INFORME SOBRE

SEGURIDAD DE LA INTELIGENCIA ARTIFICIAL 2024

Ametic
LA VOZ DE LA INDUSTRIA DIGITAL

Seguridad de la IA

Introducción

El éxito reciente de la Inteligencia Artificial (IA) en aplicaciones tales como pronósticos financieros, recomendación de productos en compras en línea, reconocimiento de imagen y voz, generación de lenguaje, etc. ha llevado a una mayor adopción de la IA en muchos escenarios, y, en consecuencia, a que sea cada vez más ubicua. Por un lado, las soluciones impulsadas por IA se utilizan cada vez más en la toma de decisiones críticas, por ejemplo, en la gestión de centrales eléctricas o en el diagnóstico médico. Por otro lado, la IA se integra como un componente en sistemas más grandes que se vuelven dependientes de las decisiones de sus algoritmos. Estos nuevos desarrollos requieren de manera urgente que las soluciones de IA sean seguras y sus decisiones fiables.

En el campo de la seguridad de la información, se definen cinco pilares principales para caracterizar la seguridad de un sistema de información, también aplicables a caracterizar los sistemas de aprendizaje automático: Confidencialidad, los modelos IA y sus sets de datos solo deben ser accesibles a las personas autorizadas; Integridad, el modelo no es alterado por terceros; Disponibilidad, el modelo es siempre accesible; Autenticidad, las predicciones y los sets de datos son legítimos. Finalmente, No repudio, que permite asegurar la trazabilidad y autoría del algoritmo. Si durante el ciclo de vida de los sistemas de IA no se cumple alguna de las propiedades anteriormente mencionadas el sistema será vulnerable y quedará alterado el correcto funcionamiento del modelo, poniendo en riesgo la confianza que tienen los usuarios en los modelos; causando sesgos y efectos discriminatorios.

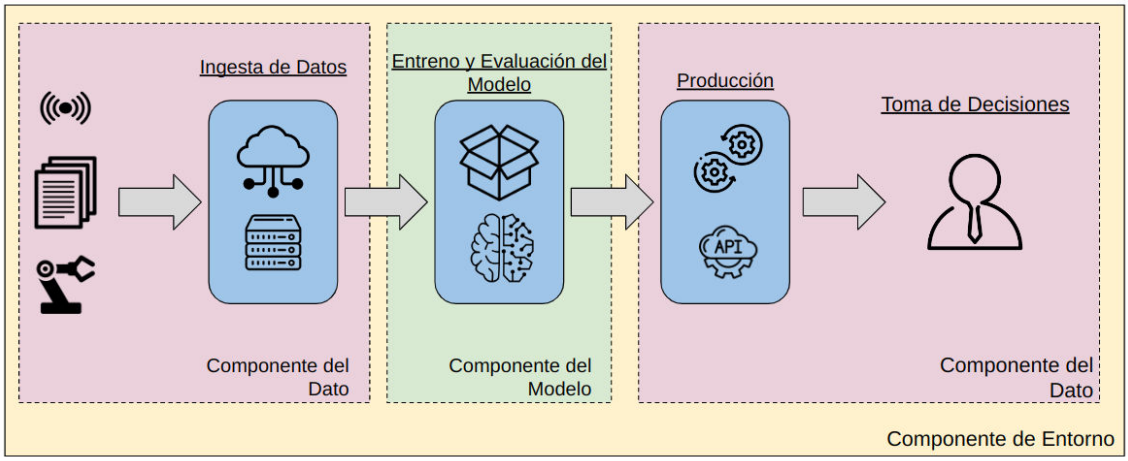

Figura 1: Ciclo de vida de las aplicaciones basadas en Inteligencia Artificial.

En los últimos años se ha hecho eco de casos de comportamientos no esperados y ataques contra estos sistemas, causando daños económicos, fuga de información y, de forma no cuantificable, pérdida de la confianza que tienen los usuarios en los algoritmos de IA.

- **Microsoft Azure Service Disruption** - ataque sucedido en 2020 y que formó parte de un ejercicio en el que el Red Team de Azure logró comprometer el funcionamiento de un servicio interno de Azure. El ataque consistió en una primera fase en la que el equipo atacante, mediante robo de credenciales, consiguió acceder al modelo de datos ML y al dataset de entrenamiento. A partir de ahí, confeccionaron un set de “evasive adversarial data” que lanzaron contra el modelo objetivo mediante una API expuesta, consiguiendo realizar un ataque de evasión y alterando la función de clasificación del modelo ML.

- **VirusTotal Poisoning** - ataque sucedido en 2020, en el que se envenenó el dataset usado por los modelos de ML de VirusTotal para clasificar familias de ransomware, mediante el uso de ficheros mutados a partir de una muestra original de ransomware. El ataque provocó que muchos proveedores de soluciones antivirus empezaran a clasificar como ransomware ficheros que no lo eran, y que ni siquiera se ejecutaban.

