Terminología y taxonomía de la inteligencia artificial

Aprendizaje automático adversario (ataque adversario)

Una práctica que se preocupa por el diseño de algoritmos de aprendizaje automático que pueden resistir desafíos de seguridad, el estudio de las capacidades de los atacantes y la comprensión de las consecuencias de los ataques. En el aprendizaje automático adversarial, las entradas están diseñadas intencionadamente para inducir un error en las predicciones, a pesar de que parezcan entradas válidas para un humano.

Sistema de IA autónomo

Se refiere a sistemas que mantienen un conjunto de capacidades basadas en inteligencia para responder a situaciones que no fueron programadas o anticipadas previamente antes de la implementación del sistema. Los sistemas autónomos tienen un grado de autogobierno y comportamiento autodirigido.

Big data

Un término que engloba conjuntos de datos digitales grandes y complejos que requieren medios tecnológicos igualmente complejos para su almacenamiento, análisis, gestión y procesamiento, junto con una potencia informática considerable. A veces, los conjuntos de datos se vinculan para ver cómo los patrones en un dominio afectan a otras áreas. Los datos pueden estar estructurados en campos fijos o ser no estructurados como información de flujo libre. El análisis de

grandes conjuntos de datos, a menudo utilizando la inteligencia artificial, puede revelar patrones, tendencias o relaciones subyacentes que no eran aparentes previamente para los investigadores.

Clasificador

Un modelo que predice (o asigna) etiquetas de clase a datos de entrada.

Envenenamiento de datos (data poisoning)

Un tipo de ataque de seguridad en el cual usuarios malintencionados inyectan datos de entrenamiento falsos con el objetivo de corromper el modelo aprendido, haciendo que el sistema de IA aprenda algo que no debería aprender.

Aprendizaje profundo (deep learning)

Un subconjunto del aprendizaje automático basado en redes neuronales artificiales que utiliza estadísticas para identificar tendencias subyacentes o patrones de datos y aplica ese conocimiento a otras capas de análisis. Algunos lo han calificado como una forma de «aprender por ejemplo» y como una técnica que «realiza tareas de clasificación directamente a partir de imágenes, texto o sonido» y luego aplica ese conocimiento de manera independiente.

Privacidad diferencial (differential privacy)

La privacidad diferencial es un método para medir cuánta información revela la salida de una computación sobre un

individuo. Produce resultados de análisis de datos que son casi igualmente probables, ya sea que un individuo esté incluido o no en el conjunto de datos. Su objetivo es oscurecer la presencia o ausencia de cualquier individuo (en una base de datos) o pequeños grupos de individuos, al tiempo que preserva la utilidad estadística.

Datos de entrada (input data)

Datos proporcionados o adquiridos directamente por un sistema de IA en base a los cuales el sistema genera una salida.

Aprendizaje automático (machine learning)

El Aprendizaje Automático es una rama de la inteligencia artificial (IA) y la informática que se enfoca en el desarrollo de sistemas capaces de aprender y adaptarse sin seguir instrucciones explícitas, imitando la forma en que los humanos aprenden, mejorando gradualmente su precisión mediante algoritmos y modelos estadísticos para analizar e inferir patrones en los datos.

Entrenamiento de modelo (model training)

Proceso para establecer o mejorar los parámetros de un modelo de aprendizaje automático, basado en un algoritmo de aprendizaje automático, utilizando datos de entrenamiento.

Validación de modelo (model validation)

Confirmación a través de la presentación de pruebas objetivas de que se han cumplido los requisitos para un uso o aplicación

específica prevista.

Procesamiento de lenguaje natural (natural language processing)

La capacidad de una máquina para procesar, analizar y simular el lenguaje humano, ya sea hablado o escrito.

Análisis predictivo (predictive analysis)

La organización de análisis de datos estructurados y no estructurados para inferencia y correlación que proporciona una capacidad predictiva útil para nuevas circunstancias o datos.

