AI Technology Research

Directrices para el desarrollo y la implementación seguros de sistemas de IA

kaspersky bring on the future

Agradecimientos

Kaspersky desarrolló este documento y lo presentó durante **el Taller n.° 31 "Ciberseguridad en IA: cómo equilibrar la innovación y los riesgos"** en la 19.ª reunión anual del **Foro de Gobernanza de Internet** (del 15 al 19 de diciembre de 2024).

El documento contó con el aporte de las siguientes personas:

- Yury Shelikhov, director de Ciberseguridad y Protección de Datos, Kaspersky
- Vladislav Tushkanov, director del Grupo de Desarrollo de la Investigación de Tecnología de Aprendizaje Automático, Kaspersky
- Alexey Antonov, responsable de Investigación de Datos, Grupo de Análisis de Métodos de Detección, Kaspersky
- Allison Wylde, miembro del equipo (interoperabilidad), Red de Políticas sobre Inteligencia Artificial (PNAI) del Foro de Gobernanza de Internet
- Dra. Melodena Stephens, profesora de Gobernanza de la Innovación y la Tecnología, Mohammed Bin Rashid School of Government, Emiratos Árabes Unidos
- Sergio Mayo Macías, director de Programas de Innovación, Instituto Tecnológico de Aragón (ITA), España
- Igor Kumagin, experto en ciberseguridad
- Dmitry Fonarev, gerente sénior de Asuntos Públicos, Kaspersky

Contenidos

Introducción 4

Objetivo 4

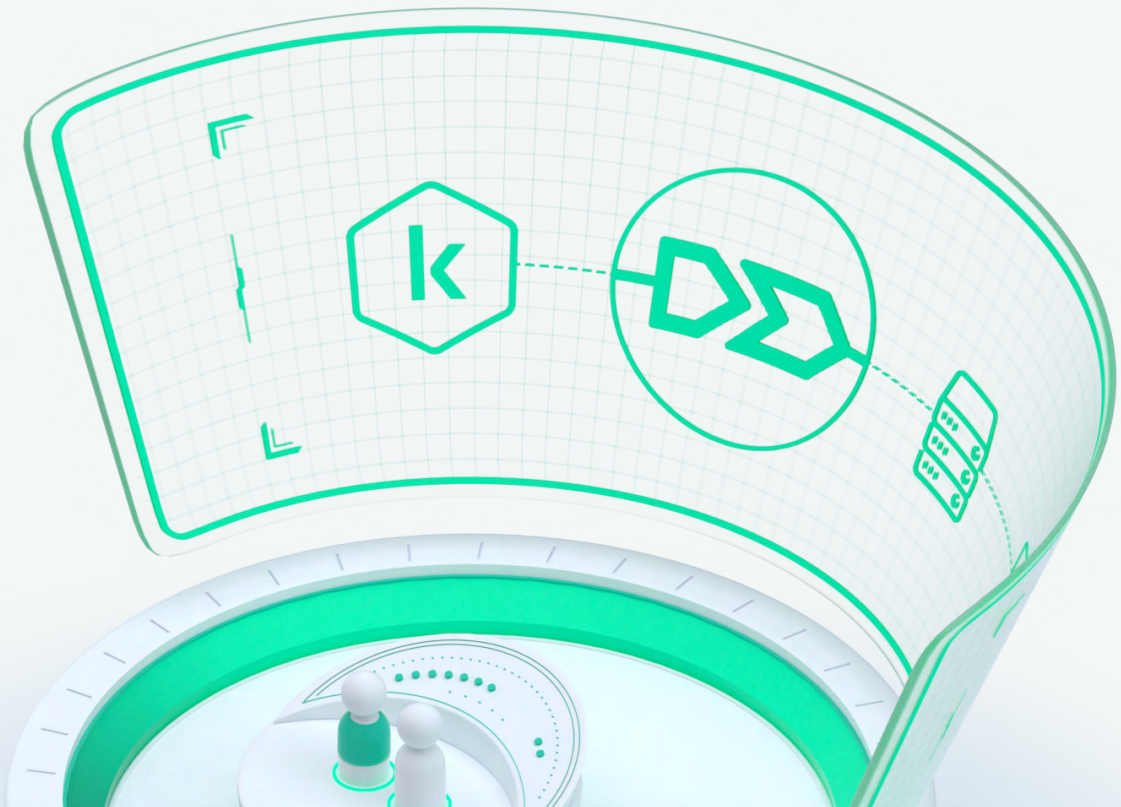
Panorama de las amenazas de la IA 5

Problemas con el desarrollo de modelos 5
Ataques a modelos de IA 6
Vulnerabilidades de seguridad tradicionales 7

Directrices 8

Concientización y capacitación en ciberseguridad 8
Modelado de amenazas/Evaluación de riesgos 9
Seguridad de la infraestructura (nube) 10
Cadenas de suministro y seguridad de datos 11
Pruebas y validaciones 12
Informes de vulnerabilidad 13
Defensa contra ataques a modelos de aprendizaje automático 14
Actualizaciones de seguridad y mantenimiento periódicos 15
Cumplimiento con normas internacionales 16

Conclusión 16



Introducción

Objetivo

Las nuevas tecnologías digitales vienen con **nuevos riesgos de ciberseguridad y vectores de ataque.** Por lo tanto, las empresas deben asegurarse de que la integración de la IA esté protegida de estas amenazas. El concepto de seguridad en el desarrollo de sistemas de IA pasó a primer plano en diversas iniciativas regulatorias, como la Ley de IA de la UE o el Marco de Gobernanza de IA para la IA Generativa en Singapur, a fin de minimizar los riesgos cibernéticos asociados. La UE está estableciendo regulaciones estrictas sobre IA con la Ley de IA, que tiene como objetivo garantizar la transparencia, la seguridad y los estándares éticos. Estados Unidos se centra en desarrollar estándares industriales y fomentar la innovación en lugar de imponer leyes estrictas. China está formulando normas y regulaciones que apoyan el desarrollo de tecnologías de IA, pero también limitan su uso en ciertas áreas.