Al igual que para los sistemas de información, existe una interminable lista de potenciales ataques donde la IA es vulnerable. Se pueden categorizar estos ataques dependiendo del componente vulnerable. Si se compromete la entrada o la salida del modelo - es decir que los datos que entran/salen del modelo pierden la integridad, autenticidad y no-repudio - hablamos de Componente del Dato. Si se intenta poner en riesgo la integridad y autenticidad de la algoritmia, hablamos de Componente de Modelo. Finalmente, si se compromete la confidencialidad, disponibilidad y la integridad de este, hablamos de Componente de Infraestructura.

- Componente del Dato
  - **Sesgos en los datos** Unos de los problemas recurrentes en las aplicaciones basadas en los sistemas de Inteligencia Artificial es la confianza que tiene el usuario en el dato, tanto la entrada al modelo como las predicciones automáticas que produce.
  - **Recolección de datos**: Un ataque conocido contra los sistemas basados en IA son los de suplantación (del Inglés, Spoofing Attacks). En este tipo de ataques un atacante se hace pasar por un actor legítimo con la finalidad de introducir nuevos datos, corrompiendo así la autenticidad de los datos.

- Componente de Modelo
  - **Ataques de envenenamiento**: Esta tipología de ataques busca corromper el conjunto de entrenamiento del modelo para reducir la precisión del modelo IA. Los principales ataques son los de disponibilidad e integridad, enfocados a inyectar datos orientados a degradar partes específicas del modelo.
  - **Ataque del adversario**: Este tipo de ataques busca aprender los patrones internos del algoritmo con el fin de poder inferir información alterada. En esta línea, los ataques de white-box buscan determinar información del modelo conociendo la arquitectura y los vectores de entrada. Por otro lado, en los ataques de black-box solo se conocen las entradas del modelo.

- Componente de Entorno
  - **Legitimidad**: Estos ataques explotan puertas traseras introducidas en el código fuente de librerías y repositorios no autorizados.
  - **Acceso al modelo:** Se tiene que asegurar que solo las personas autorizadas tienen acceso al modelo y que la API que se utiliza es segura, evitando así accesos no deseados y exfiltración de información.

Actualmente, MITRE, organización sin ánimo de lucro enfocada al estudio de amenazas y fraudes digitales, está elaborando un marco de ataques enfocado hacia los sistemas de inteligencia artificial, llamado ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems), de forma similar a los otros marcos que ya son una referencia en sus respectivos campos, tales como ATT&CK, PRE-ATT&CK, CAPEC y DEFEND. Otras iniciativas, también en desarrollo, son *NISTIR 8269 - A Taxonomy and Terminology of Adversarial Machine Learning* y *ENISA Big Data Threat Landscape and Good Practice Guide*.

Expectativas (ventajas)

Todo y que en un futuro disponer de aplicaciones seguras será un requisito indispensable para que estas puedan ser puestas en funcionamiento, hoy en día, dos dimensiones, Trustworthiness y Marco legal, han de permitir aumentar la calidad de los sistemas de Inteligencia Artificial:

Trustworthiness: La ventaja directa de desarrollar aplicaciones teniendo en cuenta la seguridad de la IA es que la Confianza (Trustworthiness, en inglés) que tiene el usuario final en el modelo se ve reforzada, ya que este puede tanto entender el modelo (transparencia algorítmica), como asegurar que no tiene sesgos y que se evitan problemas de discriminación. De igual modo, avanzar hacia una IA segura ayuda a democratizarla, de manera que pueda llegar a más usuarios y no sea vista como una amenaza.

Marco Legal: Si bien tener una IA segura es un requisito deseable, las Instituciones Europeas están trabajando en diferentes marcos normativos para que la seguridad sea un prerrequisito para todo sistema basado en proveer decisiones automatizadas. Con el objetivo de fomentar la confianza y abordar estos riesgos inherentes de la IA, la Comisión Europea ha propuesto la creación de un marco de trabajo legal para la regulación de la IA en la UE, llamado inicialmente *Artificial Intelligence Act (AIA)*. En dicho Marco se propone categorizar las aplicaciones de IA de acuerdo a sus riesgos, adaptando los requisitos al nivel de riesgo para que no se dañe al usuario final. Otra iniciativa reguladora de la UE, también en curso, es la Civil Liability, que tiene el propósito de adaptar las normas de responsabilidad civil a la era digital y de la IA.

La IA comporta en algunos casos el tratamiento masivo de datos, que pueden contener información personalmente identificable de personas físicas. Normativas como la RGPD protegen los derechos y libertades de las personas físicas en relación con el tratamiento de sus datos personales y la privacidad, y afectan también a la IA.

Obstáculos (inconvenientes, carencias de desarrollo)

Una de las principales carencias de esta incipiente, pero necesaria, tecnología es la escasa investigación al respecto y la falta de validación de ataques contra modelos IA en entornos de producción. Esta carencia también se traslada al mundo de DevOps, donde no existe ningún framework de amplia adopción que permita agilizar y estandarizar la seguridad para sistemas basados en IA.

