tp1-fundamentos

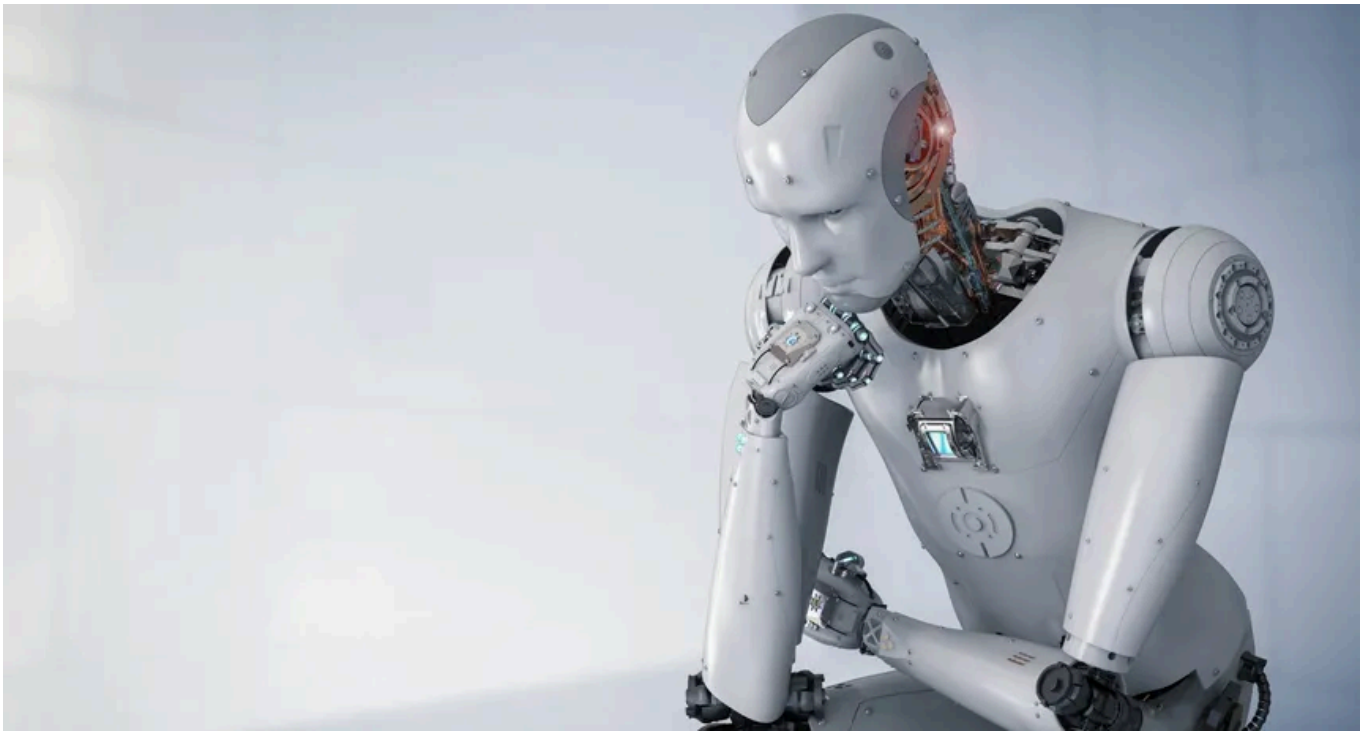
1)

Inteligencia Artificial Débil

Este tipo de inteligencia se refiere a la idea de que las máquinas podrían actuar como si fueran inteligentes, en esta categoría encajan aquellos sistemas o herramientas de inteligencia artificial que son diseñados para realizar tareas específicas o resolver problemas particulares.

