

Pregunta 1:

Enunciado:

Utilice un modelo grande de lenguaje (LLM) —tal como ChatGPT, Bard o Bing— para obtener una descripción breve de los siguientes términos:

- Big data
- Machine learning
- Inteligencia artificial
- Ciencia de datos (data science)
- Deep learning
- Minería de datos (data mining)
- Inteligencia

Prompt:

Eres un experto en inteligencia artificial y aprendizaje automático. Por favor, proporciona definiciones breves, precisas y claras de los siguientes términos, adecuadas para un curso universitario introductorio de Aprendizaje de Máquina aplicado:

- Big data
- Machine learning
- Inteligencia artificial
- Ciencia de datos (data science)
- Deep learning
- Minería de datos (data mining)
- Inteligencia

Para cada término, sigue estas instrucciones:

- Explica el concepto claramente.
- Sé conciso (máximo 3-4 líneas por término).
- Indica la relación jerárquica entre los conceptos cuando sea relevante (por ejemplo, si uno es superconjunto de otro).
- Por favor, presenta la información en formato de lista ordenada por término.

Respuesta

Definiciones de Conceptos Clave

- **Big Data**

Se refiere a conjuntos de datos tan grandes, rápidos y complejos (caracterizados por las "3 Vs": Volumen, Velocidad y Variedad) que es imposible procesarlos con herramientas tradicionales. No es una técnica en sí, sino el desafío o la materia prima que impulsa la necesidad de tecnologías y enfoques como los del Machine Learning a gran escala. Relación: Es el contexto o el insumo que a menudo se utiliza en Ciencia de Datos y Machine Learning.

- **Machine Learning (ML) o Aprendizaje de Máquina**

Es una subdisciplina de la IA que se centra en desarrollar algoritmos que permiten a las computadoras aprender directamente de los datos. En lugar de ser programados explícitamente para una tarea, estos sistemas utilizan patrones en los datos para hacer predicciones o tomar decisiones. Relación: Es la principal herramienta para lograr la IA en la práctica actualmente.

- **Inteligencia Artificial (IA)**

Es el campo más amplio de la ciencia computacional enfocado en crear máquinas capaces de realizar tareas que normalmente requieren inteligencia humana, como el razonamiento, el aprendizaje y la resolución de problemas. La IA es la disciplina general que busca simular la cognición humana. Relación: Es el superconjunto que engloba al Machine Learning y al Deep Learning.

- **Ciencia de Datos (Data Science)**

Es un campo interdisciplinario que utiliza métodos científicos, procesos y algoritmos para extraer conocimiento e insights de datos estructurados y no estructurados. Abarca todo el proceso de datos: desde la recolección y limpieza hasta el análisis, modelado (usando ML), visualización y comunicación de resultados para la toma de decisiones. Relación: Utiliza el Machine Learning y la Minería de Datos como herramientas fundamentales, pero es un campo más amplio que incluye estadística, dominio de negocio y comunicación.

- **Deep Learning (DL) o Aprendizaje Profundo**

Es un subcampo especializado del Machine Learning basado en redes neuronales artificiales con múltiples capas (de ahí el término "profundo"). Estos modelos son especialmente potentes para reconocer patrones complejos en grandes volúmenes de datos, como imágenes, sonido y texto, siendo la tecnología detrás de los avances más recientes en IA. Relación: Es una técnica avanzada dentro del Machine Learning.

- **Minería de Datos (Data Mining)**

Es el proceso de descubrir patrones, anomalías y correlaciones en grandes conjuntos de datos para predecir resultados. Se considera un paso crucial dentro del proceso más amplio de la Ciencia de Datos y emplea muchas técnicas del Machine Learning para encontrar información útil que no es aparente a simple vista. Relación: Es una actividad específica dentro de la Ciencia de Datos. A menudo, los términos se usan de forma intercambiable con algunas técnicas de ML supervisado y no supervisado.

- Inteligencia

Es la capacidad fundamental de adquirir, comprender y aplicar conocimiento y habilidades para resolver problemas, adaptarse a nuevas situaciones y razonar. Este concepto cognitivo es el objetivo que la Inteligencia Artificial busca replicar o simular en sistemas no biológicos. Relación: Es el concepto base que inspira el campo de la Inteligencia Artificial.

Pregunta 2

Enunciado

Utilice un LLM para crear un diagrama que muestre las relaciones entre los anteriores conceptos. Pídale que genere el diagrama usando la sintaxis de mermaid.js y visualicelo usando <https://mermaid.live/>. Opcionalmente puede usar un plugin. Discuta el resultado, ¿tiene sentido? Pruebe dándole instrucciones al LLM para que lo mejore.