A pesar de estos avances regulatorios, todavía persisten brechas importantes entre los marcos generales y su implementación práctica a un nivel más técnico. En esta guía, exploraremos los **requisitos básicos de ciberseguridad** que deben tenerse en cuenta en la implementación de sistemas de IA. Estos requisitos deberían aplicarse a una **gama más amplia de empresas que utilizan componentes de IA de terceros** para construir sus propias soluciones.

Para implementar la IA de forma segura, las organizaciones necesitan orientación técnica sobre el desarrollo y la implementación de la IA dentro de su infraestructura. Implementar IA sin la orientación adecuada puede conllevar riesgos importantes. En este documento se hace hincapié en proporcionar pautas a desarrolladores y administradores de sistemas de IA, MLOps y AI DevOps, que usan modelos fundacionales existentes para crear soluciones de IA generalizadas, prestando especial atención a los sistemas de IA basados en la nube. En el trabajo se abordan **aspectos clave del desarrollo, de la implementación y del funcionamiento de los sistemas de IA**, incluido el diseño, las mejores prácticas de seguridad y la integración, sin centrarse en el desarrollo de modelos fundacionales.

La inteligencia artificial (IA) se convirtió en una tecnología crítica para la economía global y se incorporó a la vida cotidiana. La IA permite a las organizaciones automatizar tareas rutinarias, mejorar el servicio al cliente y proporcionar a los empleados un acceso más rápido y sencillo a la información.

>50%

En un estudio reciente de Kaspersky se reveló que más del 50% de las empresas implementaron soluciones basadas en IA en sus infraestructuras*.

33%

están planeando adoptar esta tecnología dentro de dos años.

* Más de la mitad de las empresas utilizan IA e IoT en sus procesos comerciales, https://www.kaspersky.com/about/press-releases/more-than-half-of-companies-use-ai-and-iot-in-their-business-processes

Las amenazas a los sistemas de IA están aumentando a medida que esta tecnología se implementa cada vez más en las organizaciones. Los ciberataques afectan todas las etapas del desarrollo de la IA, desde los conjuntos de datos hasta los algoritmos y los resultados de los modelos.

Panorama de las amenazas a la IA

Según la investigación de Kaspersky*, los sistemas de IA enfrentan desafíos de seguridad únicos y cambiantes que ponen en riesgo su funcionamiento. Los actores maliciosos explotan vulnerabilidades en los datos de entrenamiento, manipulan modelos para cambiar su comportamiento y comprometen la integridad del sistema. Esto resalta la necesidad urgente de una seguridad integral en las aplicaciones de IA.

Problemas con el desarrollo de modelos

A diferencia de la programación tradicional, donde el comportamiento del código se puede entender y probar explícitamente, los modelos de aprendizaje automático, especialmente los modelos de aprendizaje profundo con millones o miles de millones de parámetros, son inherentemente complejos y a menudo funcionan como una "caja negra". Esta complejidad hace que sea difícil predecir e interpretar completamente el comportamiento de los modelos. Por lo tanto, el riesgo de que se produzcan errores no detectados con graves consecuencias, como la estabilidad financiera de un banco o incluso la vida de un paciente, aumenta significativamente.

Otro problema es el hecho de que los modelos de IA a veces pueden basar sus decisiones en **propiedades de datos de entrada irrelevantes o insignificantes**, en lugar de centrarse en características relevantes. Por ejemplo, los modelos de reconocimiento de imágenes podrían aprender a clasificar animales como guepardos, leopardos y jaguares centrándose únicamente en los patrones de sus manchas en lugar de utilizar su anatomía general**. En un caso, un modelo clasificó erróneamente un sofá manchado como un leopardo porque asoció el patrón de manchas con el animal. Estas clasificaciones erróneas pueden dar lugar a resultados erróneos en aplicaciones críticas en áreas como la atención médica, la educación, el bienestar social, el transporte, el sector gubernamental, etc.

Las inconsistencias entre los datos utilizados para el entrenamiento y los encontrados durante la implementación pueden generar un rendimiento deficiente del modelo. Por ejemplo, si un modelo se entrena con datos recopilados de un tipo de instrumento, pero se aplica a datos de un dispositivo diferente, puede aprender características específicas del dispositivo en lugar de los objetos subyacentes que pretende reconocer. Este desajuste puede dar lugar a predicciones o clasificaciones inexactas.

Estos desafíos pueden enfrentarse tanto al entrenar los modelos desde cero como al perfeccionar los fundacionales. Si bien los conjuntos de datos de ajuste fino son más pequeños y más fáciles de gestionar, también pueden introducir correlaciones espurias que hacen que el modelo resultante no se alinee con su objetivo previsto.