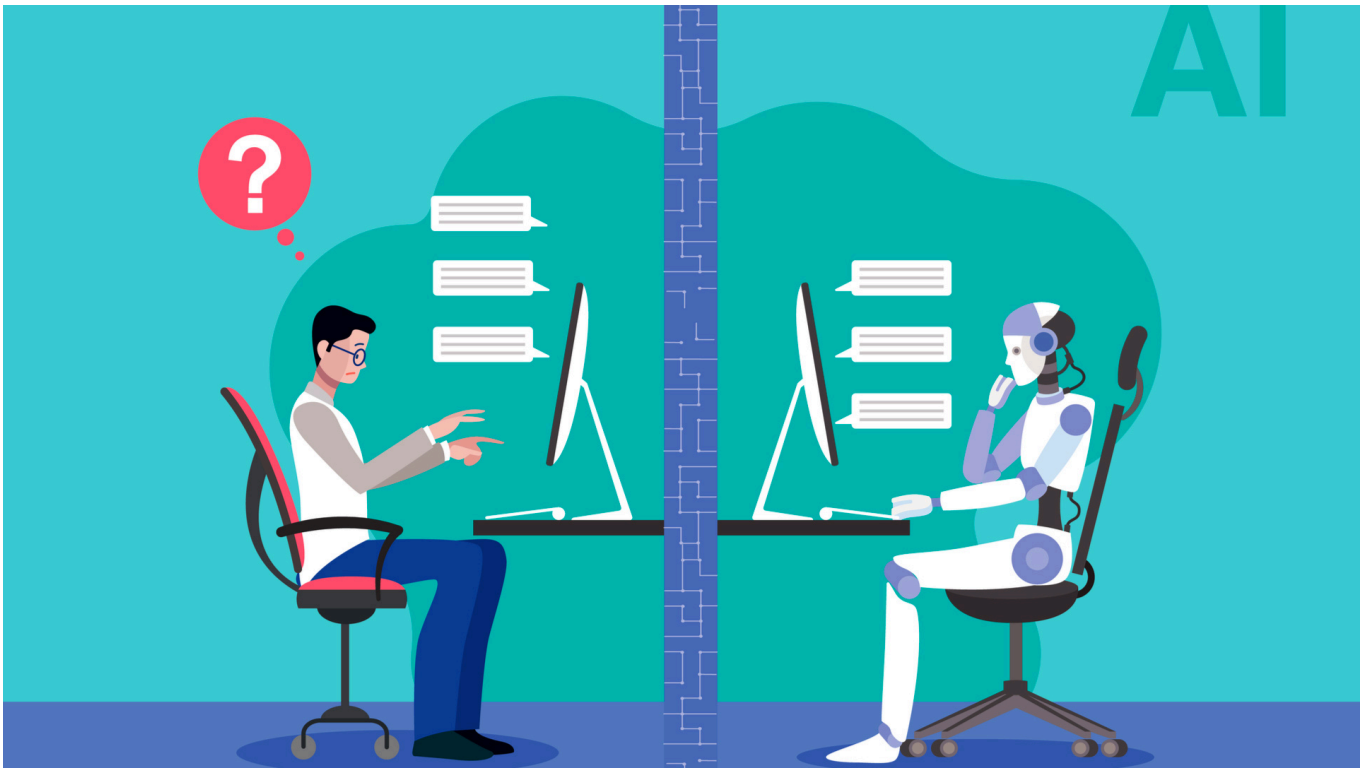
Estos sistemas no tienen conciencia del entorno o un entendimiento real del problema, simplemente son algoritmos que procesan datos y que, según los datos que se les otorgue, generen ciertos resultados.

Acá entra en juego la pregunta ¿pueden las máquinas pensar? Pero es lo mismo que preguntar si los submarinos pueden nadar (según la definición de nadar), o si los aviones pueden volar (según la definición de volar). Turing dejó de lado esa pregunta y la reemplazó por un examen de comportamiento inteligente (Turing Test).



El Test de Turing es una forma de medir la capacidad de una máquina para mostrar un comportamiento inteligente similar al de un humano: una persona tiene una conversación por texto, sin saber si del otro lado hay una persona o una máquina. Si, después de la conversación, el juez no puede distinguir con quién está hablando, se dice que la máquina ha

pasado el Test de Turing. Esto sugiere que la máquina tiene un nivel de inteligencia similar al humano en términos de lenguaje y comunicación. El objetivo del test no es que la máquina piense como un humano, sino que se comporte de manera que un humano no pueda diferenciarla de otra persona en una conversación.



El argumento de la discapacidad

Este argumento le da un enfoque al planteamiento de la pregunta principal, para ello se afirma que una máquina nunca podrá hacer ciertas cosas que los humanos pueden hacer, como ser amable, equivocarse, o enamorarse.

Con el tiempo, algunas de estas características fueron alcanzadas por las máquinas, como el hecho de equivocarse. Por otro lado, las computadoras pueden resolver problemas complejos como jugar al ajedrez o diagnosticar enfermedades, también pueden hacer tareas que parecen requerir juicio humano, como predecir el éxito de una persona o calificar exámenes. Aunque las computadoras pueden hacer estas cosas tan bien o mejor que los humanos, no significa que tengan la comprensión o el discernimiento humano, simplemente siguen algoritmos.