Perfilado (profiling)

‘Perfilado’ significa cualquier forma de procesamiento automatizado de datos personales que consiste en el uso de datos personales para evaluar ciertos aspectos personales relacionados con una persona natural, en particular para analizar o predecir aspectos relacionados con el rendimiento laboral, situación económica, salud, preferencias personales, intereses, confiabilidad, comportamiento, ubicación o movimientos de esa persona natural.

Aprendizaje por refuerzo (reinforcement learning)

Un tipo de aprendizaje automático en el que el algoritmo aprende actuando hacia un objetivo abstracto, como «obtener una alta puntuación en un videojuego» o «administrar una fábrica eficientemente». Durante el entrenamiento, cada esfuerzo se evalúa en función de su contribución hacia el

objetivo.

Datos estructurados (structured data)

Datos que tienen un modelo de datos predefinido o están organizados de una manera predefinida.

Datos no estructurados (unstructured data)

Datos que no tienen un modelo de datos predefinido ni están organizados de una manera predefinida.

Datos sintéticos (synthetic data)

Los datos sintéticos se generan a partir de datos/procesos y un modelo que está entrenado para reproducir las características y la estructura de los datos originales con el objetivo de lograr una distribución similar. El grado en que los datos sintéticos son un sustituto preciso de los datos originales es una medida de la utilidad del método y del modelo.

Aprendizaje por transferencia (transfer learning)

Una técnica en el aprendizaje automático en la que un algoritmo aprende a realizar una tarea, como reconocer automóviles, y basa ese conocimiento al aprender una tarea diferente pero relacionada, como reconocer gatos.

Aprendizaje supervisado (supervised learning)

Aprendizaje automático que utiliza datos etiquetados durante el entrenamiento.

Aprendizaje no supervisado (unsupervised learning)

Aprendizaje automático que utiliza datos no etiquetados durante el entrenamiento.

Precisión (AI) (accuracy)

La cercanía de los cálculos o estimaciones a los valores exactos o verdaderos que las estadísticas tenían la intención de medir. El objetivo de un modelo de IA es aprender patrones que se generalicen bien para datos no vistos. Es importante verificar si un modelo de IA entrenado se desempeña bien en ejemplos no vistos que no se utilizaron para entrenar el modelo. Para hacer esto, el modelo se utiliza para predecir la respuesta en el conjunto de datos de prueba y luego se compara el objetivo predicho con la respuesta real. El concepto de precisión se utiliza para evaluar la capacidad predictiva del modelo de IA. Informalmente, la precisión es la fracción de predicciones que el modelo acertó. Se utilizan varias métricas en el aprendizaje automático (ML) para medir la precisión predictiva de un modelo. La elección de la métrica de precisión a utilizar depende de la tarea de ML.

Prueba (test)

Operación técnica para determinar una o más características de un producto, material, equipo, organismo, fenómeno físico, proceso o servicio, de acuerdo con un procedimiento

especificado.

Evaluación (evaluation)

Determinación sistemática del grado en que una entidad cumple con sus criterios especificados.

Verificación (verification)

Proporciona evidencia de que el sistema o elemento del sistema realiza sus funciones previstas y cumple con todos los requisitos de rendimiento enumerados en la especificación de rendimiento del sistema.

Validación (validation)

Confirmación mediante examen y presentación de pruebas objetivas de que se cumplen los requisitos particulares para un uso específico previsto.

Prueba y evaluación, verificación y validación

Un marco para evaluar, que incorpora métodos y métricas para determinar que una tecnología o sistema cumple satisfactoriamente con sus especificaciones de diseño y requisitos, y que es adecuado para su uso previsto.

Aprendizaje adaptativo

El aprendizaje adaptativo se refiere a la capacidad de un sistema de inteligencia artificial para modificar su comportamiento mientras está en funcionamiento. Esta adaptación puede implicar ajustes en los pesos del modelo o incluso cambios en su estructura interna. El comportamiento resultante del sistema adaptado puede generar resultados

diferentes a los de su estado anterior cuando se le presenta la misma entrada.

