zertia

Términos clave para **Gobernanza de la** IA

zertia

Términos clave para **Gobernanza de la** IA

El campo de la IA está evolucionando rápidamente en sectores e industrias muy diversos, dejando a profesionales de negocios, tecnología y gobierno sin un léxico común ni una comprensión compartida de los términos y expresiones empleados en la gobernanza de IA. Incluso una búsqueda para definir "inteligencia artificial" devuelve una variedad de definiciones y ejemplos. Desde los más cinematográficos, como HAL 9000 de 2001: Una odisea del espacio, pasando por lo creativo, como el arte generativo de Midjourney y DALL-E, hasta lo cotidiano, como la autocorrección de correos electrónicos y los mapas en dispositivos móviles, los casos de uso y aplicaciones de la IA continúan creciendo y expandiéndose a todos los aspectos de la vida.

Nuestro glosario de términos clave es una adaptación de la versión publicada en julio de 2024 de "Key Terms for AI Governance" de la IAPP. Esta adaptación ha sido creada para ofrecer una referencia accesible y práctica sobre conceptos fundamentales en la gobernanza de IA, manteniendo la calidad y el rigor de la publicación original. Las explicaciones buscan presentar tanto perspectivas políticas como técnicas, aportando al sólido debate en curso sobre la gobernanza de IA.

zertia

Algoritmo (Algorithm)

Un procedimiento o conjunto de instrucciones y reglas diseñadas para realizar una tarea específica o resolver un problema particular mediante una computadora.

Algoritmos codiciosos (Greedy algorithms)

Un tipo de algoritmo que toma la decisión óptima en función de la información disponible para alcanzar un objetivo inmediato en un paso o punto de decisión específico, sin considerar la solución óptima a largo plazo

Alucinaciones (Hallucinations)

Instancias en las que los modelos de IA generativa producen resultados que, aunque parecen plausibles, son incorrectos desde un punto de vista fáctico. También conocido como confabulaciones.

Aprendizaje adaptativo (Adaptive learning)

Método que ajusta y personaliza el contenido educativo según las necesidades específicas, habilidades y ritmo de aprendizaje de cada estudiante. Su propósito es ofrecer una experiencia de aprendizaje personalizada y optimizada, adaptándose a los distintos estilos de aprendizaje.

Aprendizaje activo (Active learning)

Subcampo de la IA y el aprendizaje automático en el que un algoritmo selecciona parte de los datos de los que aprende. En lugar de utilizar todos los datos disponibles,, un modelo de aprendizaje activo solicita puntos de datos adicionales que le ayudarán a aprender de la mejor manera.

Aprendizaje automático (Machine learning)

Subcampo de la IA que involucra algoritmos que aprenden de manera iterativa y luego toman decisiones, recomendaciones, inferencias o predicciones basadas en datos de entrada. Estos algoritmos construyen un modelo a partir de datos de entrenamiento para realizar una tarea específica en nuevos datos sin estar explícitamente programados para hacerlo.

zertia

Aprendizaje de refuerzo con retroalimentación humana (Reinforcement learning with human feedback)

Proceso que combina la técnica de aprendizaje de refuerzo con retroalimentación humana durante el proceso de entrenamiento. Se proporciona retroalimentación humana sobre la salida del modelo, generalmente comparando diferentes salidas para la misma indicación y seleccionando la que mejor se ajusta con las preferencias humanas.

→ Acrónimo: RLHF.

Aprendizaje federado (Federated learning)

Método de aprendizaje automático que permite entrenar modelos (ver también modelo de aprendizaje automático) utilizando en los datos locales de múltiples dispositivos. Solo se envían al servidor central las actualizaciones del modelo local, no los datos de entrenamiento en sí mismos.

Aprendizaje no supervisado (Unsupervised learning)

Subconjunto del aprendizaje automático en el que el modelo se entrena buscando patrones en un conjunto de datos no clasificado con una supervisión humana mínima.

Aprendizaje por refuerzo (Reinforcement learning)

Método de aprendizaje automático que entrena un modelo para optimizar sus acciones dentro de un entorno dado con el fin de alcanzar un objetivo específico, guiado por mecanismos de retroalimentación de recompensas y sanciones.

AutomatAsesoría de IA (AI assurance) ed decision-making

Conjunto de marcos, políticas, procesos y controles para medir, evaluar y promover una IA segura, confiable y de confianza. Los esquemas de aseguramiento de IA pueden incluir evaluaciones de conformidad, impacto y riesgos, auditorías de IA, certificaciones, pruebas y evaluaciones, asi como el cumplimiento de normas pertinentes.

zertia

Auditoría de IA (AI audit)

Revisión y evaluación de un sistema de IA para garantizar que funcione según lo previsto y cumpla con las leyes, regulaciones y estándares relevantes.