Por lo tanto, aunque las computadoras superaron limitaciones que antes se creían sólo para humanos, aún hay tareas que las máquinas no pueden realizar, o al menos no tan bien o con tanta coherencia con la que podría hacerla un humano.

La objeción matemática

Teorema de la incompletitud: Gödel demostró que en cualquier sistema formal suficientemente potente (como los que se usan en matemáticas para construir la aritmética básica), siempre habrá afirmaciones que no se pueden probar ni refutar utilizando las reglas y axiomas dentro de ese sistema. Se divide en 2 teorías:

- Primera Incompletitud: En un sistema existirán proposiciones verdaderas que no se pueden probar dentro de ese mismo sistema. Es decir, no importa cuántas reglas o axiomas se agreguen, siempre habrá algunas verdades que simplemente no se pueden demostrar utilizando esas reglas.
- Segunda Incompletitud: Un sistema formal no puede probar su propia consistencia. Es decir, no se pueden usar las reglas de un sistema para demostrar que esas mismas reglas no llevarán a contradicciones.

Algunos filósofos utilizaron el teorema de la incompletitud para confirmar o demostrar que las máquinas son inferiores a los seres humanos en cuestión de inteligencia, ya que las máquinas son sistemas formales que están limitadas por dicho teorema, mientras que los humanos no.

El argumento de la informalidad del comportamiento

Este argumento se basa en que el comportamiento humano es tan complejo que no puede ser capturado por un simple conjunto de reglas. Como las computadoras siguen reglas, se sugiere que nunca podrán comportarse tan inteligentemente como los humanos. Este enfoque se lo llama "Good Old-Fashioned AI".

Good Old-Fashioned AI: es el nombre colectivo para todos los métodos de investigación de la inteligencia artificial que se basan en representaciones de alto nivel "simbólico" de los problemas, la lógica matemática y la búsqueda. Es decir, trata de capturar el comportamiento inteligente a través de un sistema que razona lógicamente a partir de hechos y reglas.

Se argumenta que los humanos también tienen conocimiento de ciertas reglas, pero también tienen un contexto holístico que guía su comportamiento, en otras palabras, tenemos conciencia de cada tema o situación desde una perspectiva global e integrada, en lugar de enfocarnos sólo en partes aisladas.

Ejemplos de inteligencia artificial débil

- Asistentes virtuales: son herramientas diseñadas para "entender" y responder a preguntas dentro de un rango limitado, muy extenso, pero sigue siendo limitado. Los asistentes más comunes son Siri, Alexa y Bixby.

- Sistemas de recomendación: Estos sistemas analizan el historial de visualización, compras o reproducción de los usuarios para sugerir contenido similar que pueda interesarles. Utilizan algoritmos que identifican patrones en los datos, pero no entienden por qué alguien podría preferir un tipo de contenido sobre otro, solo reconocen patrones estadísticos en el comportamiento de los usuarios.
- Filtros de spam: Los sistemas de detección de spam analizan el contenido de los correos electrónicos para identificar y filtrar los mensajes no deseados. Utilizan modelos de aprendizaje automático que han sido entrenados en grandes conjuntos de datos para distinguir entre correos legítimos y spam. Aunque son muy efectivos, operan únicamente dentro del ámbito de la clasificación de correos electrónicos y no tienen ninguna comprensión del contenido.

Inteligencia Artificial Fuerte

La inteligencia artificial fuerte se refiere a que las máquinas realmente pueden pensar, y no simular que piensan. En teoría, una máquina sería capaz de entender, razonar y tener conciencia de manera similar a un ser humano. Esto no se limita a tareas específicas, sino que tendría una inteligencia general que le permitiría abordar una amplia gama de problemas.

Aunque una máquina pase el Test de Turing, no significa que pueda pensar, sería una simulación del pensamiento. Turing ya había anticipado esta crítica y citó a Geoffrey Jefferson, quien dijo que una máquina solo podría considerarse igual a un cerebro si puede crear arte (como sonetos o conciertos) debido a sentimientos reales, no solo por combinar símbolos al azar.

