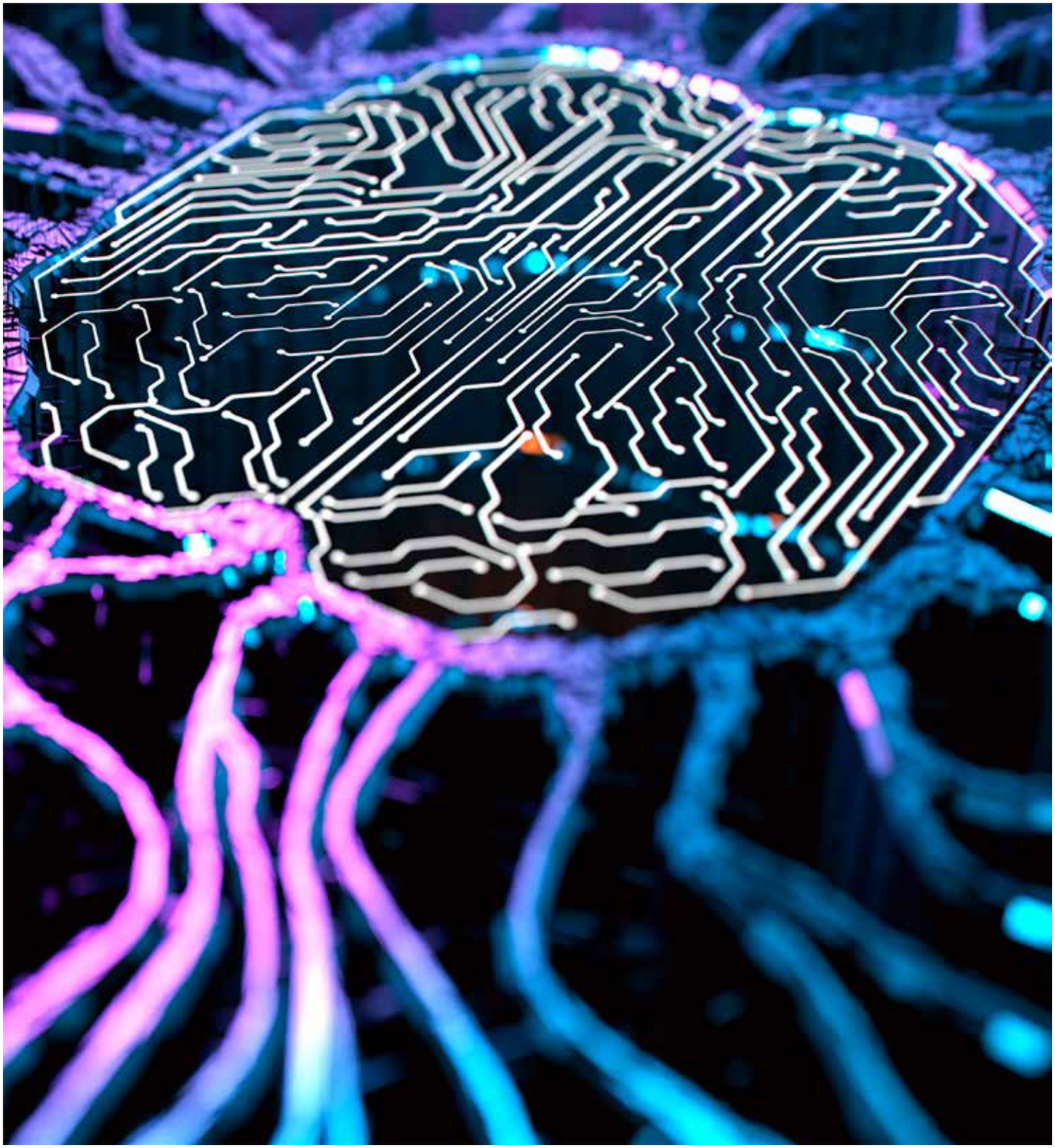
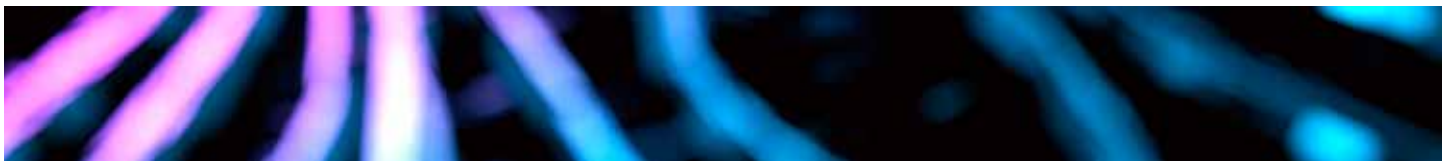
Directrices para el desarrollo de sistemas de IA seguros





National Cyber Security Centre
a part of GCHQ

CISA
CYBERSECURITY & INFRASTRUCTURE SECURITY AGENCY

Australian Government
Australian Signals Directorate

ASD AUSTRALIAN SIGNALS DIRECTORATE
ACSC Australian Cyber Security Centre

Communications Security Establishment
Canadian Centre for Cyber Security
Centre de la sécurité des télécommunications
Centre canadien pour la cybersécurité

CSIRT
Equipo de Respuesta ante Incidentes de Seguridad Informática

National Cyber and Information Security Agency

Republic of Estonia
Information System Authority

Department of Justice
Federal Bureau of Investigation

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

Federal Office for Information Security

INCD
Israel National Cyber Directorate

AGENZIA PER LA CYBERSICUREZZA NAZIONALE
ACN
SAPIENTIA ADIUVAT

NISC
内閣サイバーセキュリティセンター
National center of Incident readiness and Strategy for Cybersecurity

National Cyber Security Centre

NITDA

NSM
NORWEGIAN NATIONAL CYBER SECURITY CENTRE

National Security Agency
United States of America

NASK

국가정보원
National Intelligence Service

Ministerstwo Cyfryzacji

CSA SINGAPORE
Cyber Security Agency of Singapore

Acerca de este documento

Este documento es una publicación del UK National Cyber Security Centre, NCSC (Centro de Ciberseguridad del Reino Unido), la US Cybersecurity and Infrastructure Security Agency, CISA (Agencia de ciberseguridad y seguridad de la infraestructura de EE.UU.) y los socios internacionales siguientes:

- National Security Agency, NSA (Agencia Nacional de Seguridad)
- Federal Bureau of Investigations, FBI (Oficina Federal de Investigaciones)
- Australian Cyber Security Centre, ACSC (Centro Australiano de Ciberseguridad) de la Australian Signals Directorate (Dirección Australiana de Señales)
- Canadian Centre for Cyber Security, CCCS (Centro de Ciberseguridad de Canadá)
- New Zealand National Cyber Security Centre, NCSC-NZ (Centro Nacional de Ciberseguridad de Nueva Zelandia)
- Equipo de Respuesta ante Incidentes de Seguridad Informática (CSIRT) del Gobierno de Chile
- Agencia Nacional de Ciberseguridad y Seguridad de la Información de la República Checa (NUKIB)
- Autoridad del Sistema de Información de Estonia (RIA) y Centro Nacional de Ciberseguridad de Estonia (NCSC-EE)
- Agencia de Ciberseguridad de Francia (ANSSI)
- Oficina Federal de Seguridad de la Información de Alemania (BSI)
- Dirección Nacional de Cibernética de Israel (INCD)
- Agencia Nacional de Ciberseguridad de Italia (ACN)
- Centro Nacional de Preparación para Incidentes y Estrategia de Ciberseguridad de Japón (NISC)
- Secretaría de Ciencias, Tecnología e Innovación de Japón, Oficina del Gabinete
- National Information Technology Development Agency, NITDA (Agencia Nacional de Desarrollo de Tecnología de la Información de Nigeria)
- Centro Nacional de Ciberseguridad de Noruega (NCSC-NO)
- Ministerio de Asuntos Digitales de Polonia
- Instituto Nacional de Investigación de Polonia (NASK)
- Servicio Nacional de Inteligencia de la República de Corea (NIS)
- Cyber Security Agency of Singapore, CSA (Agencia de Ciberseguridad de Singapur)