* La IA bajo ataque, https://content.kaspersky-labs.com/se/media/en/business-security/enterprise/machine-learning-cybersecurity-whitepaper.pdf

** De repente, aparece un sofá con estampado de leopardo, https://web.archive.org/web/20200208171948/http://rocknrollnerd.github.io/ml/2015/05/27/leopard-sofa.html

Ataques a los modelos de IA

Los actores maliciosos pueden atacar los modelos de IA a través de varios métodos. A continuación, se muestran ejemplos de cómo los atacantes explotan vulnerabilidades en el diseño, el entrenamiento y los mecanismos de interacción de la IA:

Maneras de atacar	Descripción
Envenenamiento de datos: riesgo a la integridad del modelo	El envenenamiento de datos implica que un atacante inyecte datos maliciosos en el conjunto de datos de entrenamiento para influir en el comportamiento del modelo. Al elaborar y agregar cuidadosamente muestras envenenadas*, los atacantes pueden hacer que el modelo tome decisiones erróneas o clasificaciones incorrectas en ciertas entradas. Este tipo de ataque puede poner en riesgo la integridad del modelo y socavar su confiabilidad. Esto también se aplica al perfeccionamiento de los modelos base.
Ataques adversarios: manipulación invisible de la IA	Los ataques adversarios implican modificaciones sutiles de los datos de entrada que hacen que el modelo de IA los clasifique erróneamente, mientras que los humanos no perciben los cambios**. Los atacantes agregan ruido especialmente diseñado a las entradas, lo que hace que el modelo produzca salidas incorrectas mientras que la entrada parece inalterada para los observadores humanos.
Memorización de datos por la IA: riesgo de exposición involuntaria	Los modelos de IA modernos pueden memorizar inadvertidamente ciertos detalles de sus datos de entrenamiento, especialmente si los datos contienen muestras únicas o excepcionales. Los atacantes pueden aprovechar esto utilizando técnicas para extraer información confidencial que el modelo almacenó de manera accidental. Esto podría llevar a la exposición de datos personales del usuario o información comercial confidencial.
Inyección rápida: una amenaza para los modelos de lenguaje extensos	La inyección rápida es una amenaza específica para los modelos de lenguaje extensos (LLM, por sus siglas en inglés) como ChatGPT. Los desarrolladores programan los LLM para realizar tareas proporcionando indicaciones iniciales en lenguaje natural. Dado que los usuarios también interactúan con el modelo utilizando lenguaje natural, éste no puede distinguir por su naturaleza entre las instrucciones del desarrollador y las entradas del usuario. Los atacantes pueden crear entradas que anulen o manipulen el comportamiento del modelo, lo que provoca que realice acciones no deseadas o que revele información confidencial. Estos avisos maliciosos pueden ser introducidos directamente por el usuario o incrustados en los datos que procesa el modelo, como documentos o páginas web.

Estos son solo los ataques más relevantes; una descripción completa de todos los posibles ataques a los sistemas de IA está fuera del alcance de este documento.

* Cómo entender los ataques de envenenamiento de datos, https://bdtechtalks.com/2020/10/07/machine-learning-data-poisoning/
Ataques contra el aprendizaje automático: una descripción general, (https://elie.net/blog/ai/attacks-against-machine-learning-an-overview/

** Explicación y aprovechamiento de ejemplos adversarios, https://arxiv.org/abs/1412.6572
Ataques y defensas adversarios en aprendizaje profundo, https://arxiv.org/abs/2201.06192
Cómo confundir las redes neuronales antimalware: ataques adversarios y protección, https://securelist.com/how-to-confuse-antimalware-neural-networks-adversarial-attacks-and-protection/102949/

Vulnerabilidades de seguridad tradicionales

Los modelos de IA pueden ser susceptibles a debilidades de seguridad tradicionales:

Vulnerabilidades de la IA provenientes de recursos de terceros

Los sistemas de IA a menudo se basan en modelos de terceros o conjuntos de datos obtenidos de repositorios abiertos. Estos recursos pueden contener errores involuntarios o puertas traseras deliberadas insertadas por actores maliciosos. La incorporación de componentes en riesgo como estos puede introducir vulnerabilidades en el sistema de IA, lo que puede afectar su seguridad.