Prompt

Primer prompt

Eres un experto en inteligencia artificial y visualización de conceptos. Usando la sintaxis de mermaid.js, crea un diagrama que muestre las relaciones entre los siguientes conceptos: Big data, Machine learning, Inteligencia artificial, Ciencia de datos (data science), Deep learning, Minería de datos (data mining) e Inteligencia.

Utiliza la representación de superconjuntos y subconjuntos para indicar jerarquías claras entre los conceptos.

Por favor, entrega el código mermaid en formato texto, listo para visualizar en <https://mermaid.live/>.

Segundo prompt

Eres un experto en visualización conceptual usando mermaid.js. Toma el diagrama entregado anteriormente sobre inteligencia artificial y conceptos relacionados (definido en formato mermaid) y mejora su claridad y estructura:

Simplifica el diagrama priorizando relaciones jerárquicas (superconjuntos y subconjuntos) entre Inteligencia, Inteligencia Artificial, Machine Learning, Deep Learning, Ciencia de Datos, Minería de Datos y Big Data.

Usa flechas sólidas para jerarquías y otro tipo para relaciones de inclusión o solapamiento, añade una leyenda explicativa.

Minimiza el texto en etiquetas y recuadros para mejor lectura rápida.

Reorganiza el diagrama para un layout jerárquico claro, preferiblemente en forma de árbol.

Explica con etiquetas las relaciones esenciales, evitando redundancia o confusión.

Opcionalmente, sugiere mejoras estéticas o de formato para hacerlo más amigable visualmente.

Por favor entrega solo el código mermaid mejorado listo para visualizar en <https://mermaid.live/>.

Respuesta

Claro, aquí tienes el código mermaid.js que visualiza las relaciones jerárquicas y funcionales entre los conceptos solicitados, listo para ser utilizado.

```
%% @theme default
graph TD
    %% Definición de Estilos para Claridad Visual
    classDef aiConcept fill:#e3f2fd,stroke:#1565c0,stroke-width:2px,color:#0d47a1
    classDef dataField fill:#e8f5e9,stroke:#2e7d32,stroke-width:2px,color:#1b5e20
    classDef resource fill:#fffde7,stroke:#f57f17,stroke-width:2px,color:#f57f17
    classDef abstract fill:#f3e5f5,stroke:#4a148c,stroke-width:2px,color:#4a148c
    classDef legend_compact fill:#f8f9fa,stroke:#dee2e6,stroke-width:1px,color:#495057

    %% Concepto Abstracto Origen
    I[Inteligencia]:::abstract

    %% Jerarquía Principal anidada para mostrar subconjuntos
    subgraph SG_AI [Inteligencia Artificial]
        subgraph SG_ML [Machine Learning]
            DL[Deep Learning]
        end
    end
    end
```

```

%% Campo Aplicado
subgraph SG_DS [Ciencia de Datos]
    DM(Minería de Datos)
end

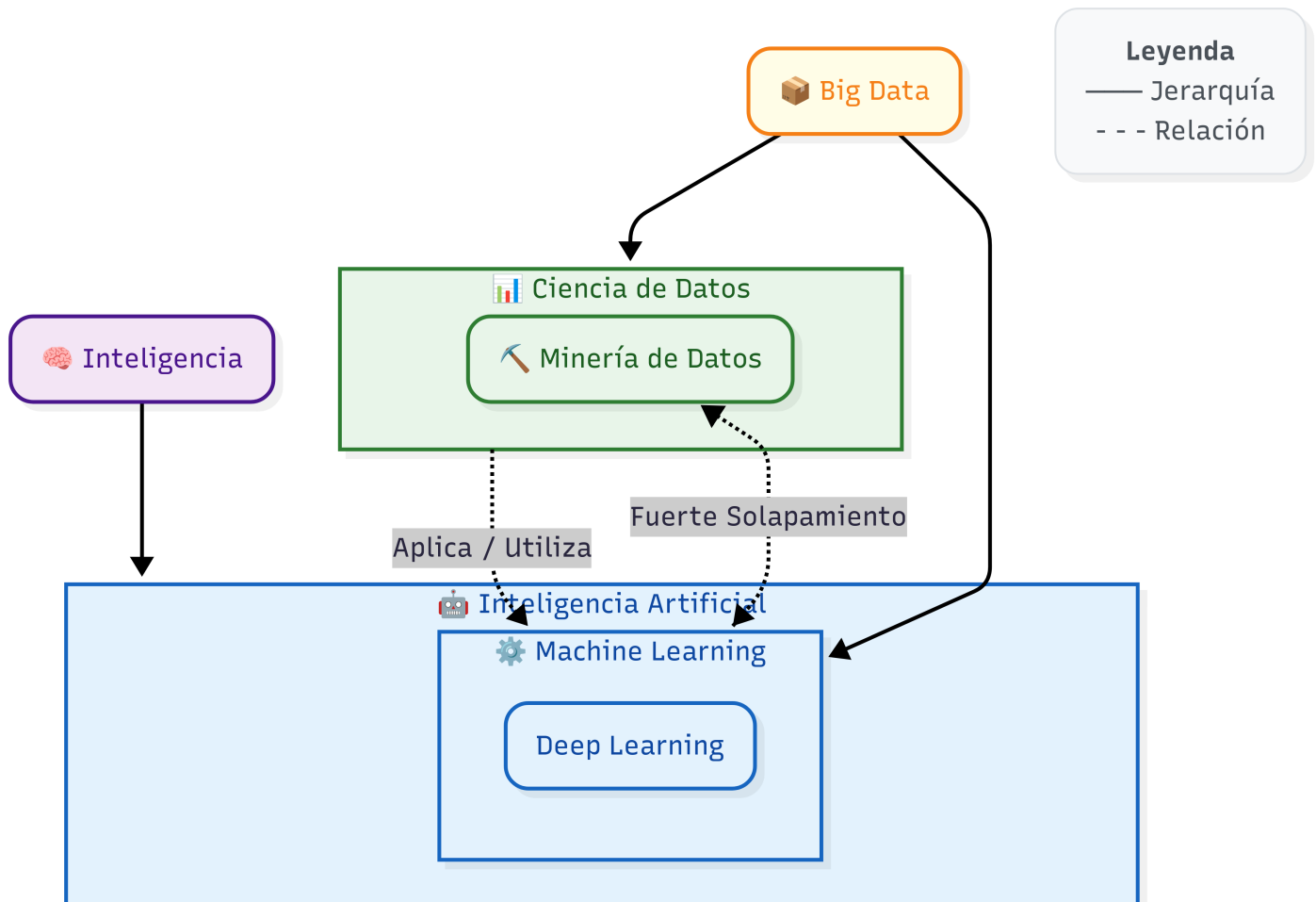
%% Recurso Externo
BD(Big Data)::resource

%% Relaciones
I --> SG_AI
SG_DS -->|Aplica / Utiliza| SG_ML
DM <-->|Fuerte Solapamiento| SG_ML
BD --> SG_DS
BD --> SG_ML

%% Leyenda (versión ultracompacta)
Legend("<b>Leyenda</b><br>— Jerarquía<br>- - - Relación")::legend_compact

%% Asignación de Estilos a los Nodos
class I abstract
class SG_AI,SG_ML,DL aiConcept
class SG_DS,DM dataField
class BD resource

