sus conjuntos de datos.

8.3 Los desarrolladores deben asegurarse de tener un registro de auditoría de los cambios en las indicaciones del sistema u otra configuración del modelo (incluidas las solicitudes) que afecten al funcionamiento subyacente de los sistemas. Los desarrolladores pueden ponerlo a disposición de cualquier operador del sistema y usuario final que tenga acceso al modelo.

Principio 9: Realizar pruebas y evaluaciones apropiadas

Se aplica principalmente a: desarrolladores y operadores de sistemas

[OWASP 2024, WEF 2024, Nvidia 2023, NCSC 2023, ENISA 2023, Google 2023, G7 2023]

9.1 Los desarrolladores se asegurarán de que todos los modelos, aplicaciones y sistemas que se publiquen a los operadores del sistema y/o a los usuarios finales hayan sido probados como parte de un proceso de evaluación de la seguridad.

9.2 Los Operadores del Sistema deberán realizar pruebas antes de que el sistema se implemente con el apoyo de los Desarrolladores.

9.2.1 Para las pruebas de seguridad, los operadores y desarrolladores de sistemas deben utilizar probadores de seguridad independientes con habilidades técnicas relevantes para sus sistemas de IA.

9.3 Los desarrolladores deben asegurarse de que los resultados de las pruebas y la evaluación se compartan con los operadores del sistema, para informar sus propias pruebas y evaluaciones.

9.4 Los desarrolladores deben evaluar los resultados del modelo para asegurarse de que no permiten que los operadores del sistema o los usuarios finales realicen ingeniería inversa de aspectos no públicos del modelo o de los datos de entrenamiento.

9.4.1 Además, los Desarrolladores deben evaluar los resultados del modelo para asegurarse de que no proporcionan a los Operadores del Sistema o a los Usuarios Finales una influencia no deseada sobre el sistema.

Implementación segura

Principio 10: Comunicación y procesos asociados con los usuarios finales y las entidades afectadas

Como parte de las prácticas de implementación más amplias de una organización, también deben considerar la posibilidad de realizar pruebas previas a la implementación de los sistemas de IA junto con los requisitos que se indican a continuación.

10.1 Los Operadores de Sistemas deberán transmitir a los Usuarios Finales de forma accesible dónde y cómo se utilizarán, accederán y almacenarán sus datos (por ejemplo, si se utilizan para el reentrenamiento de modelos o si son revisados por empleados o socios).8 Si el Desarrollador es una entidad externa, deberá proporcionar esta información a los Operadores de Sistema.

10.2 Los operadores de sistemas proporcionarán a los usuarios finales orientaciones accesibles para respaldar su uso, gestión, integración y configuración de los sistemas de IA. Si el Desarrollador es una entidad externa, deberá proporcionar toda la información necesaria para ayudar a los Operadores del Sistema.

10.2.1 Los operadores del sistema incluirán orientación sobre el uso adecuado del modelo o sistema, que incluye destacar las limitaciones y los posibles modos de fallo.

10.2.2 Los operadores de sistemas informarán de forma proactiva a los usuarios finales de cualquier actualización relevante para la seguridad y proporcionarán explicaciones claras de forma accesible.

10.3 Los desarrolladores y operadores de sistemas deben apoyar a los usuarios finales y a las entidades afectadas durante y después de un incidente de ciberseguridad para contener y mitigar los impactos de un incidente. El proceso para llevar a cabo esto debe estar documentado y acordado en contratos con los usuarios finales.

Mantenimiento seguro

Principio 11: Mantener actualizaciones de seguridad, parches y mitigaciones periódicas

Se aplica principalmente a: desarrolladores y operadores de sistemas

[ICO 2020]

11.1 Los desarrolladores proporcionarán actualizaciones y parches de seguridad, siempre que sea posible, y notificarán a los operadores del sistema de las actualizaciones de seguridad. Los Operadores del Sistema entregarán estas actualizaciones y parches a los Usuarios Finales.

11.1.1 Los desarrolladores deberán contar con mecanismos y planes de contingencia para mitigar los riesgos de seguridad, especialmente en los casos en que no se puedan proporcionar actualizaciones para los sistemas de IA.

11.2 Los desarrolladores deben tratar las actualizaciones importantes del sistema de IA como si se hubiera desarrollado una nueva versión de un modelo y, por lo tanto, emprender un nuevo proceso de evaluación y pruebas de seguridad para ayudar a proteger a los usuarios.

11.3 Los desarrolladores deben apoyar a los operadores del sistema para evaluar y responder a los cambios en el modelo (por ejemplo, proporcionando acceso a la vista previa a través de pruebas beta y API con versiones).