Agradecimientos

Las siguientes organizaciones contribuyeron a la formulación de estas directrices:

- Alan Turing Institute (Instituto Alan Turing)
- Anthropic
- Databricks
- Center for Security and Emerging Technology, Georgetown University (Centro de Tecnología Emergente y de Seguridad de la Universidad de Georgetown)
- Google
- Google DeepMind
- IBM
- ImBue
- Microsoft
- OpenAI
- Palantir
- RAND
- Scale AI
- Software Engineering Institute at Carnegie Mellon University (Instituto de Ingeniería informática de la Universidad Carnegie Mellon)
- Stanford Center for AI Safety (Centro de Stanford de Seguridad de la IA)
- Stanford Program on Geopolitics, Technology and Governance (Programa de Geopolítica, Tecnología y Gobernanza de Stanford)

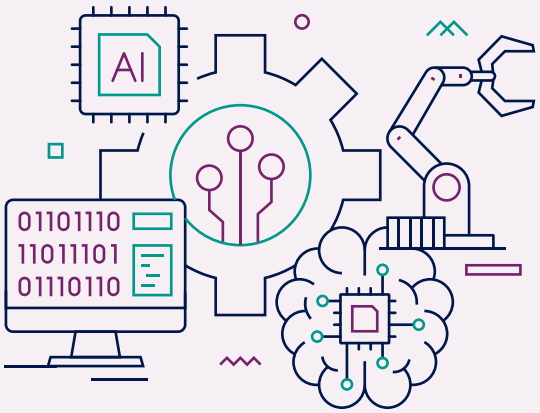
Descargo de responsabilidad

La información de este documento se ofrece “tal como está” y el NCSC y las organizaciones redactoras no aceptan responsabilidad alguna por pérdidas, lesiones o daños de cualquier tipo causados por su utilización, con excepción de lo dispuesto por la legislación. La información contenida en este documento no constituye ni implica endosos o recomendaciones algunos por parte del NCSC y los entes redactores con respecto a terceras organizaciones, productos o servicios. Los enlaces y referencias a páginas web y material de terceros se suministran para fines de información únicamente y no constituyen endosos o recomendaciones de dichos recursos más que otros.

Este documento se ofrece como TLP:CLEAR (https://www.first.org/tlp/).

Índice

Resumen 5
Introducción 6
Por qué es diferente la seguridad de la IA 6
Quiénes deberían leer este documento 7
Quién tiene la responsabilidad de desarrollar IA segura 7
Directrices para el desarrollo de sistemas de IA seguros 8
1. Diseño seguro 9
2. Desarrollo seguro 12
3. Distribución segura 14
4. Operación y mantenimiento seguros 16
Más lectura 17



Resumen

Este documento recomienda directrices para los proveedores de sistemas que usan inteligencia artificial (IA), independientemente de si dichos sistemas fueron creados desde el principio o construidos a partir de herramientas y servicios suministrados por otros. La implementación de dichas directrices ayudará a los proveedores a construir sistemas de IA que funcionen de la manera prevista, estén disponibles cuando se los necesite y funcionen sin revelar datos confidenciales a partes no autorizadas.

Este documento está dirigido en primer lugar a los proveedores de sistemas de IA que están usando modelos albergados por una organización, o que están usando interfaces de programación de aplicaciones (IPA) externas. Instamos a **todos** los interesados (incluidos los científicos de datos, diseñadores, administradores, responsables de decisiones y propietarios de riesgo) a que lean estas directrices que les facilitarán la toma de decisiones informadas sobre el **diseño**, **desarrollo**, **distribución** y **operación** de sus sistemas de IA.

Acerca de las directrices

Los sistemas de IA tienen el potencial de aportar muchos beneficios a la sociedad. No obstante, para que se concreten plenamente las oportunidades que ofrece la IA, su desarrollo, distribución y operación deben efectuarse de una manera segura y responsable.

Los sistemas de IA están sujetos a nuevas vulnerabilidades de seguridad que requieren la misma consideración que los peligros típicos que afectan a la ciberseguridad. Cuando el ritmo de desarrollo es acelerado, como es el caso de la IA, la seguridad suele quedar en segundo plano. La seguridad debe ser un requisito central, no solo en la fase de desarrollo, sino a lo largo de todo el ciclo de vida del sistema.

Es por este motivo que presentamos las directrices desglosadas en cuatro áreas clave del ciclo de vida del desarrollo del sistema de IA: **diseño seguro**, **desarrollo seguro**, **distribución segura** y **operación y mantenimiento seguros**. Cada sección ofrece consideraciones y mitigaciones que contribuirán a reducir el riesgo general para el proceso de desarrollo del sistema de IA de la organización.

1. **Diseño seguro**
   Esta sección contiene directrices para la etapa de diseño del ciclo de vida de desarrollo del sistema de IA. Abarca la comprensión del modelado de los riesgos y peligros, como así también temas específicos y compensaciones que se han de tomar en cuenta al diseñar el sistema y modelo.

2. **Desarrollo seguro**
   Esta sección contiene directrices para la etapa de desarrollo del ciclo de vida de desarrollo del sistema de IA, incluidas la seguridad de la cadena de suministro, documentación y gestión de activos y de la deuda técnica.

3. **Distribución segura**
   Esta sección contiene directrices para la etapa de distribución del ciclo de vida de desarrollo del sistema de IA, incluida la protección de la infraestructura y los modelos contra los ataques, peligros o pérdidas, la formulación de procesos de gestión de incidentes y la publicación responsable.

4. **Operación y mantenimiento seguros**
   Esta sección contiene directrices para la etapa de operación y mantenimiento seguros del ciclo de vida de desarrollo del sistema de IA. Ofrece pautas para las medidas especialmente pertinentes una vez que el sistema se ha distribuido, lo que incluye la conexión y el control, la gestión de las actualizaciones y el intercambio de información.