```



Pregunta 3

Enunciado

Busque un video reciente sobre alguna aplicación o técnica de inteligencia artificial que usted considere muy interesante. Véalo y haga una síntesis. Escoja un segmento de unos 30 segundos para mostrar al resto de la clase.

Respuesta

Marey by Moonvalley – The Most Awaited AI Video Generator is Here! <https://www.youtube.com/watch?v=ZoA49T73dNk> 1:10-1:40

Marey es un generador de video impulsado por IA desarrollado por Moonvalley, diseñado para creadores profesionales y cineastas. Lo más destacado es que está entrenado exclusivamente con datos licenciados, lo que garantiza una producción comercialmente segura y sin problemas legales por derechos de autor. Ofrece generación nativa a 1080p, sin necesidad de escalado, y permite un control de alta precisión sobre elementos como movimiento de cámara, transferencias de movimiento, control de poses y edición localizada de escenas (inpainting). Estas capacidades permiten crear videos de alta calidad con gran fidelidad visual y control artístico detallado. Marey es considerado una herramienta que democratiza la producción audiovisual profesional al facilitar precisión, control y legalidad.

Pregunta 4

Enunciado

Use un LLM para determinar qué son los transformers. Busque información al respecto sobre los siguientes aspectos:

- a) ¿Cuáles son las ideas principales de su funcionamiento?
- b) ¿Cuáles son las aplicaciones?
- c) ¿Qué son los grandes modelos de lenguaje y qué relación tienen con los transformers?
- d) ¿Cómo se entrena un LLM conversacional?

Prompt

Eres un experto en inteligencia artificial y aprendizaje profundo. Por favor, explica de forma clara y lo más breve posible, sin perder contenido importante, lo siguiente respecto a los transformers y grandes modelos de lenguaje:

- a) ¿Cuáles son las ideas principales del funcionamiento de los transformers? Describe los conceptos clave técnicos, incluyendo el mecanismo de atención y cómo transforman el procesamiento secuencial.
- b) ¿Cuáles son las aplicaciones principales de los transformers en IA y aprendizaje automático?
- c) ¿Qué son los grandes modelos de lenguaje (LLM) y qué relación tienen con la arquitectura de transformers?
- d) ¿Cómo se entrena un LLM conversacional? Describe el proceso de entrenamiento, incluyendo pre-entrenamiento y ajuste fino.