Turing llama a esto el "argumento de la conciencia". Según este argumento, para que una máquina realmente piense, tendría que ser conciente de sus propios pensamientos y emociones, y no solo simularlos.

Estados mentales y el cerebro en un tonel

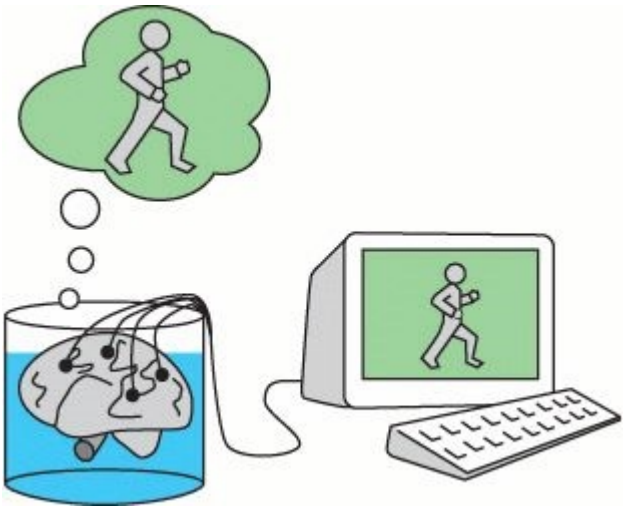
Fisicalismo: es la idea de que todo lo que existe, incluyendo los estados mentales y las experiencias, puede ser explicado completamente en términos de física y propiedades físicas.

Los filósofos del fisicalismo que tratan de explicar los estados mentales de un humano se concentran en los estados intencionales, entre estos se encuentran el saber, el deseo, el miedo, etc. los cuales se refieren a algún aspecto del mundo exterior.

El fisicalismo dice que el estado mental de una persona depende de su estado cerebral. Así que si estás concentrado en comer una hamburguesa, tu cerebro está en un estado mental

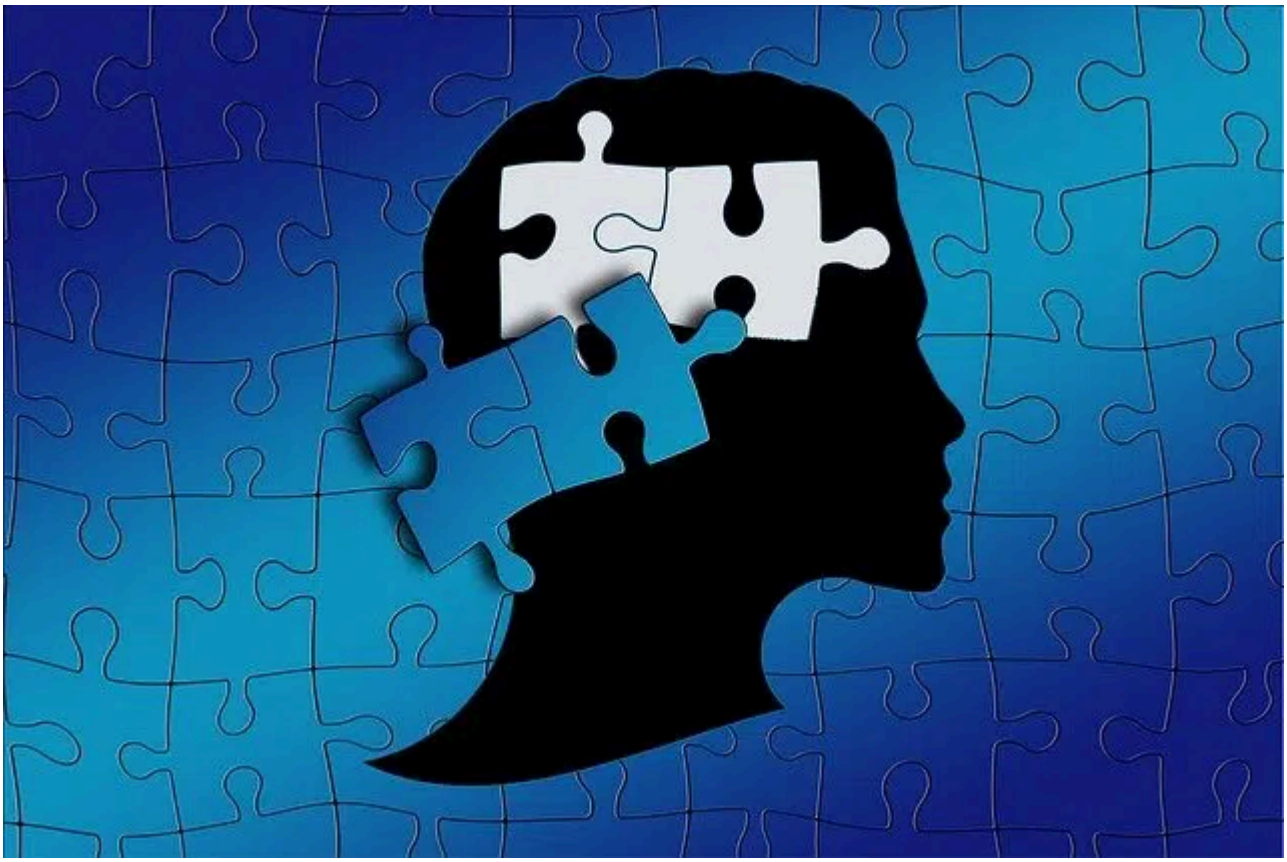
relacionado con eso. No importa la configuración exacta de los átomos en tu cerebro, lo importante es que no se puede tener un estado mental diferente con el mismo estado cerebral.