Las directrices se basan en un enfoque de "seguridad por defecto", y guardan estrecha congruencia con las prácticas definidas en la orientación del NCSC Secure development and deployment guidance, la orientación del NIST Secure Software Development Framework y los principios "secure by design principles" publicados por CISA, el NCSC y agencias internacionales de cibernética. Priorizan los siguientes:

- asumir responsabilidad por los resultados de seguridad para los clientes
- adoptar la transparencia y responsabilidad radicales
- establecer estructura y liderazgo organizacionales tan seguros por diseño es una prioridad comercial primordial

Introducción

Los sistemas de inteligencia artificial (IA) tienen el potencial de traerle muchos beneficios a la sociedad. No obstante, para que se concreten plenamente las oportunidades de la IA, el desarrollo, distribución y operación de dichos sistemas deben ser seguros y responsables. La ciberseguridad es una condición previa necesaria para la seguridad, resiliencia, privacidad, equidad, eficacia y confiabilidad de los sistemas de IA.

No obstante, los sistemas de IA están sujetos a nuevas vulnerabilidades de seguridad que requieren la misma consideración que los peligros típicos que afectan a la ciberseguridad. Cuando el ritmo de desarrollo es acelerado, como es el caso de la IA, la seguridad suele quedar en segundo plano. La seguridad debe ser un requisito central, no solo en la fase de desarrollo, sino a lo largo de todo el ciclo de vida del sistema.

Este documento recomienda directrices para los proveedores[1] de sistemas que usan inteligencia artificial (IA), independientemente de si dichos sistemas fueron creados desde el principio o construidos a partir de herramientas y servicios suministrados por otros. La implementación de dichas directrices ayudará a los proveedores a construir sistemas de IA que funcionen de la manera prevista, estén disponibles cuando se los necesite y funcionen sin revelar datos confidenciales a partes no autorizadas.

Estas directrices deberían leerse conjuntamente con las buenas prácticas establecidas de ciberseguridad, gestión del riesgo y respuesta frente a los incidentes. En particular, instamos a los proveedores a que apliquen los principios de "seguridad por defecto"[2] formulados por la US Cybersecurity and Infrastructure Security Agency (CISA), el UK National Cyber Security Centre (NCSC) y todos nuestros socios internacionales. Los principios priorizan los siguientes:

- asumir responsabilidad por los resultados de seguridad para los clientes
- adoptar la transparencia y responsabilidad radicales
- establecer estructura y liderazgo organizacionales tan seguros por diseño es una prioridad comercial primordial.

La aplicación de principios de "seguridad por defecto" exige recursos considerables a lo largo de todo el ciclo de vida del sistema. Significa que los diseñadores deben invertir en la priorización de **funciones**, **mecanismos** y **ejecución** de herramientas que protegen al cliente en todas las capas del diseño del sistema y en todas las etapas del ciclo de vida del desarrollo. Esto prevendrá costosos rediseños más adelante, y también protegerá a los clientes y sus datos en el futuro próximo.

¿Por qué es diferente la seguridad de la IA?

En este documento utilizamos "IA" para referirnos específicamente a las aplicaciones de aprendizaje automático (machine learning, ML)[3]. Esto abarca todos los tipos de aprendizaje automático. Las aplicaciones de aprendizaje automático son aquellas que:

- incluyen componentes de software (modelos) que permiten a las computadoras reconocer y contextualizar patrones en los datos sin que un ser humano tenga que programar las reglas explícitamente
- generan predicciones, recomendaciones o decisiones basadas en razonamiento estadístico

Además de estar sujetos a peligros de ciberseguridad existentes, los sistemas de IA están sujetos a nuevos tipos de vulnerabilidades. El término "aprendizaje automático adverso" describe la explotación de vulnerabilidades fundamentales en los componentes de aprendizaje automático, lo que incluye equipo, software, procesos de trabajo y cadenas de suministro. El aprendizaje automático adverso permite a los atacantes provocar conductas involuntarias en los sistemas de aprendizaje automático, por ejemplo:

- afectar el rendimiento de clasificación o regresión del modelo
- permitir a los usuarios hacer cosas no autorizadas
- extraer información confidencial del modelo

Estos efectos se pueden lograr de muchas maneras, por ejemplo mediante ataques de inyección inmediata en el dominio del modelo grande de lenguaje (large language model o LLM por sus siglas en inglés) o mediante la corrupción deliberada de los datos de entrenamiento o comentarios de los usuarios (conocido como "envenenamiento de datos").

¿Quiénes deberían leer este documento?

Este documento está dirigido en primer lugar a los proveedores de sistemas de IA que están usando modelos albergados por una organización, o que están usando interfaces de programación de aplicaciones (IPA) externas. No obstante, instamos a **todos** los interesados (incluidos los científicos de datos, diseñadores, administradores, responsables de decisiones y propietarios de riesgo) a que lean estas directrices que les facilitarán la toma de decisiones informadas sobre el **diseño**, **distribución** y **operación** de sus sistemas de IA de aprendizaje automático.

No obstante, no todas las directrices podrán aplicarse directamente en todas las organizaciones. El nivel de complejidad y las metodologías de ataque varían según el adversario que ataca al sistema de IA; por lo tanto, las directrices deben considerarse juntamente con los casos de uso y el perfil de peligro de la organización.

¿Quién tiene la responsabilidad de desarrollar IA segura?

Las cadenas de suministro de IA suelen tener muchos actores. Un enfoque simple supone dos entidades:

- el "proveedor" que es responsable por la gestión de los datos, y por el desarrollo, diseño, distribución y mantenimiento de los algoritmos
- el "usuario" que proporciona datos y recibe productos

Si bien este enfoque de proveedor y usuario se utiliza en muchas aplicaciones, se está tornando cada vez menos común[4], pues los proveedores pueden optar por incorporar en sus propios sistemas, software, datos, modelos y/o servicios a distancia proporcionados por terceros. Debido a la complejidad de estas cadenas de suministro, al usuario final se le hace más difícil comprender quién tiene la responsabilidad por la IA segura.

Normalmente, los usuarios (sean estos "usuarios finales" o proveedores que incorporan un componente de IA externo[5]) no tienen visibilidad y/o conocimientos y experiencia suficientes para comprender, evaluar o responder cabalmente a los riesgos asociados con los sistemas que están usando. Por ello, y de conformidad con los principios de "seguridad por defecto", **los proveedores de componentes de IA deberían asumir la responsabilidad por los resultados de seguridad de los usuarios que se encuentran más adelante en la cadena de suministro.**

Cuando ello sea posible, los proveedores deberían implementar controles y mitigaciones de seguridad en sus modelos, procesos y/o sistemas y, cuando se utilicen configuraciones, deberían implementar por defecto la opción más segura. Cuando no sea posible mitigar los riesgos, el proveedor deberá asumir responsabilidad por los siguientes:

- informar a los usuarios siguientes en la cadena de suministro sobre los riesgos que están aceptando ellos y (si corresponde) sus propios usuarios
- informarles de cómo usar el componente de manera segura

