5. Identificar los problemas de seguridad de la Inteligencia Artificial

Las técnicas de IA tienen un potencial tremendo. Están creciendo rápidamente e infiltrándose gradualmente en todas las profesiones y aspectos de nuestra sociedad. La IA también plantea **preguntas importantes sobre la seguridad y la responsabilidad social y ambiental.**

La IA puede facilitar **actos maliciosos** como la difusión de noticias falsas o ciberataques. Han surgido nuevas cuestiones de seguridad en torno a cómo podemos estar seguros de que la **IA hará lo que esperamos de ella.** También existen otras preocupaciones con respecto al uso de datos, especialmente datos personales o sesgados. Además, hay consideraciones sobre el **impacto ambiental de la IA.**

Exploraremos estos problemas con más detalle en este capítulo y el siguiente. En cada capítulo, sugeriremos algunas cosas a considerar para asegurarnos de que estés atento y te comportes de forma responsable. Al final de estos dos capítulos, veremos cómo la **gobernanza de la IA** puede alentar o incluso obligar a las empresas a desarrollar sistemas de IA seguros, responsables y contables. ¿Estás listo?

Explorar el uso malicioso de la IA

La IA es una herramienta que proporciona un servicio a los humanos. Como hemos visto, podemos utilizarla para lograr avances significativos en muchos sectores, como la atención médica, la educación y el medio ambiente. Sin embargo, también puede facilitar el trabajo de personas con intenciones maliciosas. Vamos a explorar dos ejemplos de uso malicioso de la IA que es probable que te encuentres: las noticias falsas y los ciberataques.

Noticias Falsas

Es posible que hayas encontrado videos extraños o improbables circulando en línea: Barack Obama insultando a Donald Trump o Mark Zuckerberg hablando de manipular a los usuarios de Facebook. Estos **videos falsos** comenzaron a aparecer en 2018 y pueden parecer sorprendentemente reales. La tecnología subyacente se basa en potentes técnicas de inteligencia artificial. La gente se refiere a estos medios sintéticos como **"deepfakes".**

La tecnología se basa en el **Deep Learning**, sobre lo cual aprenderás más en un capítulo posterior.

¿Pero las noticias falsas no existían mucho antes de la inteligencia artificial?

Sí, pero en el pasado, solo un puñado de expertos manipulaba fotos. Ahora, un número creciente de personas puede hacerlo, y los resultados son cada vez más convincentes. Visita el sitio "This person does not exist" (Esta persona no existe) para ver retratos creados utilizando esta técnica de inteligencia artificial. Sería fácil confundir a estas personas generadas artificialmente con humanos reales.

Los *deepfakes* pueden referirse a fotos, grabaciones de audio o videos manipulados. En 2019, una aplicación china llamada Zao causó un gran revuelo. Los usuarios pueden utilizar Zao para reemplazar el rostro de un actor en un video musical o una película con cualquier foto que elijan. Con el crecimiento de la IA generativa, que exploraremos en la próxima parte del curso, se han difundido en las redes sociales una gran cantidad de videos e imágenes falsos generados por IA. Un ejemplo destacado fue la imagen del Papa Francisco con una chaqueta blanca.

Sin embargo, la falsificación de la IA no se detiene en imágenes divertidas. La IA crea textos falsos, videos y mucho más. Estas nuevas técnicas conllevan riesgos, incluida la **desinformación generalizada.**

¿Qué podemos hacer al respecto?

Hay dos herramientas que puedes usar: el sentido común y el pensamiento crítico. Para protegerte de los *deepfakes*, equípate con estos, junto con algunos consejos:

- **Cuestiona tus fuentes y asegúrate de que la información sea legítima.**
- **Verifica tu información** consultando otros sitios web de noticias.
- También puedes utilizar **sitios web de verificación de hechos** como FactCheck.org del Centro de Política Pública Annenberg o Fact Checker del Washington Post.

Ciberataques

Con las crecientes capacidades de la IA y su mayor accesibilidad, los ciberdelincuentes están utilizando cada vez más la IA para llevar a cabo ciberataques. La utilizan para detectar vulnerabilidades en tus dispositivos o para automatizar ataques de phishing, revelando datos personales, contraseñas o detalles de cuentas bancarias. Veamos algunos ejemplos de posibles y exitosos ciberataques.