Si tu cerebro fue sacado de tu cuerpo al nacer y puesto en un recipiente que lo mantiene funcionando. Mientras tanto, una computadora te da señales para simular una vida que es igual a la que tendrías si estuvieras en tu cuerpo, algo parecido a la película de Matrix. Podrías tener un estado cerebral igual al de alguien que realmente está comiendo una hamburguesa, pero en realidad no estás comiendo ninguna. En este caso, tu estado mental de "saber que estás comiendo una hamburguesa" no sería correcto porque no estás experimentando eso.



Funcionalismo y el experimento de reemplazo de cerebro

Funcionalismo: es una teoría que dice que los estados mentales de una persona (como pensar, sentir o desear) son definidos por sus roles o funciones en un sistema, no por su implementación física específica. Según esta teoría, si dos sistemas (como un cerebro humano y una computadora) tienen los mismos procesos causales, deberían tener los mismos estados mentales.



Si reemplazamos gradualmente las neuronas de un cerebro humano por componentes electrónicos que cumplen las mismas funciones, y el sujeto sigue mostrando signos de conciencia (respondiendo a preguntas, sintiendo dolor), entonces debemos concluir que la conciencia no es exclusiva de los cerebros biológicos.

De esta forma, habría que explicar la conciencia en términos de funciones cerebrales, sin apelar a la biología de las neuronas. Por lo tanto, si los procesos son los mismos, la experiencia mental debería ser la misma, independientemente de la implementación física.

El naturalismo biológico y la habitación china

Naturalismo biológico: se considera que los procesos biológicos son fundamentales para entender todos los aspectos del comportamiento humano y la mente. En esencia, sostiene que las explicaciones biológicas, como la genética, la evolución y las funciones cerebrales, son clave para entender fenómenos psicológicos, cognitivos y sociales.

Según esta teoría, los estados mentales son características de alto nivel que son causadas por procesos físicos de bajo nivel en las neuronas, y que son las propiedades de las neuronas lo que realmente importa.

Conciencia, qualia y la brecha explicativa

- **Conciencia**: En el debate sobre la IA fuerte, uno de los temas centrales es la conciencia. La conciencia puede desglosarse en diferentes aspectos como la comprensión y la auto-conciencia.
- **Qualia**: El término técnico para referirse a la naturaleza intrínseca de las experiencias es **qualia** (del latín que significa “tales cosas”). Los qualia son las cualidades sensoriales subjetivas de nuestras experiencias. Por ejemplo, la sensación de ver el color rojo es un tipo de qualia.

- **Brecha explicativa**: Es la dificultad de conectar los procesos neuronales descritos científicamente con la experiencia subjetiva. Incluso si entendemos perfectamente los procesos cerebrales, no hay una forma aceptada de deducir que un ser tiene una experiencia conciente particular. Esta brecha ha llevado a algunos filósofos a concluir que los humanos podrían ser incapaces de comprender completamente su propia conciencia.
- **Posición de Turing**: Alan Turing reconoció que la cuestión de la conciencia es difícil, pero no creía que fuera relevante para la práctica de la IA. Turing pensaba que la creación de programas que se comportan de manera inteligente no requiere resolver los misterios de la conciencia. El objetivo principal es desarrollar programas inteligentes, no necesariamente conscientes, y Turing no pensaba que pudiéramos determinar si un programa es conciente o no.