Algoritmo

Un algoritmo consiste en una serie de instrucciones sistemáticas diseñadas para resolver un problema, sin incluir datos. Este conjunto abstracto de pasos se puede implementar en diversos lenguajes de programación y bibliotecas de software.

Clasificación

La clasificación implica organizar objetos en «categorías» o grupos distintos. Las clasificaciones son coherentes, poseen principios de clasificación únicos y son mutuamente excluyentes. En el diseño de inteligencia artificial, cuando la salida pertenece a un conjunto finito de valores (por ejemplo, soleado, nublado o lluvioso), el desafío de aprendizaje se denomina clasificación. Se le llama clasificación booleana o binaria cuando solo hay dos valores involucrados.

Aprendizaje federado

El aprendizaje federado es un modelo de aprendizaje automático diseñado para abordar problemas de gobernanza de datos y privacidad. Lo logra al entrenar algoritmos de manera colaborativa sin transferir los datos a una ubicación externa. Cada dispositivo federado comparte los parámetros de su modelo local en lugar de revelar todo el conjunto de datos utilizado para su entrenamiento. La topología del aprendizaje federado dicta la forma en que se comparten estos parámetros.

Red Generativa Adversativa (GAN)

Las Redes Generativas Adversativa, o GANs en resumen, representan un enfoque de modelado generativo que utiliza técnicas de aprendizaje profundo, como las redes neuronales convolucionales. El modelado generativo es una tarea de aprendizaje no supervisado en el campo del aprendizaje automático, que implica descubrir y aprender automáticamente patrones o regularidades en los datos de entrada. Esto permite que el modelo genere ejemplos nuevos que se asemejen de manera convincente a los del conjunto de datos original.

Valores humanos para la inteligencia artificial

Los valores son cualidades o condiciones idealizadas en el mundo que las personas consideran deseables. Los sistemas de inteligencia artificial no son neutrales en cuanto a valores. La incorporación de valores humanos en los sistemas de inteligencia artificial implica la identificación de la extensión, la forma y el propósito que queremos que la inteligencia artificial tenga en nuestras sociedades. Esto implica establecer principios éticos, políticas de gobernanza, incentivos y regulaciones. También implica ser conscientes de las diferencias en intereses y objetivos que subyacen a los sistemas de inteligencia artificial desarrollados por otros, basados en diferentes culturas y principios.

Inteligencia artificial centrada en el ser humano

Un enfoque de la IA que prioriza la responsabilidad ética humana, las cualidades dinámicas, la comprensión y el significado. Fomenta el empoderamiento de los seres humanos en el diseño, uso y implementación de sistemas de IA. Los

sistemas de IA centrados en el ser humano se basan en el reconocimiento de una interacción significativa entre humanos y tecnología. Están diseñados como componentes de entornos sociotécnicos en los que los seres humanos asumen una agencia significativa. La IA centrada en el ser humano no se concibe como un fin en sí misma, sino como herramientas para servir a las personas con el objetivo final de aumentar el bienestar humano y ambiental, respetando el estado de derecho, los derechos humanos, los valores democráticos y el desarrollo sostenible.

Modelo de lenguaje

Un modelo de lenguaje es una descripción aproximada que captura patrones y regularidades presentes en el lenguaje natural y se utiliza para hacer suposiciones sobre fragmentos de lenguaje previamente no vistos.

Modelo grande de lenguaje (MLG)

Una clase de modelos de lenguaje que utilizan algoritmos de aprendizaje profundo y se entrenan en conjuntos de datos textuales extremadamente grandes que pueden ser de varios terabytes de tamaño. Los MLG se pueden clasificar en dos tipos: generativos o discriminatorios. Los MLG generativos son modelos que generan texto, como respuestas a preguntas o incluso escriben un ensayo sobre un tema específico. Suelen ser modelos de aprendizaje no supervisado o semisupervisado que predicen cuál es la respuesta para una tarea dada. Los MLG discriminatorios son modelos de aprendizaje supervisado que suelen centrarse en clasificar texto, como determinar si un texto fue creado por un humano o por una IA.