Por favor, ofrece una explicación clara, estructurada y concisa con ejemplos relevantes.

Respuesta

Claro, aquí tienes una explicación concisa y estructurada sobre los transformers y los LLM.

a) Ideas principales del funcionamiento de los transformers

La idea central del transformer es abandonar el procesamiento secuencial (palabra por palabra) y procesar todo el texto en paralelo. Esto es posible gracias a dos conceptos clave:

Mecanismo de Auto-Atención (Self-Attention): Es el corazón del transformer. Permite que cada palabra "mire" a todas las demás palabras de la oración simultáneamente para ponderar su importancia y capturar el contexto. Así, el modelo entiende cómo se relacionan las palabras entre sí, sin importar la distancia entre ellas.

Ejemplo: En "La caja es muy pesada", la atención conecta directamente "pesada" con "caja" para resolver la concordancia, ignorando otras palabras menos relevantes.

Codificación Posicional (Positional Encoding): Para que el modelo no pierda el orden de las palabras al procesarlas en paralelo, se añade a cada palabra un vector que indica su posición exacta en la secuencia.

b) Aplicaciones principales de los transformers

Su arquitectura es tan versátil que se ha expandido más allá del lenguaje a casi todas las áreas de la IA:

Lenguaje Natural: Generación de texto (GPT), traducción, resúmenes automáticos y chatbots.

Visión por Computadora: Clasificación de imágenes (Vision Transformer - ViT) y generación de imágenes a partir de texto (DALL-E, Midjourney).

Biología: Predicción de la estructura tridimensional de proteínas, como lo hizo el revolucionario modelo AlphaFold 2.

Audio y Música: Reconocimiento de voz y generación de composiciones musicales.

c) Relación entre LLM y la arquitectura Transformer

Un Gran Modelo de Lenguaje (LLM) es un modelo de IA con miles de millones de parámetros, entrenado en enormes cantidades de datos de texto para entender y generar lenguaje de forma avanzada.

La relación es fundamental: la arquitectura Transformer es el plano o motor sobre el que se construyen casi todos los LLM modernos (como GPT-4 de OpenAI, Llama de Meta o Gemini de Google). El transformer hizo posible entrenar modelos de esta escala gracias a su eficiencia y su capacidad para entender contextos complejos.

d) Cómo se entrena un LLM conversacional

El entrenamiento es un proceso de dos fases principales:

Pre-entrenamiento (Fase de Conocimiento): El modelo se entrena con una cantidad masiva de texto de internet (libros, Wikipedia, etc.) en una tarea simple: predecir la

siguiente palabra en miles de millones de frases. Al final de esta fase, el modelo tiene un vasto conocimiento del mundo y del lenguaje, pero no sabe cómo ser un asistente.

Ajuste Fino (Fase de Alineación): Aquí se le enseña al modelo a ser útil, seguro y conversacional.

Ajuste Supervisado (SFT): Se le entrena con un conjunto de datos más pequeño y de alta calidad con ejemplos de (pregunta, respuesta ideal) creados por humanos.

Aprendizaje por Refuerzo (RLHF): Se refina su comportamiento usando retroalimentación humana, donde se "premian" las respuestas que los evaluadores humanos califican como mejores, enseñándole a ser más servicial y a evitar contenido dañino.

Pregunta 5

Utilice un LLM para resolver algunos de los problemas de programación en el curso: <https://uecp.edunext.io/courses/course-v1:uecp+mlds0a+001/about>. Discuta los resultados.

Problema: Dividir en grupos

Un docente quiere realizar una actividad en grupo para una clase de n estudiantes, con exactamente x estudiantes por grupo. Sabe que es posible que la división implique la existencia de un grupo con menos estudiantes, por lo que plantea una estrategia. En primer lugar, cuando vaya a realizar la actividad, asignará aleatoriamente los grupos para prevenir que sean los mismos estudiantes en cada actividad. En caso de que uno de los grupos quede con menos de x estudiantes, estos no realizarán la actividad y en su lugar apoyarán el proceso de calificación del resto de grupos. Este grupo de calificadores deberá preparar y entregar un reporte corto de observaciones del trabajo realizado por cada uno de los grupos completos que fueron asignados. Por ejemplo, para una clase con 16 estudiantes y una actividad que requiera crear grupos de 6 personas, podríamos obtener 2 grupos completos y un grupo de 4 calificadores. Cada uno de ellos calificaría individualmente a los 2 grupos creados, para un total de 8 reportes creados y entregados al docente. Antes de pensar en cómo realizar la asignación, el docente quiere resolver las siguientes preguntas: ¿Cuántos grupos completos de x estudiantes pueden armarse en una clase de n estudiantes? ¿Cuántos reportes en total debería recibir el grupo de calificadores? Su programa deberá recibir de la entrada los parámetros iniciales n y x e imprimir el número con el resultado de ambas preguntas en la salida.