Automatización de decisiones (Automated decision-making)

Proceso mediante el cual se toman decisiones a traves de medios tecnológicos sin la intervención humana, ya sea en su totalidad o en parte.

Base de conocimientos (Corpus)

Colección extensa de textos o datos que una computadora utiliza para identificar patrones, hacer predicciones o generar resultados específicos. Puede incluir datos estructurados como no estructurados.

Bosque aleatorio (Random forest)

Algoritmo supervisado de aprendizaje automático que construye múltiples árboles de decisión y los combina para obtener una predicción más precisa y estable.

Calidad de datos (Data quality)

Medida en la que un conjunto de datos cumple con los requisitos y expectativas específicos para el uso previsto.

Centrado en el ser humano (Human-centric AI)

Enfoque en el diseño, desarrollo, implementación y uso de IA que prioriza el bienestar, la autonomía, los valores y las necesidades humanas.

zertia

Chatbot

Sistema de IA diseñado para simular conversaciones e interacciones similares a las humanas que utiliza el procesamiento de lenguaje natural y el aprendizaje profundo.

Clasificación (Classification model)

Tipo de modelo utilizado en el aprendizaje automático,diseñado para clasificar datos de entrada en diferentes categorías o clases.

Computación (Compute)

Recursos de procesamiento disponibles en un sistema informático. Incluye componentes de hardware como la unidad central de procesamiento (CPU) o la unidad de procesamiento gráfico (GPU).

Confiabilidad (Reliability)

Aitributo de un sistema de IA que garantiza que se comporte como se espera y realice su función prevista de manera consistente y precisa.

Confirmidad (Conformity assessment)

Análisis, a menudo generalmente por una entidad independiente, de un sistema de IA para determinar si se han cumplido los requisitos relacionados con la gestion de riesgos , gobernanza de datos, registro, transparencia y prácticas de ciberseguridad.

Contestabilidad (Contestability)

El principio que asegura que los sistemas de IA y sus procesos de toma de decisiones pueden ser cuestionados o impugnados por los humanos.

Datos de entrada (Input data)

Datos proporcionados a o adquiridos directamente por un algoritmo o modelo de aprendizaje para generar una salida.

zertia

Datos de prueba (Testing data)

El conjunto de datos utilizado para probar y evaluar un modelo que ya ha sido entrenado

Datos de entrenamiento (Training data)

El conjunto de datos utilizado para entrenar un modelo permitiéndole predecir resultados con precisión.

Datos sintéticos (Synthetic data)

Datos generados por un sistema o modelo que imitan las estructuras y propiedades estadísticas de datos reales.

Desinformación (Disinformation)

Contenido de audio o visual que se manipula o crea intencionadamente con el propósito de causar daño.

Diferenciación (Diffusion model)

Un modelo generativo utilizado en la creación de imágenes que refina iterativamente una señal de ruido para transformarla en una imagen realista.

Equidad (Fairness)

Un atributo de un sistema de IA que asegura un tratamiento equitativo y justo de individuos o grupos.

Escasez de ajuste (Underfitting)

Ocurre cuando un modelo de aprendizaje automático no es capaz de capturar patrones subyacentes de los datos de entrenamiento, lo cual resulta en un bajo rendimiento del modelo.

zertia

Evaluación de conformidad (Conformity assessment)

Un análisis, generalmente realizado por una entidad independiente, de un sistema de IA para verificar si cumple con los requisitos relacionados con la gestión de riesgos, gobernanza de datos, registro, transparencia y prácticas de ciberseguridad.

Evaluación de impacto (Impact assessment)

Un proceso de evaluación diseñado para identificar, comprender, documentar y mitigar las posibles implicaciones éticas, legales, económicas y sociales de un sistema de IA.

Explicabilidad (Explainability)

La capacidad de describir o proporcionar información suficiente sobre cómo un sistema de IA genera un resultado específico, permitiendo comprender las razones detrás de las decisiones o acciones del sistema.

Gobernanza de IA (AI governance)

Un sistema de leyes, políticas, marcos, prácticas y procesos a nivel internacional, nacional y organizacional para gestionar y regular el uso y desarrollo de la inteligencia artificial.

Ingeniería de indicaciones (Prompt engineering)

El proceso deliberado de estructurar una indicación o serie de indicaciones con el fin de influir en el comportamiento del modelo de IA, logrando generar salidas más precisas o deseables.

Inferencia (Inference)

Un proceso en el aprendizaje automático donde se utiliza un modelo entrenado para hacer predicciones o decisiones basadas en nuevos datos de entrada.

zertia

Inteligencia artificial (Artificial intelligence)

Un término amplio que que describe un un sistema diseñado para realizar o automatizar tareas utilizando diversas técnicas computacionales.

Inteligencia artificial general (Artificial general intelligence)

IA que se considera que tiene un nivel de inteligencia humano y una gran capacidad de generalización para alcanzar objetivos.