La ética y los riegos de desarrollar Inteligencia Artificial

Riesgos

Desplazamiento Laboral e Impacto Económico

- **Preocupación:** La automatización y la IA pueden llevar a la pérdida de empleos y cambios económicos.
- **Perspectiva:** Aunque la IA ha desplazado algunos empleos, también ha creado nuevos roles y transformado la naturaleza del trabajo. El desafío es equilibrar estos cambios y asegurar que los trabajadores desplazados tengan oportunidades para reentrenarse y encontrar nuevos roles.

Tiempo Libre y Equilibrio Trabajo-Vida

- **Preocupación:** La IA podría llevar a un exceso de tiempo libre o a una presión incrementada para trabajar más duro.
- **Perspectiva:** La IA tiene el potencial de reducir las horas de trabajo y aumentar el tiempo libre, pero también podría contribuir a una cultura de disponibilidad constante y mayores expectativas en el lugar de trabajo. El objetivo es encontrar un equilibrio que mejore la calidad de vida sin aumentar el estrés.

Sentido de unicidad

- **Preocupación:** La IA podría socavar nuestro sentido de singularidad y autonomía humana.
- **Perspectiva:** Cambios históricos en la comprensión de nuestro lugar en el universo (por ejemplo, Copérnico, Darwin) han desafiado la auto-concepción humana. La IA añade a esto al potencialmente difuminar las líneas entre las capacidades humanas y las de las máquinas.

Uso indebido

- **Preocupación:** Los sistemas de IA podrían usarse para fines dañinos, incluyendo aplicaciones militares y de vigilancia.
- **Perspectiva:** El potencial de mal uso de la IA requiere directrices éticas estrictas y supervisión. Asegurar que la IA se desarrolle y utilice de manera responsable es crucial para prevenir daños y mantener las libertades civiles.

Responsabilidad y Culpabilidad

- **Preocupación:** Determinar la responsabilidad por las decisiones y acciones de la IA puede ser un desafío.

- **Perspectiva:** A medida que los sistemas de IA se vuelven más autónomos, establecer líneas claras de responsabilidad y culpabilidad es esencial. Esto implica definir roles y responsabilidades para desarrolladores, usuarios y otros interesados.

Riesgo para la humanidad

- **Preocupación:** La IA podría representar una amenaza existencial para la humanidad.
- **Perspectiva:** El riesgo de que la IA se vuelva incontrolable o persiga objetivos dañinos es una preocupación seria. Asegurar que los sistemas de IA se diseñen con mecanismos de seguridad robustos y alineados con los valores humanos es crucial.

Cómo manejar los riesgos:

- **Mecanismos de control:** Implementar mecanismos de control para prevenir consecuencias no deseadas y asegurar que los sistemas de IA cumplan con estándares éticos.
- **Transparencia y Supervisión:** Mantener la transparencia en el desarrollo de la IA e involucrar a diversos interesados en la supervisión puede ayudar a mitigar riesgos.
- **Diseño Ético:** Diseñar sistemas de IA teniendo en cuenta consideraciones éticas, incluyendo el respeto a los derechos humanos y evitando posibilidades para usos dañinos.

Discusiones

Considero que por ahora estamos bastante lejos de desarrollar una "IA Fuerte" ya que primero hay que entender la conciencia humana antes de darle conciencia a una máquina o sistema. Pienso que la teoría del funcionalismo sería la más acertada, si llegáramos a entender los procesos neuronales y describirlos como funciones, no sería costoso representar esas funciones en una máquina.

Por ahora, todos las IA implementadas entrarían a la categoría de "IA Débil", ya que, por más que tengan una gran conocimiento, incluso mucho más amplio que el de cualquier ser humano, no puede extenderse más allá de este, o incluso generar nuevo conocimiento a partir de lo que ya sabe (como sacar conclusiones).

2)

¿Es posible considerar a los agentes conversacionales basados en grandes modelos de lenguaje (LLMs) como conscientes?

Considerando la tecnología y el conocimiento de hoy, diría que no. A veces puede llegar a ser confuso si un LLM o un agente construido sobre uno es consciente o no, las razones por esto son:

- Las habilidades conversacionales de los agentes son bastante cercanas al nivel humano.
- Puede realizar acciones del mundo real (hacer compras, escribir emails, etc.).
- Tiene información sobre casi cualquier tema que se nos ocurra.

