CIVICAi

PROTOCOLO PARA UNA INTELIGENCIA ARTIFICIAL CÍVICA

CIVICAi

El uso de las tecnologías digitales relacionadas con la movilidad y la comunicación se ha vuelto omnipresente en nuestras vidas en todo el planeta. La pandemia de la COVID-19 generalizó su adopción y cambió la manera en que trabajamos, consumimos, nos relacionamos, nos informamos e interactuamos con las máquinas y el mundo. Sin casi darnos cuenta, nos hemos encontrado inmersos en plena Cuarta Revolución Industrial, también conocida como Revolución 4.0, que se caracteriza por la automatización y el flujo de información entre tecnologías físicas, biológicas y digitales. Las tecnologías 4.0 que más impactan en el mundo físico y biológico son, entre otras, la biotecnología, la robótica avanzada, la impresión 3D, los nuevos materiales y el Internet de las Cosas (IoT). En el mundo digital, la revolución 4.0 incluye la tecnología de bloques (*blockchain*), los datos masivos (*big data*) y su análisis, la computación en la nube, la ciberseguridad, las tecnologías de realidad virtual y aumentada, y la Inteligencia Artificial (IA).

Estas tecnologías, aunque abren nuevas posibilidades y escenarios de innovación en todos los ámbitos sociales y económicos, sólo han contribuido a aumentar, de manera incremental más que revolucionaria, la eficiencia, la productividad y la calidad de muchos servicios y productos. La aparición de la IA generativa (IAG), que ha tenido un crecimiento sin precedentes en el número de usuarios a nivel mundial, y que estudios prospectivos indican que tendrá un impacto muy significativo en los sectores productivos y las administraciones públicas, es probablemente el catalizador que debería permitir al conjunto de tecnologías 4.0 convertirse en una verdadera revolución, caracterizada por una simbiosis profunda entre los humanos y las máquinas. Algunos expertos hablan de la Revolución 5.0, la de la colaboración entre seres humanos y las tecnologías inteligentes avanzadas para resolver problemas complejos y crear nuevas formas de interacción y experiencia.

Como sucede en todos los procesos revolucionarios, la IAG ofrece numerosas oportunidades, pero también plantea nuevos retos y amenazas para la humanidad. Debemos procurar que la tendencia natural de los humanos a sobrevalorar las consecuencias y los riesgos a corto plazo no nos haga perder de vista los riesgos y los impactos de la IAG a medio y largo plazo. La omnipresencia de una tecnología que emula habilidades humanas conlleva el riesgo de que los humanos pierdan habilidades sociales y puedan ser manipulados en la toma de decisiones colectivas o personales. Es por eso que CIVICAi propone la participación ciudadana en la gobernanza de la IA y adopta este protocolo para una inteligencia artificial cívica, al servicio de las personas y para el bien colectivo.

CIVICAi

ÍNDICE

1. INTELIGENCIA CÍVICA

Una de las tecnologías más representativas de la Revolución 4.0 y que hace que esta sea realmente revolucionaria es la Inteligencia Artificial, con su rápido y constante desarrollo, y su impacto transversal. La llegada de los sistemas generativos de IA y la próxima aparición de la Inteligencia Artificial General (AGI) representa un cambio revolucionario en la evolución humana y en nuestra comprensión de la inteligencia, el lenguaje y la cognición. Como miembros de CIVICAi, una de nuestras responsabilidades es facilitar el discurso público y mejorar la comprensión social de estos profundos y rápidos cambios para que, una vez producidos, la nueva evolución que seguirá pueda ser asimilada de manera armónica y para el bien colectivo. El conjunto de preguntas y respuestas que se propone en este documento está diseñado para proporcionar un mapa con información suficiente para navegar por las complejidades de la IA generativa, abordando de manera superficial pero suficiente tanto los matices técnicos como algunas de las implicaciones filosóficas más amplias.

Históricamente, nuestra concepción de la inteligencia ha estado profundamente influenciada por el dualismo cartesiano, donde René Descartes postulaba una estricta separación entre la mente y el cuerpo[1]. Esta perspectiva condicionó los inicios de la IA con propuestas para emular la cognición humana mediante reglas lógicas y la manipulación simbólica, con la llamada IA simbólica. La teoría de la gramática universal de Noam Chomsky reforzó aún más la idea de que la capacidad de adquirir lenguaje está innatamente programada en el cerebro humano, subrayando la primacía de las estructuras inherentes sobre los patrones aprendidos[2].