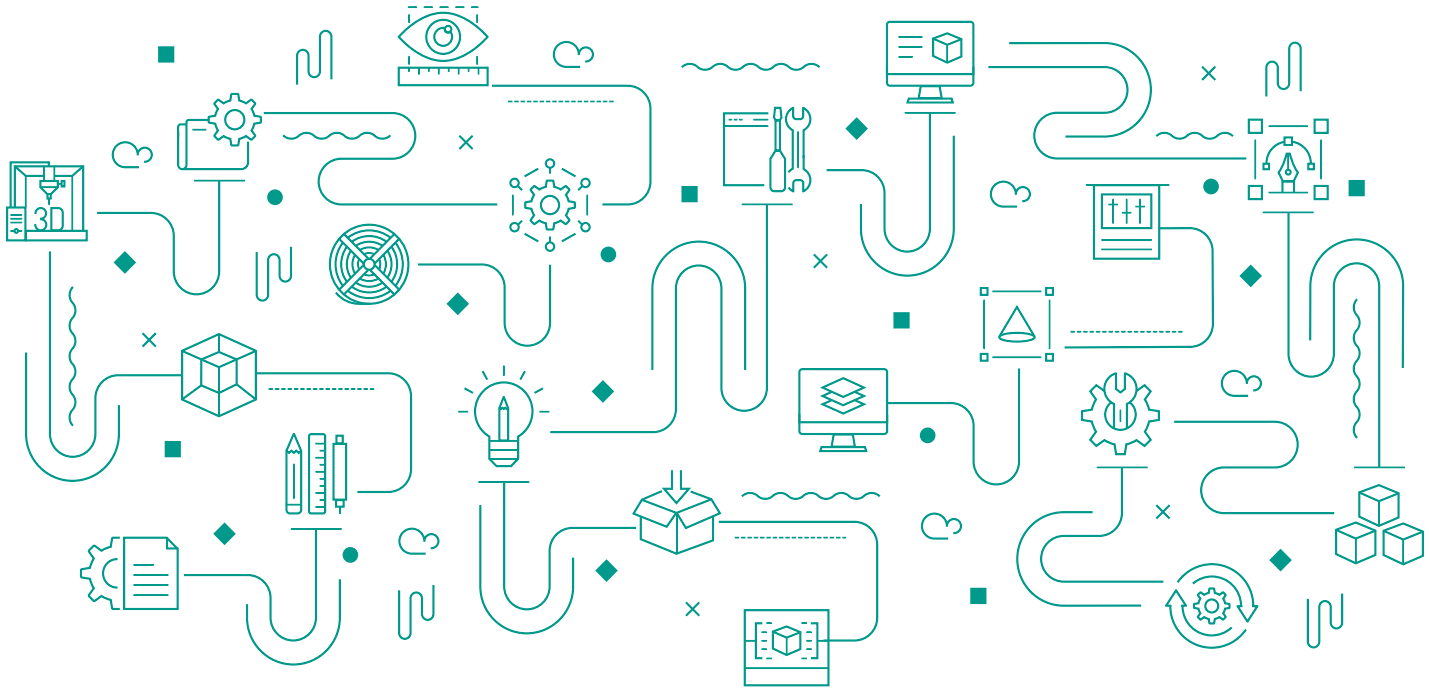
Cuando los ataques al sistema podrían causar daños físicos o de reputación tangibles o amplios, pérdida significativa de operaciones comerciales, fuga de información delicada o confidencial y/o repercusiones jurídicas, los riesgos de ciberseguridad de la IA deberán tratarse como riesgos de importancia **crítica.**

Directrices para el desarrollo de sistemas de IA seguros

Las directrices se desglosan en cuatro áreas clave del ciclo de vida de desarrollo del sistema de IA: **diseño seguro**, **desarrollo seguro**, **distribución segura** y **operación y mantenimiento seguros**. Ofrecemos para cada área consideraciones y mitigaciones que contribuirán a reducir el riesgo general al proceso de desarrollo del sistema de IA de la organización.

Las directrices presentadas en este documento guardan estrecha congruencia con las prácticas del ciclo de vida de desarrollo de software definidas en los siguientes:

- la orientación del NCSC Secure development and deployment guidance
- el marco del National Institute of Standards and Technology, NIST (Instituto Nacional de Normas y Tecnología) Secure Software Development Framework (SSDF)[6]



1. Diseño seguro

Esta sección contiene directrices para la etapa de **diseño** del ciclo de vida de desarrollo del sistema de IA. Abarca la comprensión del modelado de los riesgos y peligros, como así también temas específicos y compensaciones que se han de tomar en cuenta al diseñar el sistema y modelo.

Concientice al personal con respecto a los peligros y riesgos

Los propietarios y líderes superiores del sistema comprenden los peligros para la seguridad de la IA y sus mitigaciones. Sus científicos de datos y desarrolladores se mantienen conscientes de los peligros pertinentes para la seguridad y las modalidades de falla, y ayudan a los propietarios del riesgo a tomar decisiones informadas. Usted proporciona orientación a los usuarios sobre los riesgos únicos para la seguridad de los sistemas de IA (por ejemplo, como parte de la formación estándar InfoSec) y capacita a los desarrolladores en técnicas de cifrado seguro y prácticas de IA seguras y responsables.

Modele los peligros para su sistema

Como parte de su proceso de gestión del riesgo, usted aplica un proceso integral para evaluar los peligros para su sistema, lo que incluye comprender los eventuales efectos para el sistema, los usuarios, organizaciones y la sociedad general si se ataca un componente de IA o se comporta de manera inesperada[7]. Este proceso implica la evaluación del impacto de los peligros específicos para la IA[8] y la documentación de sus decisiones.

Reconoce que la confidencialidad y los tipos de datos usados en su sistema pueden influenciar el valor que pueda tener como objetivo de un atacante. Su evaluación debería tomar en cuenta que algunos peligros pueden aumentar a medida que los sistemas de IA cobran más interés como objetivo de gran valor y a medida que la IA propiamente dicha permite nuevos vectores de ataque automatizados.

Diseñe su sistema pensando en la seguridad como así también la funcionalidad y el rendimiento

Usted está seguro/a de que es más apropiado realizar la tarea en cuestión utilizando IA. Una vez determinado esto, puede evaluar la pertinencia de sus selecciones de diseño de IA. Toma en consideración su modelo de peligro y las mitigaciones de seguridad correspondientes, además de las funciones, experiencia del usuario, entorno de distribución, rendimiento, garantía, supervisión, requisitos éticos y legales, y otras consideraciones. Por ejemplo:

- toma en consideración la seguridad de la cadena de suministros cuando decide si va a desarrollar componentes en su organización o usar componentes externos, por ejemplo:
    - su opción de entrenar un modelo nuevo, usar un modelo existente (con o sin adaptación) o acceder a un modelo por medio de una aplicación externa es apropiada para sus necesidades
    - su opción de colaborar con el proveedor de un modelo externo incluye una evaluación de debida diligencia de la actitud de dicho proveedor con respecto a la seguridad
    - si utiliza una biblioteca externa, realiza una evaluación de debida diligencia (por ejemplo, para asegurarse de que la biblioteca tiene controles que impiden que el sistema cargue modelos no confiables sin exponerse inmediatamente a la ejecución de códigos arbitrarios[9])
    - implementa el escaneado y aislamiento/espacios reservados cuando importa modelos o valores serializados de terceros que deberían tratarse como código de terceros no confiable y podrían permitir la ejecución de códigos a distancia

  - si usa un interfaz de programación de aplicaciones, aplica controles apropiados a los datos que pueden ser enviados a servicios ajenos al control de su organización, como exigir que los usuarios se conecten y confirmen antes de enviar información eventualmente confidencial
  - aplica controles apropiados y saneamiento de datos y entradas, incluso cuando incorpora comentarios de usuarios o datos de aprendizaje constante a su modelo, reconociendo que los datos de entrenamiento definen la conducta del sistema