Ataques de Phishing

Las personas pueden utilizar la IA para identificar fácilmente perfiles de posibles víctimas que son propensas a hacer clic en enlaces falsos o para personalizar correos electrónicos utilizando datos capturados tanto de filtraciones de datos previas como de plataformas de redes sociales como LinkedIn, Facebook y Twitter.

Una herramienta automatizada, **SNAP_R**, puede generar tweets de phishing realistas dirigidos a ciertos usuarios, como se describe en este artículo de Forbes. Los investigadores que desarrollaron esta herramienta notaron tasas de clic considerablemente más altas que las técnicas anteriores que no utilizaban inteligencia artificial.

Suplantación de Identidad

Los ciberataques pueden utilizar las mismas técnicas de falsificación descritas anteriormente. Una tecnología conocida como “Deep voice” utiliza la IA para suplantar la voz de una persona basada en muestras de audio de su voz. Los atacantes pueden obtener estas muestras de grabaciones de reuniones en línea o discursos públicos (de fácil acceso para periodistas, políticos o ejecutivos corporativos).

En 2020, un gerente de un banco de Hong Kong recibió una llamada de su jefe con noticias muy positivas: la empresa planeaba llevar a cabo una importante adquisición y, para hacerlo, el gerente necesitaba la aprobación de su banco para realizar varias transferencias por un total de 35 millones de dólares. El empleado reconoció la voz de su jefe, pensó que todo era genuino y envió el dinero. Sin embargo, la voz era de inteligencia artificial.

Un Ciberataque de Extremo a Extremo Utilizando IA

Las diversas formas de ciberataque discutidas anteriormente son solo una pequeña fracción de las posibilidades. Cada una representa una amenaza por sí sola, pero también pueden combinarse para crear ataques en los que la IA realiza todas las acciones necesarias de principio a fin, ya sea reconocimiento de objetivos, intrusión en sistemas, ejecución de comandos, elevación de privilegios o exfiltración encubierta de datos.

Explorar los desafíos asociados con la seguridad de la IA

Los modelos de inteligencia artificial basados en Machine Learning son convincentes, pero plantean importantes problemas de seguridad: no siempre se comportan como se espera, a veces de maneras peligrosas. Este problema de seguridad es crucial en la IA porque estos sistemas se están volviendo cada vez más autónomos y aún son en gran medida impredecibles, a diferencia de otros objetos como los automóviles. Las empresas y gobiernos han invertido muy poco en la seguridad de la IA, pero esto podría ser uno de los principales desafíos del siglo XXI. La seguridad de la IA abarca tres problemas importantes: robustez, explicabilidad y definición de propósito.

Profundicemos un poco más en estos conceptos.

Problema N.°1: Falta de robustez en la IA

La robustez de un sistema de IA indica cuán confiable es su comportamiento en situaciones desconocidas, es decir, en casos que no ha encontrado durante el entrenamiento. Sin embargo, los sistemas de Machine Learning se basan en correlaciones estadísticas en lugar de comprender la realidad. Como resultado, cuando la realidad cambia y las correlaciones ya no son válidas, el sistema de IA puede reaccionar de manera inapropiada.

Ejemplo práctico: Supongamos que alguien entrena a un sistema de IA para detectar objetos en imágenes de perros. Se le ha proporcionado un conjunto de datos de entrenamiento que consta de imágenes de diferentes razas de perros. El sistema se desempeña bien y puede reconocer y clasificar correctamente las otras razas de perros en las imágenes proporcionadas.

Sin embargo, si de repente se le muestra una imagen de un animal desconocido que se parece a un perro, pero no es una raza enumerada en su conjunto de datos de entrenamiento, podría tener dificultades para clasificarlo correctamente. No comprende completamente la realidad subyacente de este animal desconocido. El sistema de IA podría tomar una decisión incorrecta o dar una respuesta impredecible.