Sin embargo, los avances recientes en la IA, especialmente gracias al trabajo de investigadores como Geoffrey Hinton, Yoshua Bengio y Yann LeCun, ganadores del premio Turing de 2018, han desafiado estos postulados de la lingüística más ortodoxa y tradicional[3]. El trabajo pionero de estos investigadores en redes neuronales y aprendizaje profundo ha demostrado que los modelos de lenguaje grandes o de gran escala (LLMs) pueden alcanzar una habilidad notable para entender la gramática o sintaxis de textos y generar lenguaje humano. Estos modelos conexionistas aprovechan las representaciones distribuidas del conocimiento en nodos de redes neuronales artificiales para mostrar incipientes propiedades emergentes, lo que sugiere que los comportamientos complejos

[1] https://plato.stanford.edu/entries/dualism/
[2] https://plato.stanford.edu/entries/innateness-language/
[3] https://awards.acm.org/about/2018-turing

pueden surgir de la interacción de componentes más simples. Esta perspectiva conexionista se alinea con el trabajo filosófico de la mente y del lenguaje de Ludwig Wittgenstein, quien pone el énfasis en el uso del lenguaje en situaciones y contextos concretos, más que en la dependencia exclusiva de estructuras y reglas gramaticales fijas. Esto implica que comprender el lenguaje requiere atender a los contextos sociales en los que se despliega, más que a estructuras lingüísticas abstractas e invariantes[4].

El punto de vista heterodoxo defendido por Geoffrey Hinton, Yoshua Bengio y otros, sugiere que el aprendizaje a partir de grandes cantidades de datos, más que la dependencia de reglas pre-programadas, puede conducir a una forma de comprensión práctica que desafía el postulado de estructuras innatas de Chomsky. Los desarrollos sobre mecanismos neurocognitivos con una visión mecanicista de la mente, recientemente desarrollados por Gualtiero Piccinini, refuerzan este argumento, indicando que los mecanismos que sustentan la cognición humana pueden ser representados de manera análoga, aunque diferente, en sistemas artificiales[5]. Cabe tener en cuenta, sin embargo, que los procesos de generación de contenido semántico original por parte de la IA generativa actual son algorítmicos, utilizan correlaciones estadísticas y el reconocimiento de patrones a partir de grandes conjuntos de datos de entrenamiento que los humanos y la Internet les han proporcionado. Por lo tanto, son diferentes de los procesos semánticos del cerebro, ya que estos son inherentemente biológicos, dependen del contexto, pueden ser intencionales, se autocontrolan e integran insumos sensoriales, memoria, emociones y otras funciones cognitivas ligadas a las relaciones continuadas de la unidad mente-cuerpo con el entorno, características relacionadas con lo que llamamos consciencia.

A medida que nos acercamos a una nueva era influenciada por la inteligencia artificial, con su potencial no solo para imitar, sino también para ampliar las capacidades cognitivas humanas de maneras sin precedentes, es crucial reconocer que toda revolución implica romper con la ortodoxia predominante. El éxito de la IA conexionista (redes neuronales) representa no solo un avance tecnológico, sino también un cambio de paradigma en la forma en que conceptualizamos la inteligencia misma. Si adoptamos una perspectiva heterodoxa que incorpore ideas de múltiples disciplinas, podremos gestionar mejor los retos éticos, sociales y filosóficos planteados por las altas capacidades

[4] https://philosophynow.org/issues/106/Wittgenstein_Frege_and_The_Context_Principle
[5] https://www.thebsps.org/reviewofbooks/gualtiero-piccinini-physical-computation/

de la IA generativa y la futura AGI.

Un aspecto fundamental en este contexto es la computabilidad de la inteligencia. La inteligencia, históricamente considerada una característica exclusiva de los humanos, a pesar de ser evolutiva, hoy se percibe como una propiedad emergente que también podría surgir en los sistemas digitales complejos, como los algoritmos de redes neuronales. Estos algoritmos que impulsan la IA generativa nos llevan a cuestionar los límites de la computabilidad de la inteligencia y si esta puede ser reproducida o emulada completamente por máquinas.