Riesgos a la cadena de suministro en el desarrollo de la IA

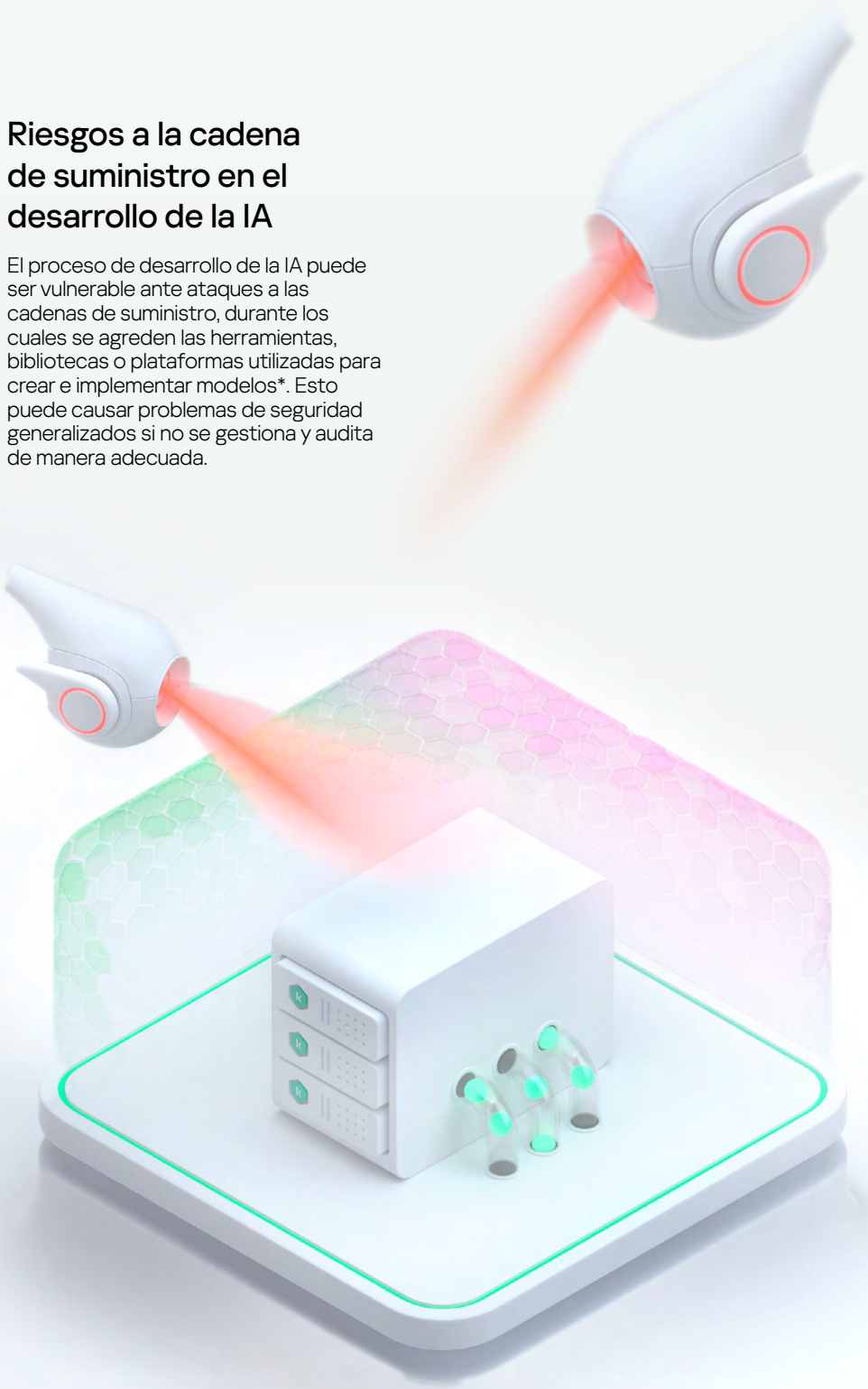
El proceso de desarrollo de la IA puede ser vulnerable ante ataques a las cadenas de suministro, durante los cuales se agreden las herramientas, bibliotecas o plataformas utilizadas para crear e implementar modelos*. Esto puede causar problemas de seguridad generalizados si no se gestiona y audita de manera adecuada.

Errores de código que exponen vulnerabilidades de la IA

Los errores en el código de las interfaces de acceso a la IA pueden provocar vulnerabilidades.

Riesgos del robo de componentes de la IA

Los sistemas de IA o sus componentes críticos, como modelos o conjuntos de datos, pueden ser robados si no se implementa la protección adecuada, que se explica en este documento.



* Cadena de dependencias de PyTorch-nightly en riesgo entre el 25 y el 30 de diciembre de 2022, https://pytorch.org/blog/compromised-nightly-dependency

Directrices

Concientización y capacitación en ciberseguridad

La implementación de nuevas tecnologías como la IA requiere **apoyo de la dirección, establecimiento de políticas internas y gobernanza, así como capacitación especializada para empleados sobre los riesgos y las amenazas asociados con la IA.** El último requisito es fundamental debido a la rápida evolución de esta tecnología. Muchos de los recursos de IA disponibles para los desarrolladores no son lo suficientemente maduros o no adoptan plenamente los principios de seguridad por defecto y seguridad por diseño. Como resultado, se impone una carga adicional sobre los desarrolladores de sistemas de IA, que deben abordar los riesgos potenciales que necesitan explicación.

Para este fin, una organización debe considerar implementar las siguientes medidas además de sus prácticas de seguridad habituales:

1

Los dirigentes de la organización deben ser conscientes de los riesgos de seguridad asociados con el uso de servicios de IA y aprender a gestionarlos.

2

Las políticas de seguridad de la organización deben actualizarse para abordar los requisitos específicos de los servicios de IA a fin de garantizar que todos los empleados y contratistas estén familiarizados con ellos.

3

Las políticas deben describir los riesgos asociados con el desarrollo y uso de servicios de IA, así como las restricciones actuales sobre su uso de conformidad con la legislación local.

4

Las políticas deben definir roles y responsabilidades relacionados con el uso de servicios de IA dentro de la empresa.

5

Se debe desarrollar o adquirir un curso de capacitación corporativa sobre el uso seguro de los servicios de IA dentro de la empresa a fin de garantizar que todos los empleados, incluidos los recién contratados, lo completen. El programa debe cubrir políticas organizacionales, amenazas existentes y ejemplos de incidentes, medidas de protección contra amenazas, leyes aplicables y otros temas relevantes, así como incluir ejercicios de simulación, si corresponde. Se debe evaluar a los empleados cuando completen el curso. El curso debe actualizarse oportunamente de forma regular.