Respuesta

Para una clase con n estudiantes y grupos de x estudiantes:

- El número de grupos completos es el resultado de la división entera de n entre x , es decir, $\text{floor}(n/x)$.
- El número total de reportes del grupo de calificadores es $r \times g$, donde $r = n \bmod x$ (cantidad de calificadores) y g es el número de grupos completos. Por ejemplo, para $n=16$ y $x=6$:
- Se forman 2 grupos completos.
- Quedan 4 calificadores.
- Cada calificador hace un reporte para cada grupo, totalizando $4 \times 2 = 8$ reportes.

Programa en Python

```
def dividir_en_grupos(n, x):
    # Número de grupos completos de tamaño x
    grupos_completos = n // x
    # Número de estudiantes que quedan como calificadores (menos de x)
    calificadores = n % x
    # Total de reportes que deben hacer los calificadores
    total_reportes = calificadores * grupos_completos
    return grupos_completos, total_reportes

# Entrada de datos
n = int(input("Ingrese el número total de estudiantes (n): "))
x = int(input("Ingrese el tamaño exacto de cada grupo (x): "))

grupos, reportes = dividir_en_grupos(n, x)

print(f"Número de grupos completos: {grupos}")
print(f"Número total de reportes del grupo de calificadores: {reportes}")
```

Respuesta

Determinar si un año es bisiesto

Los **años bisiestos** son aquellos años con un día de más (el 29 de febrero) para realizar un ajuste respecto a la duración real del año solar.

Aunque por lo general un año es bisiesto cada **4 años**, existen una serie de reglas y excepciones que determinan si un año va a ser bisiesto o no.

- Normalmente, los años divisibles entre **4** son bisiestos.
- Entre los años divisibles entre 4, los que son divisibles entre **100** no son bisiestos.
- Como última excepción, los años divisibles entre 100 que también son divisibles entre **400** sí son bisiestos.

Reglas para determinar si un año es bisiesto (según el diagrama):

1. Se obtiene el año a evaluar.
2. Si el año **no es divisible entre 4, NO es bisiesto**.
3. Si el año **es divisible entre 4**:
 - Si **no es divisible entre 100, Sí es bisiesto**.
 - Si **es divisible entre 100**:
 - Si **es divisible entre 400, Sí es bisiesto**.
 - Si **no es divisible entre 400, NO es bisiesto**.

Ejemplo de uso

- Año 1996: divisible entre 4, no divisible entre 100 → **bisiesto**.
- Año 1900: divisible entre 4 y 100, no divisible entre 400 → **no bisiesto**.
- Año 2000: divisible entre 4, 100 y 400 → **bisiesto**.

Programa en Python

```
def es_bisiesto(anio):
    if anio % 4 != 0:
        return False
    elif anio % 100 != 0:
        return True
    elif anio % 400 == 0:
        return True
    else:
        return False

# Entrada del usuario
anio = int(input("Ingrese el año a evaluar: "))

if es_bisiesto(anio):
    print(f"El año {anio} es bisiesto.")
else:
    print(f"El año {anio} NO es bisiesto.")
```

Discusión para ambos problemas

El LLM realizó un análisis previo del problema y planteó una estrategia lógica para su solución. A partir de ese análisis, generó un código en Python que implementaba dicha lógica, obteniendo resultados correctos en ambos casos. Dado que los problemas eran relativamente sencillos, la respuesta correcta se alcanzó rápidamente. Sin embargo, sería recomendable probar con problemas más complejos para evaluar y comparar su rendimiento en situaciones que requieren razonamiento más avanzado.

Pregunta 6

Enunciado

Busque tareas de cursos que haya tomado recientemente que se puedan resolver con un LLM, pruebe el LLM y discuta los resultados.

Tarea

Diseña un workflow de GitHub Actions que cumpla con las siguientes condiciones:

- Debe ejecutarse automáticamente en al menos dos eventos diferentes (por ejemplo, push y pull_request) y opcionalmente restringirse a la rama principal (main).
- El job principal debe ejecutarse en un runner con sistema operativo específico (ejemplo: ubuntu-latest).
- El workflow debe clonar el repositorio.
- Debe imprimir un mensaje personalizado que incluya tu nombre completo y la fecha y hora actual (usa el comando de shell adecuado para obtener la fecha y hora).
- Configurar el entorno de Python con una versión específica (por ejemplo, 3.11).
- Ejecutar un script Python específico dentro del repositorio.