- incorpora el desarrollo del sistema de IA en las buenas prácticas existentes de seguridad en el desarrollo y operación; todos los elementos del sistema de IA están redactados en entornos apropiados, con prácticas de cifrado y lenguajes que reducen o eliminan las clases conocidas de vulnerabilidades, cuando es razonable

- si los componentes de IA deben disparar acciones, por ejemplo enmendar archivos o dirigir productos a sistemas externos, aplica restricciones apropiadas a las posibles acciones (ello incluye sistemas externos de IA y no IA a prueba de fallas si fuera necesario)

- las decisiones respecto a la interacción con el usuario se basan en riesgos específicos de la IA, por ejemplo:

  - su sistema ofrece a los usuarios productos utilizables sin revelar niveles innecesarios de detalle a un eventual atacante
  - si fuera necesario, su sistema rodea a los productos del modelo con protecciones eficaces
  - si ofrece un interfaz de programación de aplicaciones a clientes o colaboradores externos, aplica controles apropiados que mitigan los ataques al sistema de IA por medio de dicho interfaz
  - incorpora las selecciones más seguras en el sistema por defecto
  - aplica los principios de privilegio mínimo para limitar el acceso a las funciones de un sistema
  - explica las capacidades más riesgosas a los usuarios y les exige que, para usarlas, opten específicamente por aceptarlas; comunica los casos de uso prohibido y, cuando es posible, informa a los usuarios de soluciones alternativas

Considere los beneficios de la seguridad y las compensaciones cuando seleccione su modelo de IA

Su selección de modelo de IA incluirá la búsqueda del equilibrio entre una serie de requisitos. Ello incluye la selección de la arquitectura, la configuración, los datos de formación, el algoritmo de formación y los hiperparámetros del modelo. Sus decisiones están basadas en su modelo de peligro, y las vuelve a evaluar con regularidad a medida que avanzan los estudios de seguridad de la IA y que evoluciona la comprensión del peligro.

Cuando elige un modelo de IA, sus consideraciones incluyen probablemente las siguientes, sin limitaciones:

- la complejidad del modelo que utiliza, es decir la arquitectura y el número de parámetros elegidos; entre otros factores, la arquitectura y el número de parámetros que haya elegido para su modelo afectarán cuántos datos de entrenamiento necesita y cuán resistente será el modelo frente a los cambios en los datos introducidos cuando esté en uso

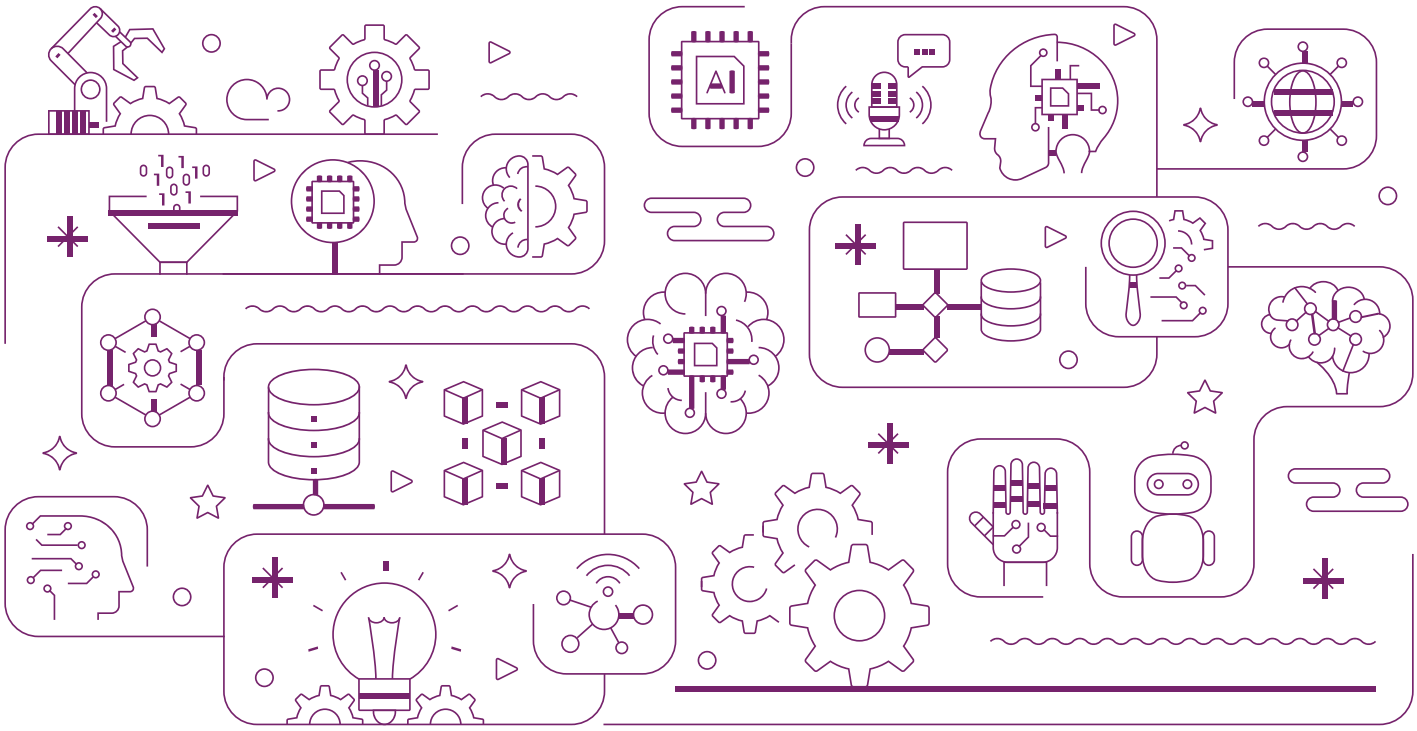
- la pertinencia del modelo para su caso y/o la viabilidad de adaptarlo a sus necesidades específicas (por ejemplo, ajustándolo)

- la capacidad de alinear, interpretar y explicar los productos de su modelo (por ejemplo, para depuración, auditorías o cumplimiento regulatorio); puede ser ventajoso usar modelos más sencillos y transparentes en lugar de modelos más grandes y complejos, que son más difíciles de interpretar

- características de la(s) serie(s) de datos de entrenamiento, por ejemplo tamaño, integridad, calidad, sensibilidad, edad, pertinencia y diversidad

- el beneficio de usar técnicas de endurecimiento del modelo (como el entrenamiento adversario), regularización y/o aumento de la privacidad
- la procedencia y las cadenas de suministro de los componentes, incluido el modelo o modelo básico, los datos de entrenamiento y las herramientas asociadas

Para obtener más información acerca del número de estos factores que repercuten en los resultados de seguridad, vea los Principios para la seguridad del aprendizaje automático del NCSC, y especialmente Design for security (model architecture) (Diseño con miras a la seguridad (arquitectura de modelos)).



2. Desarrollo seguro

Esta sección contiene directrices para la etapa de **desarrollo** del ciclo de vida de desarrollo del sistema de IA, lo que incluye la seguridad de la cadena de suministro, documentación y la gestión de activos y de la deuda técnica.