6

Los cursos de seguridad de la información existentes deben actualizarse para incluir nuevos métodos utilizados por actores maliciosos que explotan los servicios de IA, como la generación de texto, para atacar a las empresas. La clonación de voz, la manipulación de fotografías o la generación de vídeos falsos son algunos ejemplos.

7

Se debería organizar el seguimiento de la legislación relacionada con el uso seguro de los servicios de IA. Las políticas internas y los programas de capacitación deben actualizarse oportunamente.

El modelado de amenazas ayuda a identificar, comprender y mitigar los posibles riesgos de seguridad en las primeras etapas del desarrollo de un sistema de IA.

Modelado de amenazas/Evaluación de riesgos

Este proceso es particularmente importante para los sistemas de IA, ya que se trata de una tecnología emergente con riesgos que evolucionan y se adaptan constantemente. Realizar una evaluación de riesgos puede ayudar a adelantar a estos desafíos. Además, el modelado de amenazas podría ayudar a los desarrolladores principantes a prepararse mejor para los desafíos asociados con el desarrollo de sistemas de IA. Les permitiría también identificar y mitigar de forma proactiva las debilidades del sistema de IA antes de que sean explotadas.

Para garantizar un proceso de modelado de amenazas eficaz, una organización debe cumplir con las **siguientes recomendaciones:**

Elegir una metodología de evaluación de riesgos (por ejemplo, STRIDE, DREAD, LINDDUN, PASTA, TRIKE) y desarrollar procedimientos para realizar evaluaciones de riesgos y modelos de amenazas para los servicios de IA. La metodología de evaluación también debe incluir un marco para determinar los niveles de riesgo, estrategias para gestionar los riesgos dentro de la organización (incluidos los umbrales de riesgo aceptables), procedimientos para el monitoreo de riesgos y la asignación de personal responsable para supervisar el proceso.

Al gestionar el riesgo, considere clasificar las amenazas en las siguientes categorías:

- Amenazas derivadas de la no utilización de servicios de IA.
- Amenazas derivadas de no conformidad.
- Amenazas derivadas del mal uso de los servicios de IA por parte de los usuarios.
- Amenazas a los modelos de IA y los conjuntos de datos utilizados para el entrenamiento.
- Amenazas que plantean los modelos de IA a los servicios.
- Amenazas a los datos asociados.
- Amenazas a los aspectos ambientales, sociales y de gobernanza (ESG).

Se debe realizar una evaluación de riesgos para todos los servicios de IA existentes y recientemente desarrollados.

Una evaluación de riesgos debe incluir la identificación de posibles actores amenazantes o atacantes, así como las amenazas y los riesgos identificados. Se deben utilizar materiales de referencia como NIST-AI-600-1, MITRE ATLAS, OWASP Top 10 for LLM Applications, DASF y CSA para identificar amenazas y riesgos conocidos.

La información sobre los riesgos identificados en los servicios de IA se debe comunicar a los dirigentes de la organización.

Seguridad de la infraestructura (nube)

Los servicios de IA generalmente se brindan como servicios en la nube y a menudo requieren una infraestructura especializada, por ejemplo, servidores equipados con GPU, FPGA, ASIC o TPU. Dada la sensibilidad de los sistemas de IA, deben protegerse de acuerdo con **los marcos de ciberseguridad más avanzados**, como NIST Cybersecurity Framework u otro con estándares similares. Los servicios de IA generalmente utilizan software gratuito o de código abierto como TensorFlow, PyTorch o Keras, junto con bibliotecas como Pandas, NumPy y SciPy. Para proteger este entorno, se deben considerar los **siguientes requisitos**:

1

Identificar todos los activos y mantener un inventario de activos de información como conjuntos de datos para entrenamiento y prueba de modelos, datos para ajuste fino del entrenamiento, bases de datos, tarjetas de datos, modelos y riesgos, datos entrantes y salientes hacia/desde el servicio, ponderaciones e hiperparámetros del modelo, datos de registro de los sistemas LLM.

2

Controlar el acceso en todos los niveles, incluida la red, los sistemas operativos, las bases de datos, el software, los datos y los modelos. Implementar la autenticación de dos factores (2FA) para el acceso administrativo.

3

Registrar todos los eventos y asegurarse de que los datos de registro estén protegidos. Supervisar incidentes de seguridad y posibles infracciones.

4

Implementar medidas de seguridad contra malware y otros tipos de ataques. Aplicar parches de seguridad a los componentes de infraestructura de forma periódica.

5

Segmentar la red para proteger áreas confidenciales. Utilizar cifrado para datos en tránsito y en reposo.

6

Garantizar la integridad de los datos críticos y verificar la autenticidad de las bibliotecas y los modelos en uso.

7

Proporcionar redundancia de servidor y canal de comunicación. Realizar copias de seguridad periódicas y asegurarse de su correcto funcionamiento.

8

Almacenar claves de forma segura en un KeyVault.

9

Aplicar los principios de mínimo privilegio y confianza cero en toda la infraestructura.

En función de la infraestructura que respalde los servicios de IA, los requisitos adicionales pueden incluir los siguientes:

Utilizar una puerta de enlace API para administrar el acceso a los modelos y manejar la autenticación a través de API.

Implementar medidas de seguridad específicas para entornos de Kubernetes.

Adherirse a las mejores prácticas para proteger los servicios basados en la nube.

Garantizar la integridad del código fuente, los datos de entrenamiento, los modelos y los guiones de automatización.