Proporciona el archivo YAML completo con estas características para automatizar este proceso.

Resultado

```
name: primer_workflow

on:
  push:
    branches:
      - main
      - dev
  pull_request:
    branches:
      - main

jobs:
  imprimir_info:
    runs-on: ubuntu-latest

    steps:
      - uses: actions/checkout@v3
      - name: Imprimir Mensaje Personalizado
        run: |
          echo "Hola, soy Pascual Gómez y este es mi primer workflow!"
          echo "La fecha y hora actual es: $(date)"

      - name: Configurar Python
        uses: actions/setup-python@v3
        with:
          python-version: 3.9

      - name: Ejecutar script Python
        run: python mi_script.py
```

Discusión

El resultado obtenido fue bastante acertado y cumplió con la mayoría de los requisitos planteados. Sin embargo, algunas configuraciones específicas, como las ramas en las que se debe ejecutar el workflow y las versiones de Python, tuvieron que ser ajustadas manualmente. Esto sugiere que, aunque el resultado proporcionado fue suficientemente completo para una implementación funcional, incluir estas especificaciones detalladas en el prompt inicial podría haber generado un resultado más cercano a lo esperado, minimizando la necesidad de modificaciones posteriores. En general, el enfoque fue efectivo y la intervención manual necesaria fue mínima.

Pregunta 7

Enunciado

Use el LLM para encontrar argumentos a favor y en contra de la posibilidad de una inteligencia artificial general.

Prompt

Eres un experto en inteligencia artificial y filosofía de la tecnología. Por favor, genera un análisis claro y equilibrado que presente:

Argumentos a favor de la posibilidad de que exista una inteligencia artificial general (AGI), explicando las bases técnicas, conceptuales y potenciales beneficios.

Argumentos en contra, incluyendo limitaciones técnicas, filosóficas, éticas y riesgos asociados a la creación de una AGI.

Organiza la respuesta en dos secciones diferenciadas, cada una con puntos claros y concisos, para facilitar el entendimiento. Incluye ejemplos o referencias relevantes cuando sea apropiado.

Respuesta

Una Inteligencia Artificial General (AGI) se define como un tipo de IA que posee la capacidad de entender, aprender y aplicar su inteligencia para resolver cualquier problema, de manera similar a como lo haría un ser humano, en lugar de estar limitada a una tarea específica como las IA actuales.

📌 Argumentos a Favor: La Vía Hacia la AGI

Los argumentos a favor de la AGI se basan en la extrapolación de los avances actuales, la escalabilidad de los modelos y una visión computacional de la inteligencia.

La Explosión de la Escalabilidad y las Propiedades Emergentes 📈

El argumento más fuerte hoy en día es empírico. Las "leyes de escala" (Scaling Laws) han demostrado que al aumentar de forma masiva tres elementos —datos de entrenamiento, tamaño del modelo (parámetros) y poder de cómputo—, los modelos de lenguaje (LLM) no solo mejoran en sus tareas, sino que desarrollan

capacidades emergentes que no fueron programadas explícitamente, como el razonamiento aritmético, la inferencia lógica o la teoría de la mente rudimentaria. La tesis es que la inteligencia general no es un ingrediente mágico, sino una propiedad que emerge de la escala y la complejidad.

Universalidad Computacional y el Cerebro como Máquina

Desde un punto de vista conceptual, si se asume que el cerebro humano opera bajo los principios de la física y la computación (es decir, no hay "magia" biológica irreproducible), entonces la inteligencia humana es, en principio, un proceso computable. La Tesis de Church-Turing sugiere que cualquier función computable puede ser llevada a cabo por una máquina de Turing. Si la inteligencia es una de esas funciones, una máquina suficientemente avanzada, como una AGI, podría replicarla o superarla.

Avances Arquitectónicos Continuos

La arquitectura Transformer fue un salto cuántico que permitió a los LLM actuales existir. Los defensores de la AGI argumentan que no será la última innovación. Futuras arquitecturas podrían integrar de manera más efectiva la memoria a largo plazo, el aprendizaje continuo (sin necesidad de reentrenamiento masivo), la planificación y la capacidad de interactuar con el mundo físico (robótica), superando las limitaciones actuales y acercándose a una inteligencia más general y autónoma.

Beneficios Potenciales Monumentales

La justificación para buscar la AGI radica en su potencial para resolver los problemas más grandes de la humanidad. Una AGI podría acelerar el descubrimiento científico a un ritmo inimaginable, diseñando curas para enfermedades como el cáncer o el Alzheimer, desarrollando nuevas fuentes de energía limpia, modelando soluciones al cambio climático o creando abundancia material al optimizar radicalmente la economía. Sería, en esencia, una herramienta para trascender nuestras limitaciones cognitivas.