Interpretabilidad (Interpretability)

La capacidad de explicar o presentar el razonamiento de un modelo de IA de manera comprensible para los seres humanos.

Marcado de agua (Watermarking)

El proceso de incrustar patrones sutiles o visualmente imperceptibles en contenido generado por IA, con el fin de identificar o rastrear el origen del contenido.

Modelo de lenguaje grande (Large language model)

Una forma de IA que utiliza algoritmos de aprendizaje profundo para crear modelos preentrenados en grandes volúmenes de datos textuales, capaces de generar o comprender texto en lenguaje natural.

Modelo discriminativo (Discriminative model)

Un tipo de modelo utilizado en el aprendizaje automático que asigna directamente características de entrada a etiquetas de clase y analizando patrones para distinguir entre diferentes clases.

Modelo fundacional (Foundation model)

Un modelo de gran escala que ha sido entrenado utilizando extensos y diversos conjuntos de datos, y que puede ser adaptado a múltiples tareas de IA sin necesidad de ser entrenado desde cero para cada una.

zertia

Modelo multimodal (Multimodal models)

Un tipo de modelo de aprendizaje automático capaz de procesar y combinar diferentes tipos de datos de entrada o salida simultáneamente, como texto, imágenes, audio, etc.

Posprocesamiento (Post processing)

Acciones realizadas después de ejecutar un modelo de aprendizaje automático para ajustar, refinar o modificar su salida antes de su uso final.

Precisión (Accuracy)

El grado en que un sistema de IA realiza correctamente la tarea para la que ha sido diseñado, evaluando la proporción de resultados correctos en relación con el total de casos.

Preprocesamiento (Preprocessing)

Etapas previas a la fase de entrenamiento de un modelo de aprendizaje automático, donde se preparan y organizan los datos, incluyendo la limpieza y normalización, para garantizar que el modelo reciba la mejor entrada posible.

Procesamiento de lenguaje natural (Natural language processing)

Un subcampo de la inteligencia artificial que permite que las computadoras comprendan, interpreten y generen lenguaje humano de manera que puedan interactuar de forma natural con los usuarios

Red de equipo (Red teaming)

El proceso de evaluar la seguridad, el rendimiento y la confiabilidad de un sistema de IA desde una perspectiva adversaria, simulando posibles ataques o fallos para identificar vulnerabilidades y áreas de mejora.

zertia

Redes neuronales (Neural networks)

Un tipo de modelo de aprendizaje automático basado en la estructura de neuronas artificiales que imita el funcionamiento del cerebro humano. Estas redes utilizan múltiples capas de procesamiento para aprender patrones complejos a partir de los datos.

Responsabilidad (Accountability)

La obligación de los desarrolladores y responsables del despliegue de un sistema de IA de garantizar que el sistema funcione de manera ética, justa y conforme a las regulaciones, asumiendo las consecuencias de sus decisiones y acciones.

Robustez (Robustness)

Un atributo de un sistema de IA que describe su capacidad para resistir y recuperarse de ataques de seguridad, manteniendo su integridad y funcionamiento

Seguridad (Safety)

Un término amplio que puede referirse al diseño, desarrollo y despliegue de sistemas de IA con el objetivo de minimizar los riesgos y daños, asegurando que el sistema opere de manera segura y controlada en diversas condiciones.

Sistema experto (Expert system)

Un tipo de IA basado en un conjunto de reglas que permite extraer inferencias a partir de una base de conocimiento proporcionada por expertos humanos, con el fin de resolver problemas específicos o tomar decisiones.

Supervisión (Oversight)

El proceso de monitorear, revisar y controlar de manera efectiva un sistema de IA para garantizar que se minimicen los riesgos, se cumplan las regulaciones pertinentes y se mantengan prácticas responsables en su uso.

zertia

Tarjeta del modelo (Model card)

Un documento conciso que proporciona información sobre un modelo de IA, incluyendo detalles sobre su entrenamiento, desempeño, limitaciones y posibles riesgos, con el objetivo de aumentar la transparencia y la comprensión del modelo

Transparencia (Transparency)

Un término amplio que implica que los sistemas de IA sean abiertos, comprensibles y responsables en cuanto a cómo funcionan, cómo toman decisiones y cómo se utilizan los datos en su desarrollo.

Transferencia de aprendizaje (Transfer learning)

Un enfoque en el aprendizaje automático en el que un modelo entrenado para realizar una tarea específica utiliza el conocimiento adquirido para facilitar el aprendizaje de una tarea diferente, aprovechando las similitudes entre ambas.

Variable (Variable)

En el contexto del aprendizaje automático, una variable es un atributo, característica o unidad medible que puede asumir diferentes valores durante el proceso de análisis o entrenamiento del modelo.

Varianza (Variance)

Una medida que indica el grado de dispersión de un conjunto de datos con respecto a su valor promedio. Cuanto mayor es la varianza, mayor es la dispersión de los datos.