Proteja su cadena de suministros

Usted evalúa y controla la seguridad de sus cadenas de suministro de IA a lo largo de todo el ciclo de vida del sistema, y exige a los proveedores que cumplan las mismas normas que aplica su propia organización a otro software. Si los proveedores no pueden cumplir las normas de su organización, usted obra de conformidad con sus propias normativas de gestión del riesgo existentes.

Cuando no se producen en la organización, adquiere y mantiene componentes de equipo y software bien protegidos y documentados (por ejemplo, modelos, datos, bibliotecas de software, módulos, middleware, marcos e interfaces de programación de aplicaciones externos) de desarrolladores comerciales verificados, de fuente abierta y otros terceros para garantizar la seguridad robusta de sus sistemas.

Si no se satisfacen los criterios de seguridad, está dispuesto/a a optar por soluciones alternativas para los sistemas de importancia crítica para la misión. Utiliza recursos como Supply Chain Guidance del NCSC y marcos como Supply Chain Levels for Software Artifacts (SLSA)[10] para rastrear los certificados de los ciclos de vida de la cadena de suministros y de desarrollo del software.

Defina, rastree y proteja sus activos

Usted comprende el valor que tienen para su organización sus activos de IA, como los modelos, datos (incluidos los comentarios de los usuarios), instrucciones, software, documentación, registros y evaluaciones (incluida la información sobre capacidades eventualmente inseguras y modalidades de falla), y reconoce cuándo representan inversiones importantes y cuándo el acceso a ellos habilita a un atacante. Trata los registros como datos confidenciales e implementa controles para proteger su confidencialidad, integridad y disponibilidad.

Sabe dónde residen sus activos y ha evaluado y aceptado todo riesgo pertinente. Tiene procesos y herramientas para rastrear, autenticar, controlar la versión y proteger sus activos, y puede restablecer las funciones a una buena condición conocida en caso de ataque.

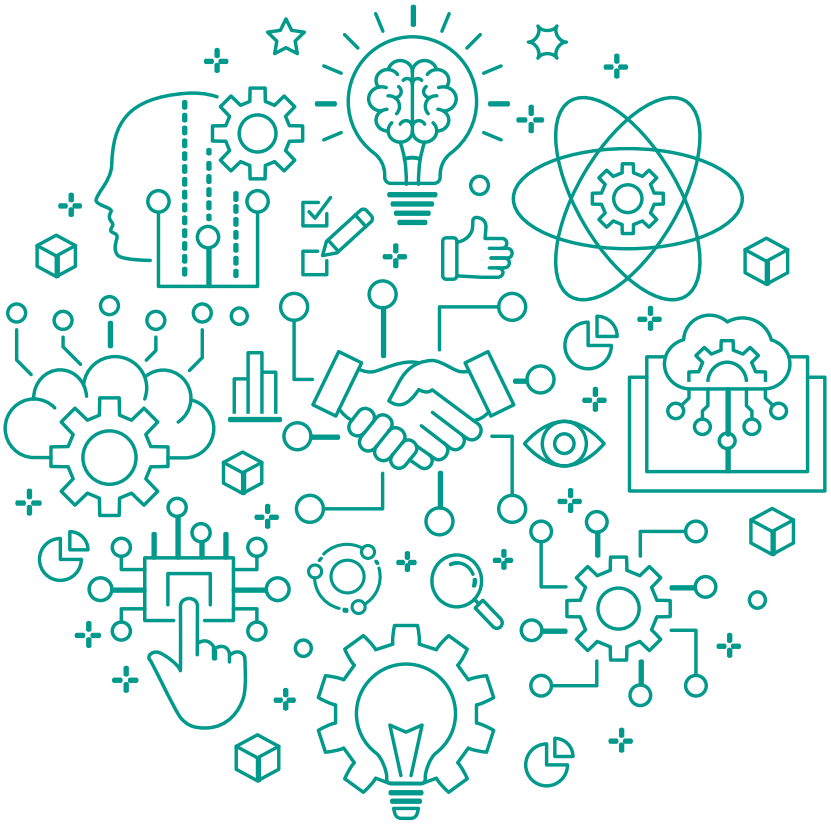
Tiene procesos y controles para gestionar a qué datos tienen acceso los sistemas de IA, y para gestionar el contenido generado por IA según su nivel de confidencialidad (y la confidencialidad de las entradas que permitieron generarlo).

Documente sus datos, modelos e instrucciones

Usted documenta la creación, operación y gestión del ciclo de vida de todo modelo, serie de datos e instrucciones meta o de sistema. Su documentación incluye información pertinente a la seguridad, como las fuentes de datos de entrenamiento (incluidos los datos ajustados y la retroalimentación operativa humana y otras), el ámbito de aplicación esperado y las limitaciones, protecciones, hashes criptográficos o firmas, tiempo de retención, frecuencia de revisión sugerida y eventuales modalidades de falla. Entre otras, las estructuras útiles que le ayudarán a lograrlo son las tarjetas de modelo, tarjetas de datos y listas de materiales de software. La producción de documentación exhaustiva respalda la transparencia y gestión responsable[11].

Gestione su deuda técnica

Al igual que todo sistema de software, usted define, rastrea y gestiona su "deuda técnica" a lo largo del ciclo de vida del sistema de IA (la deuda técnica se define como las decisiones de ingeniería que, por alcanzar resultados a corto plazo, no aplican la buena práctica a costas de los beneficios a largo plazo). Al igual que las deudas financieras, las deudas técnicas no son algo negativo de por sí, pero es preciso gestionarlas desde las primeras etapas del desarrollo[12]. Usted reconoce que esto puede resultar más complejo en el contexto de IA que con el software estándar, y que probablemente su nivel de deuda técnica sea alto debido a los rápidos ciclos de desarrollo y a la falta de protocolos e interfaces bien establecidos. Se asegura de que sus planes de ciclo de vida (incluidos los procesos para desmantelar sistemas de IA) evalúen, reconozcan y mitiguen los riesgos para sistemas similares futuros.



3. Distribución segura

Esta sección contiene directrices para la etapa de **distribución** del ciclo de vida de desarrollo del sistema de IA, incluida la protección de la infraestructura y los modelos contra los ataques, peligros o pérdidas, la formulación de procesos de gestión de incidentes y la publicación responsable.

Proteja su infraestructura

Usted aplica buenos principios de seguridad de la infraestructura a la infraestructura utilizada en todas las partes del ciclo de vida de su sistema. Aplica controles de acceso apropiados a sus interfaces de programación de aplicaciones, modelos y datos, y a sus procesos de entrenamiento y transformación, en investigación y desarrollo como así también en la distribución. Ello incluye la segregación apropiada de los entornos que contienen código o datos confidenciales. Esto también ayuda a mitigar los ataques comunes a la ciberseguridad, que intentan robar un modelo o dañar su rendimiento.

Proteja su modelo continuamente

Es posible que los atacantes puedan reconstruir las funciones de un modelo[13] o los datos que se utilizaron para entrenarlo[14] si acceden al modelo directamente (adquiriendo valores del modelo) o indirectamente (cuestionando el modelo por medio de una aplicación o un servicio). También es posible que los atacantes manipulen los modelos, datos o instrucciones durante o después del entrenamiento, causando así la falta de confiabilidad de los productos.