Principio 12: Supervise el comportamiento de su sistema

Se aplica principalmente a: desarrolladores y operadores de sistemas

[OWASP 2024, WEF 2024, Nvidia 2023, ENISA 2023, BSI1 2023, Cisco 2022, Deloitte 2023, G7 2023, Amazon 2023, ICO 2020]

12.1 Los operadores del sistema registrarán las acciones del sistema y del usuario para respaldar el cumplimiento de la seguridad, las investigaciones de incidentes y la corrección de vulnerabilidades.

12.2 Los operadores del sistema deben analizar sus registros para asegurarse de que los modelos de IA siguen produciendo los resultados deseados y para detectar anomalías, violaciones de seguridad o comportamientos inesperados a lo largo del tiempo (por ejemplo, debido a la deriva o el envenenamiento de datos).

12.3 Los operadores y desarrolladores de sistemas deben supervisar los estados internos de sus sistemas de IA cuando esto pueda permitirles

hacer frente mejor a las amenazas de seguridad o permitir futuros análisis de seguridad.

12.4 Los operadores y desarrolladores de sistemas deben supervisar el rendimiento de sus modelos y sistemas a lo largo del tiempo para que puedan detectar cambios repentinos o graduales en el comportamiento que puedan afectar a la seguridad.

Fin de vida útil seguro

Principio 13: Garantizar la eliminación adecuada de datos y modelos
Se aplica principalmente a: desarrolladores y operadores de sistemas

13.1 Si un Desarrollador u Operador del Sistema decide transferir o compartir la propiedad de los datos de entrenamiento y/o un modelo a otra entidad, deberá involucrar a los Custodios de Datos y disponer de forma segura de estos activos. Esto protegerá los problemas de seguridad de la IA que puedan transferirse de una instancia de un sistema de IA a otra.

13.2 Si un Desarrollador u Operadores de Sistemas decide retirar un modelo y/o sistema, deberán involucrar a los Custodios de Datos y eliminar de forma segura los datos y detalles de configuración aplicables.

Anexo A: Glosario de términos

IA adversaria: describe técnicas y métodos que explotan las vulnerabilidades en la forma en que funcionan los sistemas de IA, por ejemplo, mediante la introducción de entradas maliciosas para explotar su aspecto de aprendizaje automático y engañar al sistema para que produzca resultados incorrectos o no deseados. Estas técnicas se utilizan habitualmente en los ataques adversarios, pero no son un tipo distinto de sistema de IA.

Ataque adversario: Un intento de manipular un modelo de IA mediante la introducción de entradas especialmente diseñadas para hacer que el modelo produzca errores o resultados no deseados.

Interfaz de programación de aplicaciones (API): Un conjunto de herramientas y protocolos que permiten que diferentes sistemas de software se comuniquen e interactúen.

Inteligencia artificial (IA): Sistemas diseñados para realizar tareas que normalmente requieren inteligencia humana, como la toma de decisiones, la comprensión del lenguaje y el reconocimiento de patrones. Estos sistemas pueden operar con diferentes niveles de autonomía y adaptarse a su entorno o datos para mejorar el rendimiento.

Envenenamiento de datos: un tipo de ataque adversario en el que se introducen datos maliciosos en conjuntos de datos de entrenamiento para comprometer el rendimiento o el comportamiento del sistema de IA.

Explicabilidad: La capacidad de un sistema de IA para proporcionar información comprensible para los humanos sobre su proceso de toma de decisiones.

Barandillas: Restricciones o reglas predefinidas implementadas para controlar y limitar los resultados y comportamientos de un sistema de IA, lo que garantiza la seguridad, la fiabilidad y la alineación con las directrices éticas u operativas.

Ataque de inferencia: Un ataque a la privacidad en el que un adversario recupera información confidencial sobre los datos de entrenamiento, o los usuarios, mediante el análisis de los resultados de un modelo de IA.

Inversión del modelo: un ataque a la privacidad en el que un adversario infiere información confidencial sobre los datos de entrenamiento mediante el análisis de los resultados del modelo de IA.

Indicación: Una entrada proporcionada a un modelo de IA, a menudo en forma de texto, que dirige o guía su respuesta. Las indicaciones pueden incluir preguntas, instrucciones o contexto para el resultado deseado.

Evaluación de riesgos: proceso de identificación, análisis y mitigación de amenazas potenciales para la seguridad o la funcionalidad de un sistema de IA.

Higienización: El proceso de limpieza y validación de datos o entradas para eliminar errores, inconsistencias y contenido malicioso, garantizando la integridad y seguridad de los datos.

Mensaje del sistema: Una entrada predefinida o un conjunto de instrucciones proporcionadas para guiar el comportamiento de un modelo de IA, que a menudo se utilizan para definir su tono, reglas o contexto operativo.

Modelado de amenazas: Un proceso para identificar y abordar las posibles amenazas de seguridad para un sistema durante sus fases de diseño y desarrollo.

Entrenamiento: el proceso de enseñar a un modelo de IA a reconocer patrones, tomar decisiones o generar resultados exponiéndolo a datos etiquetados y ajustando sus parámetros para minimizar los errores.