Aislar los datos de entrenamiento, los modelos y los entornos de entrenamiento para evitar fugas o contaminación de datos.

Cadenas de suministro y seguridad de datos

Los ataques a las cadenas de suministro suponen una amenaza importante para la infraestructura de cualquier organización. La arquitectura de la IA no es una excepción. Se conocen casos en los que bibliotecas especializadas para el entrenamiento de redes neuronales fueron el objetivo de dichos ataques. Sin embargo, cuando se trata de la IA, surge una preocupación específica con respecto tanto a la seguridad del proveedor de servicios como a la protección de los modelos de aprendizaje automático.

El acceso a modelos de IA avanzados, especialmente los LLM, a menudo depende de soluciones basadas en la nube. Sin embargo, indisponibilidad de ciertos modelos en algunas regiones, junto con otras restricciones, pueden impulsar a los empleados y desarrolladores de la empresa a optar en sus tareas diarias por servicios de terceros (proxies) que revenden el acceso a los modelos de IA a través de API. Esta práctica conlleva riesgos significativos, desde abrir un vector adicional para fugas de datos en caso de un incidente de seguridad en el servicio proxy hasta la práctica poco ética de abusar los datos obtenidos para revenderlos o entrenar sus propias versiones de LLM. **Es importante comprender estos riesgos.** Se debe revisar cuidadosamente las políticas de privacidad tanto del proveedor principal como del proxy y realizar capacitaciones de concientización en toda la empresa sobre los peligros de utilizar servicios de terceros para tareas laborales.

Para mitigar estos riesgos, una organización puede elegir implementar un servicio LLM local. Este enfoque garantiza que los datos confidenciales procesados por el LLM permanecerán dentro de la empresa si se cumplen ciertas condiciones específicas (por ejemplo, aislamiento de la red, desactivación de la telemetría, etc.). Sin embargo, además de los riesgos asociados con las vulnerabilidades en los marcos de aprendizaje automático, este método implica amenazas relacionadas con puertas traseras en los modelos. Eso significa que los formatos de datos utilizados para distribuir modelos de aprendizaje automático pueden tener diferentes niveles de seguridad. Algunos formatos formatos permiten potencialmente la incorporación de código arbitrario que puede ejecutarse cuando se ejecutan los modelos. Varias investigaciones demostraron que existen modelos disponibles en repositorios públicos (aunque su número es limitado) que pueden ejecutar código específico al cargarse. El uso de modelos de repositorios de terceros, en lugar de los originales, puede deberse a la falta de disponibilidad de modelos descargables en ciertas regiones o a restricciones de licencia. El uso de formatos seguros, como los safetensors, ayuda a abordar este problema, pero es necesario generar conciencia entre los desarrolladores y analistas de datos sobre la importancia de seleccionar fuentes confiables para los modelos y usar formatos seguros para intercambiar ponderaciones de los modelos.

- Asegurarse de que los modelos de IA se obtengan de fuentes confiables y legítimas. Evitar utilizar repositorios de terceros.
- Utilizar formatos seguros como safetensors para intercambiar ponderaciones de modelos y evitar el riesgo de ejecución de código arbitrario.
- Implementar medidas para detectar y responder a ataques a las cadenas de suministro que afectan componentes relacionados con la IA.
- Evaluar y revisar las políticas de privacidad de los servicios de terceros y servidores proxy utilizados para acceder a los modelos de IA a fin de garantizar que cumplan con los estándares de seguridad.
- Implementar modelos de IA a nivel local en condiciones que garanticen la privacidad de los datos, como el aislamiento de la red y la desactivación de funciones de telemetría.
- Establecer protocolos para la implementación segura de modelos locales con el objetivo de minimizar los riesgos asociados con posibles puertas traseras en los modelos de aprendizaje automático.
- Actualizar los marcos de aprendizaje automático y aplicar parches periódicos para abordar vulnerabilidades conocidas.
- Implementar medidas para garantizar que los datos confidenciales procesados por los modelos de IA no salgan de la infraestructura de la organización.
- Al utilizar API de terceros, se deben realizar auditorías de seguridad de acuerdo con los principales estándares internacionales, como OWASP API Security Top 10.

Pruebas y validaciones

Una vez realizada la evaluación e identificados los riesgos, es fundamental entender cómo protegerse contra errores accidentales o deliberados en el entrenamiento y la aplicación del modelo. Para lograr este objetivo, una organización podría considerar implementar las siguientes medidas:

1

Evaluar el daño potencial que podrían causar errores accidentales o deliberados en el sistema. Evaluar el valor de los datos utilizados para entrenar el modelo de IA y los datos que procesa.

2

Determinar si se están utilizando marcos, modelos o conjuntos de datos de código abierto para construir el sistema de IA.

3

Identificar la base de usuarios potenciales: empleados de la empresa, clientes o acceso público abierto.

4

Verificar el cumplimiento de las mejores prácticas de aprendizaje automático en la construcción de modelos. Asegurarse de que los conjuntos de datos estén divididos correctamente en conjuntos de entrenamiento, de prueba y de validación dependiendo de cómo funciona el modelo.

5

Al validar los modelos de IA y sus métricas (falsos positivos y falsos negativos), se debe comprobar que los criterios para la división del conjunto de datos sean apropiados para la naturaleza de los datos (por ejemplo, partición cronológica para datos temporales y prevención de fugas de datos).