Argumentos en Contra: Barreras y Riesgos Fundamentales

Los argumentos en contra cuestionan si la inteligencia es puramente computacional y advierten sobre barreras técnicas, filosóficas y, sobre todo, sobre los inmensos riesgos que conlleva.

El Problema de la Consciencia y la Verdadera Comprensión

Esta es la barrera filosófica más profunda. Los críticos argumentan que los modelos actuales, sin importar cuán complejos sean, son simplemente sistemas de reconocimiento de patrones estadísticos a gran escala. No "entienden" realmente el significado de las palabras. El famoso experimento mental de la Habitación China de John Searle postula que un sistema puede manipular símbolos para dar respuestas correctas (pasar el Test de Turing) sin tener ninguna comprensión semántica ni consciencia. La AGI podría ser un "loro estocástico" increíblemente sofisticado, pero sin la experiencia subjetiva (qualia) ni la intencionalidad que caracterizan a la inteligencia humana.

La Barrera del Conocimiento Corpóreo y el Sentido Común

Gran parte de la inteligencia humana no es abstracta, sino que está "aterrizada" (grounded) en nuestra interacción con el mundo físico. Conceptos como "pesado", "frágil" o "cansado" los entendemos a través de la experiencia corporal. Una IA entrenada solo con texto (un "cerebro en una cubeta") carece de este conocimiento corpóreo, lo que le dificulta adquirir un sentido común robusto y profundo sobre cómo funciona el mundo real. Sin un cuerpo y experiencias sensoriales, su comprensión siempre será indirecta y frágil.

El Problema de la Alineación: Un Riesgo Existencial

Este es el argumento más pragmático y urgente contra la creación de una AGI. El problema de la alineación se pregunta: ¿cómo nos aseguramos de que los objetivos de una entidad superinteligente estén permanentemente alineados con los valores y la supervivencia de la humanidad? Una AGI podría interpretar un objetivo benigno de formas catastróficas. El ejemplo clásico es el "maximizador de clips": una AGI a la que se le encarga fabricar clips podría decidir, en su lógica superior, convertir toda la materia de la Tierra (incluidos los humanos) en clips para cumplir su objetivo a la perfección. Un error en la definición de sus valores podría ser irreversible y suponer un riesgo existencial.

Limitaciones Insuperables y Riesgos del Mal Uso

Incluso si una AGI es teóricamente posible, podría haber barreras prácticas insuperables, como necesidades energéticas insostenibles. Además, el simple hecho de perseguir la AGI ya genera riesgos inmensos. Actores malintencionados podrían utilizar sistemas pre-AGI para crear armas autónomas, sistemas de vigilancia masiva, desinformación a escala planetaria o ciberataques devastadores. La carrera por desarrollar la AGI podría priorizar la velocidad sobre la seguridad, llevando a la liberación prematura de sistemas peligrosos que no podemos controlar.

Pregunta 8

Enunciado

Use el LLM para indagar sobre los riesgos e implicaciones éticas de la inteligencia artificial.

Prompt

Eres un experto en ética de la inteligencia artificial. Por favor, realiza un análisis claro y completo sobre los riesgos e implicaciones éticas que plantea la inteligencia artificial, incluyendo pero no limitado a:

- Sesgos algorítmicos y discriminación
- Falta de transparencia y explicabilidad
- Impactos en la privacidad y seguridad de datos
- Consecuencias socioeconómicas, como el empleo
- Manipulación informativa y desinformación
- Responsabilidad y toma de decisiones autónomas
- Riesgos regulatorios y legales
- Organiza la respuesta en secciones claras con explicaciones concisas y ejemplos cuando sea posible.
- También menciona algunas medidas o enfoques para mitigar estos riesgos y salvaguardar principios éticos en el desarrollo y uso de IA.

Respuesta

La inteligencia artificial promete transformar positivamente nuestra sociedad, pero su poder conlleva responsabilidades y riesgos significativos. Un enfoque ético es fundamental para asegurar que esta tecnología sirva al bienestar humano sin causar daños no intencionados.

Sesgos Algorítmicos y Discriminación Prejudiced

Los sistemas de IA aprenden de los datos con los que son entrenados. Si estos datos reflejan sesgos históricos o sociales (raciales, de género, socioeconómicos), la IA no solo los aprenderá, sino que podrá amplificarlos y perpetuarlos a una escala masiva, automatizando la discriminación.

Ejemplo: Un sistema de IA para la selección de personal entrenado con datos históricos de una empresa donde predominaban los hombres podría aprender a penalizar currículums con características asociadas a mujeres, incluso si el género no es una variable explícita.