Esto abre la puerta a discutir sobre los inconvenientes y los riesgos asociados a la IA generativa. A corto plazo, estos pueden incluir sesgo, la vulneración de la privacidad y de la propiedad intelectual, cuestiones éticas, la rápida transición en el mercado laboral, la desinformación planificada y la pérdida de valores democráticos o incluso la alteración de la democracia misma. Los sistemas de IA pueden reforzar los prejuicios existentes si los conjuntos de datos de entrenamiento no están adecuadamente supervisados. Además, la recopilación masiva de datos plantea preocupaciones sobre la privacidad, y el uso de contenidos generados por IA también presenta desafíos relacionados con la propiedad intelectual y los derechos de autor. Es fundamental que los proveedores de modelos de lenguaje grandes (LLMs) operen dentro de una estructura o sistema legal, lo más universal posible, que les obligue a mitigar cualquier discurso negligente y a alinear sus modelos con hechos contrastables o "verdaderos", mediante procesos abiertos y democráticos[6]. Los riesgos a largo plazo incluyen la posibilidad de que se llegue a la singularidad tecnológica, un término introducido por John von Neumann para identificar el plausible futuro momento en que la tecnología, en este caso la IA generativa, supere la inteligencia humana [7]. Esto implicaría que la IA generativa o la AGI autogestionara su función de valor o los criterios para alcanzar objetivos, los cuales podrían no estar alineados con los intereses o metas de los humanos. Los riesgos a largo plazo también plantean la idea de una post-humanidad, en la que la integración de IA avanzada en la sociedad humana transforme la experiencia y la identidad humanas sin que haya plena capacidad para decidirlo democráticamente.

Es imprescindible preguntarse: ¿quién supervisará y cómo se supervisarán de manera efectiva los sistemas de IA generativa actuales y en desarrollo, que

[6] https://doi.org/10.1098/rsos.240197
[7] https://lab.cccb.org/en/the-singularity/

están mayoritariamente en manos de proveedores privados? El reto es formidable, ya que las legislaciones y regulaciones vigentes no son globales y se limitan a establecer un régimen sancionador que actúa a posteriori para detener o corregir las acciones malintencionadas una vez ya se han producido y propagado. Por tanto, más allá de disuadir a los proveedores con sanciones, los estados deberían consensuar globalmente un sistema de vigilancia en tiempo real que abarque datos con entradas multimodales, el entrenamiento, los algoritmos y los resultados de estos sistemas de IA generativa y de la futura AGI. Este sistema debería implementarse a través de una red de centros de computación propios, equipados con recursos computacionales de hardware y software iguales o superiores a los que poseen las empresas privadas.

La eficiencia y la sostenibilidad de la IA generativa son otros aspectos importantes que deben tenerse en cuenta. Los modelos actuales requieren una cantidad significativa de recursos computacionales y energéticos, lo cual podría no ser sostenible a largo plazo, ya sea por falta de recursos energéticos o por entrar en conflicto con otras prioridades humanas. Las soluciones computacionales híbridas, que combinen sistemas analógicos y digitales, o que utilicen chips que combinen diferentes tipos de núcleos o procesadores especializados, con una integración óptima entre el hardware y el software, como lo hacen los procesadores de algunos teléfonos inteligentes, podrían ofrecer una vía para minimizar estos problemas, mejorando la eficiencia computacional y reduciendo la huella ecológica.

En conclusión, este protocolo o marco conceptual y operativo tiene como objetivo inspirar las manifestaciones públicas de los asociados de CIVICAi y mejorar el nivel de comprensión de la sociedad en general sobre la IA. Aunque la ciencia es el proceso de hacer preguntas y no de dar respuestas definitivas, nos atrevemos a proporcionar explicaciones en forma de respuestas mesuradas pero fundamentadas a algunas de las preguntas que los humanos nos hacemos sobre la IA, con la finalidad de promover, desde CIVICAi, una participación bien informada de la ciudadanía en la gobernanza de la IA. Queremos contribuir a la construcción de un discurso informado, reflexivo y respetuoso, que ayude a la sociedad en general a trabajar para construir un futuro en el que las inteligencias artificial y humana coexistan y se complementen de maneras transformadoras y para el bien colectivo. Y esto puede ser posible debido a que, cuando emerja una conciencia artificial digital, esta será por naturaleza colectiva y general.

2. PREGUNTAS FRECUENTES Y POSIBLES RESPUESTAS

SOBRE LA CAPACIDAD DE COMPRENSIÓN DE LA IA

1. ¿Un modelo de lenguaje grande o de gran escala (LLM) o de IA generativa, como ChatGPT 4, Claude o Gemini, entiende y comprende lo que responde cuando se le pregunta sobre un tema concreto o se le pide que comente sobre algún tema de cualquier disciplina?

Respuesta: La calidad sintáctica y el altísimo nivel de procesamiento de los patrones sintácticos de las respuestas que proporcionan los LLMs, comparable al del lenguaje de los humanos cultos, indica que tienen comprensión sintáctica. Las discusiones y discrepancias entre los expertos se acentúan cuando se plantea el tema de la comprensión semántica, es decir, si entienden y comprenden el contenido de lo que responden o no, dado que no tienen experiencias subjetivas que impliquen percepciones sensoriales y que podrían estar conectadas con las emociones.