Usted protege el modelo y los datos contra el acceso directo e indirecto, respectivamente, mediante las siguientes acciones:

- implementa buenas prácticas normalizadas de ciberseguridad
- implementa controles del interfaz de búsquedas para detectar y prevenir los intentos de acceso, modificación y exfiltración de información confidencial

Para asegurarse de que los sistemas de consumo puedan validar modelos, calcula y comparte hashes criptográficos y/o firmas de los archivos del modelo (por ejemplo, valores del modelo) y series de datos (incluidos los puntos de control) tan pronto como el modelo ha sido entrenado. Como siempre sucede con la criptografía, es esencial la buena gestión de las claves[15].

El enfoque con respecto a la mitigación de los riesgos para la confidencialidad dependerá mucho de los casos de uso y del modelo de peligro. Algunas aplicaciones, por ejemplo las que tratan con datos muy confidenciales, pueden requerir garantías teóricas cuya aplicación puede ser difícil o costosa. En el caso apropiado se pueden utilizar tecnologías de aumento de la privacidad (como la privacidad diferencial o el cifrado homomórfico) para explorar o garantizar niveles de riesgo asociados con el acceso de los consumidores, usuarios y atacantes a los modelos y productos.

Formule procedimientos para la gestión de incidentes

La inevitabilidad de los incidentes de seguridad que afecten sus sistemas de IA se ve reflejada en sus planes de respuesta a incidentes, intensificación y corrección. Sus planes toman en cuenta diferentes casos teóricos y se reevalúan regularmente a medida que evolucionan el sistema y la investigación general. Usted guarda los recursos digitales de importancia crítica de su compañía en copias de respaldo fuera de línea. Los intervinientes han recibido capacitación para evaluar y abordar los incidentes relacionados con la IA. Proporciona registros de auditoría de buena calidad y otras funciones o información de seguridad a los clientes y usuarios sin costo adicional, para facilitar sus procesos de respuesta a incidentes.

Distribuya IA de manera responsable

Usted distribuye modelos, aplicaciones o sistemas solo después de haberlos sometido a evaluaciones apropiadas y eficaces de la seguridad, como evaluación comparativa y uso de equipos rojos (red teaming) (como así también otras pruebas que no caen bajo el ámbito de estas directrices, como la seguridad personal o la equidad), y manifiesta claramente a sus usuarios las limitaciones conocidas o las eventuales modalidades de falla. Encontrará los datos de bibliotecas de prueba de seguridad de fuente abierta en la sección de más lectura al final de este documento.

Ayude al usuario a hacer lo correcto

Usted reconoce que cada nueva selección u opción de configuración debe ser evaluada junto con el beneficio que le trae a la empresa y todo riesgo que introduce para la seguridad. En la situación ideal, el ajuste más seguro se incorpora en el sistema como única opción. Cuando es preciso configurar, la opción por defecto debería tener un amplio grado de seguridad contra los peligros comunes (es decir, seguridad por defecto). Aplica controles para impedir el uso o distribución de su sistema de maneras maliciosas.

Proporciona orientación a los usuarios sobre el uso apropiado de su modelo o sistema, lo que incluye señalar las limitaciones y eventuales modalidades de falla. Declara claramente a los usuarios qué aspectos de la seguridad son responsabilidad de ellos, y señala de manera transparente cuándo (y cómo) podrían usarse, accederse o almacenarse sus datos (por ejemplo, si se utilizan para reentrenar el modelo, o ser examinados por empleados o socios).

4. Operación y mantenimiento seguros

Esta sección contiene directrices para la etapa de **operación y mantenimiento seguros** del ciclo de vida de desarrollo del sistema de IA. Ofrece pautas para las medidas especialmente pertinentes una vez que el sistema se ha distribuido, lo que incluye la conexión y el control, la gestión de las actualizaciones y el intercambio de información.

Controle la conducta de su sistema

Usted mide los productos y rendimiento de su modelo y sistema para poder detectar cambios repentinos y graduales en la conducta que afecta la seguridad. También puede identificar y dar cuenta de eventuales intrusiones y ataques, como así también de la desviación natural de datos.

Controle las entradas a su sistema

En consonancia con los requisitos de protección de la privacidad y los datos, usted controla y registra las entradas a su sistema (como las peticiones de inferencia, las consultas o instrucciones) para facilitar las obligaciones de cumplimiento, auditorías, investigación y corrección en caso de ataque o uso no autorizado. Esto podría incluir la detección explícita de entradas fuera de distribución y/o adversarias, incluso aquellas que tienen por objeto explotar los pasos de preparación de los datos (por ejemplo, el recorte y cambio de tamaño de las imágenes).

Aplique a las actualizaciones un enfoque de diseño para la seguridad

Usted incluye actualizaciones automáticas por defecto en todos los productos y usa procedimientos de actualización seguros y modulares para distribuirlas. Sus procesos de actualización (incluidos los regímenes de prueba y evaluación) son un reflejo del hecho de que los cambios a los datos, modelos o instrucciones pueden llevar a cambios en la conducta del sistema (por ejemplo, usted trata las actualizaciones grandes como si fueran versiones nuevas). Presta apoyo a los usuarios para que evalúen y respondan a cambios del modelo (por ejemplo, otorgando acceso para vistas previas e interfaces de programación de aplicaciones en versiones).

Reúna y comparta las enseñanzas aprendidas

Usted participa en comunidades de intercambio de información, y colabora en todo el ecosistema global de la industria, el mundo académico y los gobiernos para compartir buenas prácticas según sea apropiado. Mantiene líneas de comunicación abiertas para fines de retroalimentación sobre la seguridad de los sistemas, tanto en su organización como a nivel externo; esto incluye autorizar a los investigadores de la seguridad a que investiguen y notifiquen de las vulnerabilidades. Cuando ello sea necesario, lleva los problemas a la comunidad general, por ejemplo, publicando boletines para responder a revelaciones de vulnerabilidad, incluida la enumeración detallada y completa de las vulnerabilidades comunes. Toma medidas para mitigar y corregir los problemas rápidamente y de manera apropiada.

Más lectura

Desarrollo de IA

Principles for the security of machine learning (Principios para la seguridad del aprendizaje automático)
La orientación detallada del NCSC sobre el desarrollo, distribución y operación de un sistema con un componente de aprendizaje automático.

Secure by Design - Shifting the Balance of Cybersecurity Risk: Principles and Approaches for Secure by Design Software (Seguridad por defecto – Modificación del equilibrio del riesgo para la ciberseguridad: Principios y enfoques para software de diseño seguro por defecto)
Producida en colaboración por CISA, NCSC y otras agencias, esta orientación describe las medidas que deberían tomar los fabricantes de sistemas de software, incluida la IA, para tomar en consideración la seguridad en la etapa de diseño del desarrollo del producto, y vender sistemas que vengan con seguridad incorporada.

AI Security Concerns in a Nutshell (Las inquietudes de seguridad de la IA en pocas palabras)
Producido por la Oficina Federal de Seguridad de la Información de Alemania (BSI), este documento presenta una introducción a posibles ataques a los sistemas de aprendizaje automático y las eventuales defensas contra dichos ataques.