Otro inconveniente detectado es la falta de herramientas de Benchmarking, Baselines o Checklist enfocadas a testear, establecer o auditar la seguridad de la IA. Se han dado casos como el del repositorio Keras, en el que una incorrecta comprobación de los hashes permitía la descarga de modelos sin que se verificase su integridad. Benchmarks de seguridad para los módulos y sistemas de IA, realizados por asociaciones como el *Center for Internet Security*, ayudarían a una implementación más segura de la IA.

Oportunidades de negocio y mensajes para los primeros receptores (Comisión Ciberseguridad)

Aunque la IA es una herramienta de presente y futuro, esta debe desarrollarse de forma segura, minimizando todo impacto previsible. En esta línea de trabajo, asociaciones tales como MITRE o la Unión Europea están desarrollando metodologías para cuantificar e investigar las amenazas (incluyendo técnicas de ataque) a los sistemas basados en IA, así como también medidas de protección y controles adecuados. Como ejemplo, se puede citar el Proyecto Europeo SPARTA, en su foco en la ML/DL, y actualmente en curso.

El desarrollo de mecanismos de secure IA ha de permitir:

- Hacer que los conjuntos de datos sean privados, en el sentido de que no contengan datos personalmente identificables o sensibles, por ejemplo, mediante técnicas de anonimización o de generación de datos sintéticos. También hacer uso de la agregación y de la minimización de datos.
- Hacer un uso seguro y no invasivo de los algoritmos de ML, desvinculando a los datos de los usuarios.
- Una implementación segura de la IA mediante la aplicación rigurosa de las protecciones y controles de seguridad, para impedir el acceso a los conjuntos de datos a personas no autorizadas y evitar posibles exfiltraciones de datos.

- Los sistemas de AI han de proveer mecanismos de consentimiento y eliminación ágil de datos a petición del usuario.
- Hacer una evaluación del impacto de las actividades de procesamiento en la protección de los datos personales, para asegurar que el procesamiento no genera un riesgo para los derechos y libertades de las personas.

Este nuevo paradigma es una clara oportunidad para el desarrollo de nuevas tecnologías que consoliden la IA segura. A continuación, listamos algunas tecnologías disruptivas:

- **DevSecOps para la IA:** La existencia de un marco de trabajo DevSecOps común, que incorpore herramientas automatizadas específicas para la IA que permitan la comprobación y análisis de los sistemas IA durante las fases de desarrollo y operación.
- **Federated Learning y Transfer Learning:** El uso de paradigmas de IA que permitan el entreno de modelos federados y por transferencia seguirán en aumento durante los próximos años. El uso de algoritmos por aprendizaje federado permite entrenar un algoritmo de forma descentralizada, donde cada dispositivo tiene de forma privada su conjunto de datos, minimizando potenciales problemas de compartición de datos. De forma similar, los algoritmos de transfer learning permiten minimizar la compartición de datos y por tanto crear sistemas “secure-by-design”. Estos algoritmos se basan en entrenar un modelo en un set de datos específico, pero se utilizan para realizar la inferencia sobre datos de otro entorno.

Otra oportunidad la encontramos en los sistemas basados en Blockchain. La gran ventaja de los sistemas blockchain es que proporcionan un entorno descentralizado, proporcionando evidencias que permiten verificar que el modelo IA no ha sido alterado ni manipulado. Se proporcionan mecanismos de verificación de la integridad y autenticidad de la entrada y salida de datos, la no manipulación de los modelos de la IA, y resuelve el problema de los entornos tradicionales distribuidos IA, al asegurar cuáles son las dependencias existentes entre las predicciones realizadas y sobre qué modelo y datos se entrenó.

- **Computación Verificable:** Con la aplicación de la tecnología blockchain podemos ofrecer garantías de veracidad a la transferencia del aprendizaje, al permitir verificar que el resultado del aprendizaje obtenido se realiza sí o sí sobre un determinado modelo y con los datos de entrada que se dicen que han sido proporcionados.
- **Criptografía Homomórfica:** Permite ofrecer soluciones de privacidad respecto a los datos que se utilizan en los entrenamientos de los modelos de IA. Mediante el uso de la criptografía homomórfica es posible entrenar, por ejemplo, redes neuronales sin tener acceso directo a los datos, garantizando la privacidad y confidencialidad del conjunto de datos proporcionado a terceros

Links

- (https://www.cisecurity.org/cis-benchmarks/
- https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence
- AIA - https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI(2021)698792_EN.pdf

- Civil liability - https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12979-Civil-liability-adapting-liability-rules-to-the-digital-age-and-artificial-intelligence_en

INFORME SOBRE

SEGURIDAD DE LA INTELIGENCIA ARTIFICIAL 2024

OFICINA MADRID
PRÍNCIPE DE VERGARA, 74, 4ª PLANTA
28006 - MADRID
TEL. 91 590 23 00

OFICINA BARCELONA
AVDA. SARRIÁ, 28, 1º- 1ª
08029 - BARCELONA
TEL. 93 241 80 60

Ametic
LA VOZ DE LA INDUSTRIA DIGITAL

www.ametic.es