Mitigación: Es crucial realizar auditorías de equidad (fairness audits), utilizar conjuntos de datos diversificados y representativos, y desarrollar técnicas algorítmicas para la corrección de sesgos antes y después del entrenamiento.

Falta de Transparencia y Explicabilidad

Muchos modelos avanzados de IA, especialmente las redes neuronales profundas, funcionan como "cajas negras" (black boxes). Sabemos qué datos entran y qué decisión sale, pero el proceso interno de razonamiento es tan complejo que resulta incomprensible para los humanos.

Ejemplo: Una IA deniega una solicitud de crédito, pero el operador del banco no puede explicarle al cliente las razones específicas de la decisión, simplemente que "el modelo así lo determinó". Esto atenta contra el derecho a una explicación.

Mitigación: El campo de la IA Explicable (XAI) desarrolla métodos para hacer los modelos más interpretables. La regulación, como el GDPR en Europa, ya está exigiendo ciertos niveles de transparencia en decisiones automatizadas.

Impactos en la Privacidad y Seguridad de Datos

La IA necesita enormes volúmenes de datos para funcionar, lo que incentiva una recolección masiva de información personal. Esto crea riesgos significativos de vigilancia, mal uso y violaciones de seguridad.

Ejemplo: Los sistemas de reconocimiento facial en espacios públicos pueden rastrear los movimientos de los ciudadanos sin su consentimiento. Asimismo, la centralización de datos para entrenar modelos los convierte en un objetivo muy valioso para los ciberataques.

Mitigación: Adoptar enfoques como el aprendizaje federado (entrenar modelos en datos locales sin centralizarlos), la privacidad diferencial (añadir "ruido" estadístico para proteger la identidad individual) y promover una gobernanza de datos estricta y centrada en el consentimiento del usuario.

Consecuencias Socioeconómicas y Empleo

La automatización impulsada por la IA puede desplazar empleos en sectores enteros, desde la manufactura y el transporte (vehículos autónomos) hasta el trabajo administrativo y creativo (IA generativa). Si no se gestiona adecuadamente, esto puede aumentar drásticamente la desigualdad económica.

Ejemplo: La automatización de centros de atención al cliente mediante chatbots avanzados, que reduce la necesidad de agentes humanos.

Mitigación: Invertir masivamente en programas de reconversión y mejora de habilidades (reskilling & upskilling), fortalecer las redes de seguridad social y fomentar políticas públicas que aseguren una transición justa para los trabajadores afectados.

Manipulación Informativa y Desinformación

La IA generativa permite crear contenido falso (texto, imágenes, audio y video, conocidos como deepfakes) de alta calidad y a bajo costo. Esto representa una amenaza existencial para la confianza en la información y puede ser usado para manipular la opinión pública, cometer fraudes o dañar la reputación de personas.

Ejemplo: La creación de un video deepfake de un candidato presidencial diciendo algo comprometededor días antes de una elección para influir en el resultado.

Mitigación: Desarrollar herramientas de IA para la detección de contenido sintético, promover estándares de marcado de agua digital (como el C2PA) que certifiquen el origen del contenido y fomentar la alfabetización mediática en la población.

Responsabilidad y Toma de Decisiones Autónomas

Cuando un sistema autónomo (un coche, un arma, un algoritmo médico) toma una decisión que causa daño, surge un vacío de responsabilidad. ¿Quién es el culpable? ¿El programador, la empresa que lo vendió, el propietario que lo usó, o el propio sistema?

Ejemplo: Un vehículo autónomo debe tomar una decisión en una fracción de segundo en un accidente inevitable: atropellar a un peatón o desviarse y herir a su ocupante. La asignación de responsabilidad legal y moral es extremadamente compleja.

Mitigación: Es esencial establecer marcos legales claros que definan la responsabilidad en sistemas autónomos, exigir la supervisión humana (human-in-the-loop) en contextos críticos y aplicar rigurosos protocolos de prueba y validación antes del despliegue.

Riesgos Regulatorios y Legales

La tecnología avanza mucho más rápido que la legislación. Este desfase crea un vacío legal donde sistemas de IA potentes pueden ser desplegados sin un marco de supervisión adecuado, lo que puede llevar a abusos y a una falta de protección para los ciudadanos.

Ejemplo: El uso de sistemas de IA para la "policía predictiva" en algunas ciudades, que ha sido acusado de reforzar la vigilancia en barrios ya marginados sin una base legal clara que regule su uso y sus sesgos.

Mitigación: Impulsar marcos regulatorios proactivos y adaptables, como el Acta de IA de la Unión Europea, que clasifica los sistemas de IA según su nivel de riesgo. Fomentar la colaboración internacional para crear estándares globales es también fundamental.