6

Evaluar qué características utiliza el modelo para tomar decisiones y si son consistentes con la intuición de los expertos. Utilizar métodos de interpretación de modelos, como los vectores SHAP, para comprender el proceso de toma de decisiones del modelo.

7

Evaluar el desempeño real del modelo para asegurarse de que esté entregando los resultados esperados. Monitorear el modelo de manera continua, ya que la distribución de los datos de entrada puede cambiar con el tiempo y degradar potencialmente el rendimiento del modelo.

8

Adaptar el plan de pruebas para verificar si el modelo es susceptible a vulnerabilidades específicas de los modelos de aprendizaje automático (por ejemplo, ataques adversarios y envenenamiento de datos).

Informes de vulnerabilidad

La IA es un área tecnológica relativamente nueva que está experimentando un rápido desarrollo. A pesar de los importantes beneficios que ofrecen, muchos sistemas de IA son susceptibles a vulnerabilidades específicas de esta tecnología.

Una de las principales preocupaciones se debe a que algunos sistemas de IA contienen vulnerabilidades que pueden explotarse para obtener acceso no autorizado a sus datos. Otro ejemplo de vulnerabilidades en los sistemas de IA es el sesgo, cuando los modelos se entrenan con datos que no son representativos o contienen sesgos ocultos. Por ejemplo, los sistemas de IA pueden verse afectados por sesgos de prejuicio cuando los estereotipos y suposiciones sociales erróneas se infiltran en el conjunto de datos del algoritmo, o por sesgos de medición provocados por datos incompletos. Como resultado, estos sistemas pueden tomar decisiones injustas o discriminatorias, lo que impacta negativamente a los usuarios y socava la confianza en las tecnologías de IA.

Para abordar estas cuestiones, es necesario implementar un mecanismo que permita a los usuarios **informar sobre las vulnerabilidades identificadas** y sesgos en los sistemas de IA. Este mecanismo de informes permitirá a las organizaciones recibir comentarios con rapidez y tomar medidas:

1

Establecer una política pública que defina las vulnerabilidades en los sistemas de IA y describa cómo los usuarios pueden denunciarlas.

2

Proporcionar métodos seguros para que los usuarios informen vulnerabilidades, como formularios web encriptados o direcciones de correo electrónico dedicadas.

3

Definir procedimientos para evaluar, priorizar y remediar rápidamente las vulnerabilidades reportadas.

4

Comunicarse con la persona que notificó la vulnerabilidad para informar sobre el estado y la resolución del problema.

5

Mantener informados a los usuarios sobre las vulnerabilidades conocidas y los esfuerzos de solución para generar confianza y demostrar responsabilidad.

6

Colaborar con investigadores de seguridad a través de programas de recompensas por errores. Esto también ayudará a mantenerse al tanto de las amenazas emergentes y las mejores prácticas de seguridad de IA.

Defensa contra ataques a modelos de aprendizaje automático

Dados los avances actuales en el desarrollo de IA, algunos de sus componentes pueden ser vulnerables a los ataques típicos de los modelos de aprendizaje automático. Estos ataques pueden explotar vulnerabilidades, por ejemplo, al introducir deliberadamente datos malformados o comandos ocultos en el modelo. Por lo tanto, las organizaciones que utilizan IA gratuita para desarrollar sus sistemas deben ser conscientes de estos riesgos. La protección contra los ataques típicos de los modelos de aprendizaje automático requiere **la implementación de diversas medidas de seguridad**, como las siguientes:

Incorporar ejemplos adversarios* en el conjunto de datos de entrenamiento para ayudar al modelo a aprender a manejar estas entradas de manera más eficaz.

Aplicar técnicas de destilación que ayuden a que el modelo sea más resistente a las entradas adversas y simplifique su proceso de toma de decisiones.

Considerar el uso de modelos monótonos**, que pueden mejorar la estabilidad y reducir la susceptibilidad a la manipulación adversaria.

Incorporar sistemas que puedan detectar entradas adversas o anómalas en las solicitudes de los usuarios, lo que permite que el modelo detecte y rechace intentos maliciosos antes de procesar los datos.

Para protegerse contra el envenenamiento de datos, los desarrolladores de sistemas de IA deben **analizar las muestras de entrenamiento** para objetos anómalos y comparar el rendimiento de los nuevos modelos con versiones anteriores a fin de identificar cualquier cambio brusco en sus propiedades.

Para protegerse contra ataques de inyección de prompts en los LLM, los desarrolladores de sistemas de IA pueden implementar un sistema que analice las solicitudes entrantes de los usuarios u otros datos de terceros introducidos en la entrada del LLM. Otro enfoque es analizar las respuestas a dichas solicitudes y evaluar su conformidad con la tarea actual del sistema.

* Cómo confundir las redes neuronales antimalware. Ataques adversarios y protección, https://securelist.com/how-to-confuse-antimalware-neural-networks-adversarial-attacks-and-protection/102949/

** Modelos monótonos para la detección dinámica de malware en tiempo real: https://arxiv.org/pdf/1804.03643

Actualizaciones de seguridad y mantenimiento periódicos

El campo de la IA, y en particular el de los LLM, es todavía relativamente joven y la calidad del código no siempre es elevada. Como resultado, muchos marcos y herramientas utilizados para trabajar con el aprendizaje automático pueden contener un **número significativo de vulnerabilidades**. Afortunadamente, ahora estamos en una fase en la que los marcos populares se están actualizando de manera continua para alcanzar los estándares de calidad de producción, con lanzamientos regulares y actualizaciones de seguridad. Además, surgen programas de recompensas por errores destinados a encontrar vulnerabilidades en la infraestructura de aprendizaje automático, lo que ayuda a abordar estos problemas rápidamente.