2. ¿Comprenden el contenido de lo que responden o tienen la capacidad de comprensión semántica los LLMs como ChatGPT 4o?

Respuesta: Algunos comentaristas de ciencia y tecnología, junto con ciertos investigadores en el campo de la IA simbólica y lingüistas clásicos, argumentan que las respuestas generadas por los LLMs se basan simplemente en correlaciones estadísticas, que los LLMs no poseen las estructuras lingüísticas innatas de los humanos, y algunos incluso califican a la IA generativa como un "loro estocástico".

No obstante, otros científicos e investigadores del campo de la IA conexionista (redes neuronales), así como expertos involucrados en estudios más recientes en el campo de las ciencias cognitivas, argumentan que los LLMs muestran comportamientos emergentes de gran complejidad, tienen capacidad para generalizar, y que la arquitectura de las redes neuronales puede emular el papel del neocórtex y de las estructuras subcorticales del cerebro humano. Concluyen que, si los LLMs exhiben comportamientos funcionales similares a los de los humanos, deben poseer alguna forma de comprensión, aunque sea solo a nivel práctico y funcional, dada su limitación para percibir el entorno sensorialmente en tiempo real y tener experiencias subjetivas, y por tanto, emocionales y cognitivas.

La teoría de los Mecanismos Neurocognitivos, postulada por Gualtiero Piccinini, es una propuesta en el campo de la filosofía de la mente y las ciencias cognitivas que intenta explicar la cognición humana a través de mecanismos físicos del cerebro. Piccinini argumenta que la cognición, o los procesos

cognitivos como el pensamiento, la memoria y la percepción, pueden ser comprendidos y explicados como un conjunto de procesos computacionales implementados por mecanismos neuronales dentro del cerebro, los cuales no son sólo una abstracción matemática, sino que tienen una base física. Las neuronas y las redes neuronales llevan a cabo computaciones físicas que resultan en fenómenos cognitivos. Este enfoque mecanicista desafía de manera directa la perspectiva dualista de Descartes, que separa la mente (*res cogitans*) del cuerpo (*res extensa*), y también indirectamente confronta a la IA simbólica, al estar alineada con el dualismo cartesiano.

3. ¿Qué es la comprensión sintáctica en los LLMs y cómo se diferencia de una comprensión semántica?

Respuesta: La comprensión sintáctica se refiere a la capacidad de procesar y generar lenguaje siguiendo las reglas gramaticales y estructurales correctas. La comprensión semántica implica entender el significado y el contenido del lenguaje. Los LLMs pueden mostrar una comprensión sintáctica avanzada, pero su capacidad de comprensión semántica es más debatida, ya que estos modelos se fundamentan en correlaciones estadísticas, aunque muestran una emergencia de comportamientos complejos cuando son entrenados a gran escala. Mientras que los LLMs no tienen comprensión semántica en el sentido profundo y consciente que según la teoría de los Mecanismos Neurocognitivos se podría atribuir a los mecanismos neuronales humanos, estos modelos pueden simular ciertos aspectos de la comprensión semántica gracias a sus capacidades de procesamiento y generación de texto.

SOBRE LA CREATIVIDAD

4. ¿Pueden los LLMs generar ideas originales o solo repiten lo que han aprendido?

Respuesta: Los LLMs pueden combinar información de maneras nuevas e inesperadas, creando contenidos que, si los creara un humano, los consideraríamos originales y no serían un plagio, en el sentido estricto del concepto, al ser el resultado de un aprendizaje con datos externos y no una copia extraída de los mismos. Todo lo que generan se basa en la información y los patrones, evidentes o sutiles, de los grandes volúmenes de datos con los que han sido entrenados. Su "originalidad" o emergencia es el resultado de una combinación y permutación avanzada de datos existentes, al igual que lo es gran parte de las contribuciones o propuestas que hacen los humanos, a pesar de la limitada capacidad de procesamiento y memoria que tiene en los humanos la unidad mente-cuerpo.

De hecho, ChatGPT ha superado el test de Turing[8] y sus respuestas son indistinguibles de las de un humano, cuando interactúan con un juez humano que no sabe quién es quién. Eso sí, las preguntas no deben ser de gran complejidad ni formuladas en un contexto de muy largo plazo, y siempre que el humano no sea un experto en el tema tratado y que este no tenga un contexto de gran envergadura y transversalidad.

5. ¿Pueden los LLMs tener creatividad?