Hiroshima Process International Guiding Principles for Organizations Developing Advanced AI Systems (Principios rectores internacionales del proceso de IA de Hiroshima para sistemas avanzados de IA) y Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems (Código de conducta internacional para desarrolladores de sistemas avanzados de IA del proceso de IA de Hiroshima)
Estos documentos, producidos como parte del Proceso de IA de Hiroshima del G7, ofrecen orientación a las organizaciones desarrolladoras de sistemas de IA de vanguardia, incluidos los modelos de base y los sistemas generativos de IA más avanzados, con el objeto de promover IA segura y confiable en todo el mundo.

AI Verify
Marco y caja de herramientas de prueba de software de gobernanza de la IA de Singapur, que valida el desempeño de los sistemas de IA en comparación con una serie de principios reconocidos internacionalmente por medio de pruebas normalizadas.

Multilayer Framework for Good Cybersecurity Practices for AI — ENISA (europa.eu) (Marco multicapa de buenas prácticas de ciberseguridad para la IA de ENISA, la Agencia de la Unión Europea para la ciberseguridad)
Marco de orientación para las autoridades nacionales competentes e interesados en IA sobre los pasos necesarios para la seguridad de sus sistemas, operaciones y procesos de IA.

ISO 5338: Procesos del ciclo de vida de los sistemas de IA (en curso de estudio)
Una serie de procesos y conceptos asociados que describen el ciclo de vida de los sistemas de IA sobre la base del aprendizaje automático y los sistemas heurísticos.

AI Cloud Service Compliance Criteria Catalogue (AIC4)
El Catálogo de criterios de cumplimiento del servicio de la nube de IA de BSI presenta criterios específicos para la IA que permiten evaluar la seguridad de un servicio de IA a lo largo de su ciclo de vida.

NIST IR 8269 (Draft) A Taxonomy and Terminology of Adversarial Machine Learning (Proyecto de taxonomía y terminología del aprendizaje automático adversario)
Serie de procesos y conceptos asociados para describir el ciclo de vida de los sistemas de IA sobre la base del aprendizaje automático y los sistemas heurísticos.

MITRE ATLAS
Base de conocimientos de tácticas y técnicas adversarias, y estudios de caso para los sistemas de aprendizaje automático, modelados según el marco MITRE ATT&CK y vinculados a éste.

An Overview of Catastrophic AI Risks (2023) (Panorama de los riesgos catástroficos de IA, 2023)
Producido por el Centro de seguridad de la IA, este documento presenta las áreas de riesgo representadas por la IA.

Large Language Models: Opportunities and Risks for Industry and Authorities (Modelos grandes de lenguaje: oportunidades y riesgos para la industria y las autoridades)
Documento producido por BSI destinado a las empresas, autoridades y desarrolladores que desean informarse más sobre las oportunidades y los riesgos del desarrollo, distribución y/o uso de los modelos grandes de lenguaje.

Proyectos de fuentes gratuitas para ayudar a los usuarios a poner a prueba la seguridad de los modelos de IA:

- Adversarial Robustness Toolbox (IBM)
- CleverHans (Universidad de Toronto)
- TextAttack (Universidad de Virginia)
- Prompt Bench (Microsoft)
- Counterfit (Microsoft)
- AI Verify (Infocomm Media Development Authority, Singapur)

Ciberseguridad

CISA's Cybersecurity Performance Goals (Metas de rendimiento de ciberseguridad de CISA)
Una serie común de protecciones que deberían implementar todas las entidades de infraestructura crítica para lograr la reducción significativa de la probabilidad y el efecto de los riesgos conocidos y las técnicas adversarias.

NCSC CAF Framework
El Marco de evaluación cibernética (CAF por sus siglas en inglés) del NCSC ofrece orientación a las organizaciones con responsabilidad por servicios y actividades de importancia vital.

MITRE's Supply Chain Security Framework (Marco de seguridad de la cadena de suministro de MITRE)
Marco para la evaluación de los proveedores de productos y servicios en una cadena de suministro.

Gestión del riesgo

NIST AI Risk Management Framework (AI RMF) (Marco de gestión del riesgo de la IA de NIST)
El Marco de gestión del riesgo de la IA describe cómo gestionar los riesgos sociotécnicos para las personas físicas, organizaciones y la sociedad con una asociación única con la AI.

ISO 27001: Seguridad de la información, ciberseguridad y protección de la privacidad
Esta norma ofrece a las organizaciones orientación sobre el establecimiento, implementación y mantenimiento de un sistema de gestión de la seguridad de la información.

ISO 31000: Gestión del riesgo
Norma internacional que ofrece a las organizaciones directrices y principios para la gestión del riesgo en su organización.

NCSC Risk Management Guidance (Orientación sobre la gestión del riesgo del NCSC)
Esta orientación está destinada a los profesionales del riego de ciberseguridad y les facilitará la comprensión y gestión de los riesgos para la ciberseguridad que afectan a sus organizaciones.

Notas

1. Definido aquí como persona física, autoridad, agencia u otro organismo que desarrolla un sistema de IA (o que contrata el desarrollo de un sistema de IA) y lanza dicho sistema al mercado o lo pone en funcionamiento bajo su propio nombre o marca
2. Para obtener más información sobre "seguridad por defecto", véase la página web de CISA Secure by Design y la orientación Shifting the Balance of Cybersecurity Risk: Principles and Approaches for Secure by Design Software
3. En lugar de los enfoques de IA que no utilizan aprendizaje automático, como los sistemas basados en reglas
4. CEPS (Centro europeo de estudios sociales y políticos) describe siete tipos diferentes de interacción de desarrollo de IA en su publicación "Reconciling the AI Value Chain with the EU's Artificial Intelligence Act" (Conciliación de la cadena de valor de la IA con la Ley de inteligencia artificial de la UE)
5. ISO/IEC 22989:2022(en) lo define como un "elemento funcional que construye un sistema de IA"
6. NIST tiene por misión producir directrices (y tomar otras medidas) para fomentar el avance del desarrollo y uso seguros y confiables de la Inteligencia Artificial (IA). Véanse las Responsabilidades de NIST con arreglo a la Orden ejecutiva del 30 de octubre de 2023
7. Se puede obtener más información sobre el modelado de los peligros de la OWASP Foundation
8. Véase MITRE ATLAS Adversarial Machine Learning 101
9. GitHub: RCE PoC for Tensorflow using a malicious Lambda layer
10. SLSA: "Safeguarding artifact integrity across any software supply chain" (Protección de la integridad del artefacto a lo largo de cualquier cadena de suministro de software)
11. METI (Ministerio de economía, comercio e industria del Japón, 2023), "Guía de introducción a la lista de materiales de software (SBOM) para la gestión de software"
12. Búsqueda en Google: Machine Learning: The High Interest Credit Card of Technical Debt (Aprendizaje automático: la tarjeta de crédito de alto interés de la deuda técnica)
13. Tramèr et al 2016, Stealing Machine Learning Models via Prediction APIs (El robo de modelos de aprendizaje automático por medio de interfaces de programación de aplicaciones)
14. Boenisch, 2020, Attacks against Machine Learning Privacy (Part 1): Model Inversion Attacks with the IBM-ART Framework
15. Centro Nacional de Ciberseguridad, 2020, Design and build a privately hosted Public Key Infrastructure