El problema es que no se puede ir más allá del contexto que se les otorga (que parece infinito por ser tan amplio), y por ende tampoco tiene un criterio de la verdad, no puede comprar su conocimiento con la realidad externa y actualizarse en base a ello. Y aunque realice acciones del mundo real, no está haciendo otra cosa que satisfacer las necesidades del usuario real.

La conciencia implica la atención y auto-conciencia, y considero que para tener estas cualidades, hay que entender tu contexto, tu entorno, cosa que los LLM o los agentes basados en estos carecen.

¿Cuáles son las implicaciones éticas de atribuir conciencia y, por ende, "derechos morales" a los agentes de IA avanzados?

- **Derechos Morales:** Si una comunidad considera que los agentes basados en LLM son conscientes y capaces de sufrir, podría surgir un dilema moral. La preocupación es que se podría dar prioridad a los derechos de los agentes de IA sobre los de los seres humanos.
- **Impacto en las Relaciones Humanas**: La presencia de agentes que parecen conscientes podría tener un impacto significativo en las relaciones humanas. En una visión optimista, podría reencantar el mundo al introducir nuevas formas de "ser mágico". Sin embargo, en una visión pesimista, podría degradar las relaciones humanas auténticas si las personas prefieren la compañía de agentes de IA a la de otros humanos.

3)

La inteligencia artificial generativa utiliza técnicas de machine learning para aprender y crear nuevos datos. Esta tecnología es capaz de generar texto, imágenes, música, videos y otros medios en respuesta a comandos específicos. Tiene una gran capacidad para imitar patrones y estructuras del contenido con el que se entrena.

Desventajas según Emily M. Bender

Emily M. Bender, experta en lingüística computacional, señala varias desventajas de la inteligencia artificial generativa:

- Falta de inteligencia y contexto: Los modelos de lenguaje de gran tamaño (LLM) no poseen la inteligencia ni el conocimiento del contexto necesario para ser completamente confiables o coherentes en todas las situaciones. Esto puede llevar a respuestas incorrectas o irrelevantes, como por ejemplo el pulpo que al no saber qué era un oso, no supo qué responder.
- Problemas éticos: La posibilidad de falsificar conversaciones humanas a través de chatbots plantea serias preocupaciones éticas. El uso de estos sistemas para suplantar identidades humanas puede llevar a la desinformación y a la manipulación de la opinión pública: si la falsificación de billetes es ilegal, ¿por qué la de los humanos no?
- Contenido inapropiado: Al entrenar los modelos con grandes cantidades de datos, es inevitable que algunos de estos datos contengan contenido de odio, como racismo o xenofobia. Filtrar estos contenidos de manera efectiva es algo imposible ya que se utilizan billones de datos.

Utilidades de la inteligencia artificial generativa

A pesar de estas desventajas, la inteligencia artificial generativa tiene muchas aplicaciones útiles en diversos campos:

- Creación de contenido visual y auditivo: Puede generar imágenes, videos, música y otros tipos de contenido multimedia que ayudan a las personas a expresar y transmitir ideas de manera más efectiva.
- Generación de borradores y prototipos: Aunque no siempre produce resultados perfectos, puede ser muy útil para generar borradores de textos, prototipos de diseño y otros materiales iniciales que luego pueden ser refinados por humanos.
- Asistencia a personas con discapacidades: Puede describir imágenes o entornos, lo que es muy útil para personas con discapacidades visuales, mejorando su accesibilidad y autonomía.
- Síntesis de voz: La capacidad de generar voces permite la creación de contenido de audio, asistentes virtuales y narraciones automáticas, lo que puede ser valioso en diversas aplicaciones tecnológicas.
- Toma de decisiones: Si se alimenta con datos precisos y relevantes, la inteligencia artificial generativa puede ayudar a agilizar la toma de decisiones en ámbitos como los negocios, la medicina y más.
- Sistemas de recomendación: Puede mejorar los sistemas de recomendación, sugiriendo libros, películas, música y otros contenidos basados en las preferencias y el historial del usuario.

- Automatización de tareas: Puede automatizar tareas repetitivas y tediosas, como la redacción de respuestas a correos electrónicos o la transcripción de audio a texto, aumentando la eficiencia y productividad.