La falta de robustez puede ser peligrosa. Esto se debe a que algunos ciberataques están diseñados para aprovechar esta vulnerabilidad al engañar al modelo de IA con pequeños ajustes en sus datos de entrada. Por ejemplo, un atacante podría usar calcomanías o pintura en señales de tráfico para dirigirse a la IA en vehículos autónomos y afectar cómo interpreta las señales. Esto podría hacer que el sistema de IA interprete incorrectamente un letrero de alto como un letrero de ceder, poniendo en riesgo a los pasajeros.

Este tipo de ataque se conoce como ataque adversario.

Las dos imágenes del letrero de ALTO se ven iguales a simple vista, pero algunos cambios invisibles a simple vista son suficientes para cambiar la interpretación del modelo de IA.

Problema N.°2: Falta de explicabilidad en la IA

La explicabilidad y la transparencia de un sistema de IA permiten a un humano comprender y analizar cómo funciona para garantizar que funcione de la manera deseada. Hoy en día, la mayoría de los sistemas de IA que utilizan Machine Learning son “cajas negras” que operan de manera autónoma sin que nadie sepa cómo ni por qué. Los sistemas de IA de Machine Learning utilizan datos y métodos de razonamiento estadístico para aprender correlaciones. Sin embargo, estas correlaciones no indican necesariamente una relación causal.

Por ejemplo, un sistema de Machine Learning en medicina examinará muchos casos diagnosticados previamente para establecer correlaciones en lugar de llegar a un diagnóstico médico aplicando conocimiento específico del dominio y reglas preconcebidas. Así es como algunos sistemas de IA para diagnosticar el cáncer han "aprendido" a distinguir entre imágenes de tumores malignos y benignos en función de si la imagen contiene una regla graduada. Entre las imágenes de tumores prediagnosticados que se les proporcionaron, las imágenes de tumores malignos a menudo contenían una regla graduada, que se utilizaba para medir el tamaño del tumor. ¡Se demostró la correlación entre la regla y el diagnóstico de cáncer!

En el caso del algoritmo automatizado de clasificación de lesiones cutáneas, una mayor explicabilidad y transparencia permiten a las personas comprender cómo funciona el modelo de IA y detectar cuándo no funciona correctamente.

Problema N.°3: Definir los objetivos correctos para un Sistema de IA

Cuando un modelo interactúa con personas, tiene que llevar a cabo acciones o tomar decisiones que afectan al mundo. Por lo tanto, debemos definir sus objetivos correctamente, o las acciones y decisiones no cumplirán nuestras expectativas.

Sin embargo, en la práctica, es difícil traducir la complejidad y el matiz de los objetivos humanos al lenguaje informático. Es fácil para una máquina malinterpretar la intención detrás de las instrucciones humanas al aplicarlas de manera demasiado literal.

Por ejemplo, al decirle a una IA que no pierda en un juego de Tetris, el modelo identificó la mejor forma de cumplir con la instrucción sin cumplir las expectativas de los diseñadores: pausa el juego tan pronto como cree que perderá. En un ejemplo menos trivial, muchas plataformas de video en línea intentan sugerir videos “que el usuario querrá ver” dándole a la IA el objetivo de ofrecer videos que el usuario verá de principio a fin.

Sin embargo, definir el objetivo de la IA de esta manera hace que el algoritmo favorezca videos cortos y sensacionales o refleje las opiniones fuertes del usuario. Es más probable que el usuario los vea de principio a fin, pero aún necesita ver los videos que más desea. ¿Cómo le explicas a una máquina lo que quieres decir con “videos que el usuario querrá ver”? No es tan simple.

¡Resumamos!

- La IA puede ser utilizada en tu contra con fines maliciosos. Utiliza tus habilidades de pensamiento crítico para protegerte de las noticias falsas y los ciberataques. Si algo te parece sospechoso, tómate un poco de tiempo para verificar y cruzar referencias de tus fuentes.

- La IA plantea muchos problemas de seguridad y no siempre se comporta como deseamos. La seguridad de la IA es un tema de investigación importante que está experimentando un crecimiento considerable. También se requiere gobernanza y regulación para limitar los riesgos asociados con la IA.

En el próximo capítulo, analizaremos los desafíos de responsabilidad social y ambiental en torno al uso de la inteligencia artificial.