Así, es esencial monitorear continuamente el estado de la infraestructura, desde las plataformas utilizadas para rastrear experimentos hasta las bibliotecas diseñadas para comunicarse con servicios en la nube. Mantener la infraestructura actualizada puede llevar más tiempo que en proyectos de campos de evolución más lenta. Además, el uso de las últimas versiones de las bibliotecas puede generar problemas de compatibilidad, lo que requiere una mayor inversión en el desarrollo y mantenimiento de la funcionalidad del código dependiente. Estos costos deben tenerse en cuenta al planificar iniciativas de IA.

Otro riesgo asociado con el uso de modelos de IA basados en la nube, como los LLM, es el **ciclo de vida relativamente corto de cada versión del modelo**. El modelo seleccionado para un proyecto podría reemplazarse por el proveedor de la plataforma por una nueva versión en un corto período de tiempo. Si bien se espera que la calidad general de los modelos mejore, su comportamiento para tareas específicas y su resistencia a los ataques pueden cambiar. Por ejemplo, puede sufrir inyecciones de prompts o intentos de obtener resultados no autorizados a través de jailbreaks. Esto requiere una planificación avanzada para garantizar una transición fluida entre modelos sin comprometer la calidad de las tareas posteriores ni el nivel de seguridad:

1. Asegurarse de que la infraestructura se mantenga actualizada con los últimos parches de seguridad y actualizaciones del marco.
2. Participar enérgicamente en programas de recompensas por errores y utilizar herramientas de análisis de vulnerabilidades para detectar debilidades en los marcos de aprendizaje automático y la infraestructura de la IA.
3. Revisar y aplicar actualizaciones periódicas de seguridad para herramientas y bibliotecas de aprendizaje automático a fin de reducir la exposición a vulnerabilidades conocidas.
4. Planificar para abordar posibles problemas de compatibilidad al utilizar las últimas versiones de bibliotecas y marcos, así como asignar recursos de desarrollo y prueba.
5. Implementar una estrategia para gestionar el ciclo de vida de los modelos de IA basados en la nube, con planes de transición a nuevas versiones a medida que el proveedor las publique.

Cumplimiento con normas internacionales

Dado el rápido desarrollo del marco regulatorio en el área de la IA, el cumplimiento con las leyes pertinentes y la adhesión a las mejores prácticas adquieren cada vez mayor importancia. En primer lugar, los datos de entrenamiento de IA pueden recopilarse de múltiples fuentes en diferentes jurisdicciones, lo que dificulta el procesamiento y el uso de esta información. Además, los modelos a menudo proceden de **repositorios abiertos** y agregan un nivel de incertidumbre en cuanto a su conformidad con los requisitos regulatorios de una jurisdicción particular.

Por lo tanto, los desarrolladores de IA se enfrentan a la difícil tarea de garantizar el cumplimiento de todos los requisitos legales en los países donde se utiliza el sistema. La mejor estrategia en esta situación es **seguir los estándares de los líderes en regulación de IA**, como China, la Unión Europea o Estados Unidos. Muchos países ya están compartiendo sus enfoques e implementando requisitos similares, lo que permite a los desarrolladores prepararse con antelación para la implementación global de ese sistema:

1

Establecer pautas para el uso y desarrollo ético de la IA a fin de garantizar la transparencia y la responsabilidad en los procesos relacionados.

2

Asegurarse de que todos los datos recopilados de distintas fuentes cumplan con las leyes de privacidad de datos en cada jurisdicción, como el Reglamento General de Protección de Datos (RGPD) en Europa o la Ley de Privacidad del Consumidor de California (CCPA) en EE.UU.

3

Al utilizar modelos de IA de repositorios abiertos, verificar que cumplan con los derechos de propiedad intelectual.

4

Seguir los marcos regulatorios líderes, como la Ley de IA de la Unión Europea o la Carta de Derechos de IA de EE.UU., ya que estos suelen ser utilizados como puntos de referencia por otros países.

5

Mantenerse al tanto de las regulaciones de IA nuevas y en evolución en todo el mundo.

6

Auditar periódicamente los modelos y sistemas de IA para comprobar que cumplan con las normas internacionales, a fin de identificar y mitigar posibles riesgos legales y éticos.

Conclusión

Como la mayoría de las innovaciones tecnológicas, las tecnologías de IA presentan tanto grandes oportunidades como amenazas considerables. Los riesgos de ciberseguridad asociados con la IA y su impacto en la sociedad dependen del comportamiento y las intenciones del desarrollador. Para desarrollar e implementar IA en su infraestructura de forma segura, las organizaciones deben seguir las directrices técnicas especiales, ya que llevar a cabo este proceso sin las recomendaciones adecuadas puede suponer riesgos importantes. Es vital que las organizaciones establezcan una cultura de seguridad y responsabilidad durante todo el ciclo de vida de la IA e incorporen controles de seguridad básicos, desde la evaluación de riesgos y las pruebas del sistema hasta la protección de las cadenas de suministro y el mantenimiento continuo. La implementación exitosa de los requisitos presentados ayudará a **mitigar los riesgos** relacionados con la introducción de sistemas de IA en las operaciones de una empresa.

Más información

www.kaspersky.com

#kaspersky
#bringonthefuture