Modelo

Una función que toma características como entrada y predice

etiquetas como salida. Las fases típicas del flujo de trabajo de un modelo de IA son: recopilación y preparación de datos, desarrollo del modelo, entrenamiento del modelo, evaluación de la precisión del modelo, ajuste de parámetros, uso del modelo, mantenimiento del modelo y versionado del modelo.

Red neuronal

Un sistema informático inspirado en cerebros vivos, también conocido como red neuronal artificial, red neuronal o red neuronal profunda. Consiste en dos o más capas de neuronas conectadas por enlaces ponderados con pesos ajustables, que toman datos de entrada y producen una salida. Mientras que algunas redes neuronales están diseñadas para simular el funcionamiento de las neuronas biológicas en el sistema nervioso, la mayoría de las redes neuronales se utilizan en inteligencia artificial como implementaciones del modelo conexionista.

Escalabilidad

La capacidad de aumentar o disminuir los recursos computacionales necesarios para ejecutar un volumen variable de tareas, procesos o servicios.

Sistema sociotécnico

La tecnología siempre es parte de la sociedad, al igual que la sociedad siempre es parte de la tecnología. Esto significa que no se puede entender una sin la otra. La tecnología no es solo diseño y apariencia material, sino también sociotécnica; es decir, un proceso complejo constituido por diversos factores sociales, políticos, económicos, culturales y tecnológicos.

Interoperabilidad técnica

La capacidad de sistemas o componentes de software o hardware para operar juntos con éxito con un esfuerzo mínimo por parte del usuario final.

Diseño sensible a los valores

Un enfoque teóricamente fundamentado para el diseño de tecnología que tiene en cuenta los valores humanos de manera sistemática y principiada a lo largo del proceso de diseño.

Bias en la Inteligencia Artificial (IA) (o Sesgo algorítmico)

El sesgo perjudicial en la IA describe errores sistemáticos y repetibles en los sistemas de IA que crean resultados injustos, como dar ventaja sistemática a grupos privilegiados y desventaja sistemática a grupos no privilegiados. Diferentes tipos de sesgo pueden surgir y interactuar debido a muchos factores, incluyendo, pero no limitado a, decisiones y procesos humanos o del sistema en todo el ciclo de vida de la IA. El sesgo puede estar presente en los sistemas de IA debido a expectativas culturales, sociales o institucionales preexistentes; debido a limitaciones técnicas de su diseño; por ser utilizados en contextos no anticipados; o por especificaciones de diseño no representativas.

Chatbot (Bot de Conversación)

Un programa de computadora diseñado para simular una conversación con un usuario humano, generalmente a través de Internet; especialmente uno utilizado para proporcionar información o asistencia al usuario como parte de un servicio automatizado.

Trazabilidad

Capacidad para rastrear el recorrido de una entrada de datos a través de todas las etapas de muestreo, etiquetado, procesamiento y toma de decisiones.

IA de confianza

La IA de confianza tiene tres componentes: (1) debe ser legal, garantizando el cumplimiento de todas las leyes y regulaciones aplicables; (2) debe ser ética, demostrando respeto y asegurando la adhesión a principios y valores éticos; y (3) debe ser robusta, tanto desde una perspectiva técnica como social, ya que, incluso con buenas intenciones, los sistemas de IA pueden causar daños no intencionales. Las características de los sistemas de IA de confianza incluyen ser válidos y fiables, seguros, seguros y resistentes, responsables y transparentes, explicables e interpretables, mejorados en privacidad y justos, con sesgos perjudiciales gestionados. La IA de confianza no se refiere solo a la confiabilidad del propio sistema de IA, sino que también abarca la confiabilidad de todos los procesos y actores que forman parte del ciclo de vida del sistema de IA.