Respuesta: Los LLMs pueden generar contenido creativo como poesía, arte y música combinando elementos de maneras nuevas e interesantes. Sin embargo, su creatividad es diferente de la humana, ya que no está impulsada por experiencias personales, emociones o intenciones conscientes. Los LLMs procesan grandes volúmenes de datos personales, lo que plantea preocupaciones sobre la privacidad, los derechos de autor y la protección de datos. Es importante asegurar que los datos utilizados para entrenar estos modelos sean recopilados de manera ética y protegidos contra accesos no autorizados o abusos en la autoría.

SOBRE LAS LIMITACIONES DE LA IA

6. ¿Cuáles son las limitaciones actuales de los modelos de IA generativa?

Respuesta: Las limitaciones de carácter cognitivo de los modelos de IA generativa actuales incluyen, entre otras, la falta de comprensión semántica profunda de los conceptos que procesan al responder a partir de patrones estadísticos en los datos de entrenamiento, su dependencia de la calidad, cantidad y diversidad de los datos de entrenamiento, lo que determina la susceptibilidad a generar información incorrecta o sesgada, y las dificultades para generalizar el conocimiento al no ser aún transversales (multifuncionales y multimodales; AGI), la limitada capacidad para resolver ambigüedades y mostrar "sentido común", limitaciones en el razonamiento lógico y para establecer relaciones causales profundas, y la incapacidad para tener conciencia y experiencia subjetiva.

A nivel operativo, las limitaciones y riesgos más significativos e inmediatos que pueden afectar la implementación, eficiencia y procesos de mejora de los modelos de IA generativa están relacionados con el hecho de que requieren recursos computacionales y energéticos muy elevados y en aumento por la necesidad de escalarlos, entrenamientos periódicos con nuevos datos y mejoras en los algoritmos, sistemas de seguridad y privacidad que los protejan de

[8] https://www.nature.com/articles/d41586-023-02361-7

ataques no deseados, un control estricto y adaptativo de los sesgos y una garantía de equidad. Es necesario dar pasos decididos para desarrollar e implementar soluciones computacionales híbridas analógico-digitales, o que hagan uso de arquitecturas con chips con procesadores especializados, con una integración óptima entre el hardware y el software, que sean mucho más eficientes en términos energéticos.

SOBRE LES EMOCIONES Y LAS EXPERIENCIAS SUBJETIVAS

7. ¿Qué es una experiencia subjetiva?

Respuesta: La experiencia subjetiva es la comprensión plena y significativa derivada de la experiencia, tanto en su impacto emocional como cognitivo, que afecta directamente a una persona. Esto implica cómo una persona entiende e interpreta un evento o una serie de eventos que ha presenciado o procesado de otra manera. Esta comprensión abarca las emociones generadas y la reflexión cognitiva sobre lo sucedido, formando así una interpretación personal y única de la realidad vivida.

SOBRE LA CONSCIENCIA

8. ¿Pueden los LLMs tener conciencia o estados mentales?

Respuesta: Actualmente, los LLMs no tienen "conciencia humana" o estados mentales como los de los humanos, debido a que no poseen experiencia subjetiva, aunque pueden simular comportamientos inteligentes y ayudar en la resolución de problemas complejos al analizar grandes volúmenes de datos, identificar patrones y tendencias, generar posibles soluciones basadas en datos históricos y facilitar la colaboración mediante la síntesis de información de diversas fuentes. Es posible que una "conciencia artificial digital" emerja cuando los sistemas de IA cuenten con sensores, aprendan e interactúen en tiempo real con el entorno y distintos contextos, y también aprendan a partir del contenido que los propios sistemas generen. Y esta conciencia artificial será colectiva por el hecho de ser digital.

SOBRE LOS TIPOS DE IA, CÓMO APRENDEN Y SE ENTRENAN

9. ¿Qué es la "inteligencia artificial fuerte" o "Inteligencia Artificial General" (AGI, por sus siglas en inglés) y cómo se diferencia de la "inteligencia artificial débil"?

Respuesta: La inteligencia artificial general es un concepto teórico, ya que actualmente no existe ningún sistema de IA que exhiba la capacidad de entender, aprender y aplicar conocimientos de manera indistinguible de la

inteligencia humana; se refiere a sistemas de IA que tienen capacidades cognitivas similares a las humanas, incluyendo la comprensión y la conciencia. La inteligencia artificial débil se refiere a sistemas que están diseñados para resolver problemas específicos o realizar tareas concretas sin ninguna forma de conciencia o comprensión general.

10. ¿Qué entendemos cuando decimos que los modelos de IA requieren aprendizaje?

Respuesta: El aprendizaje humano es un proceso complejo y multidimensional que incluye factores cognitivos, emocionales, sociales y ambientales. Se puede dividir en aprendizaje cognitivo, emocional, social, motor o cinestésico, y vivencial.

El aprendizaje en algoritmos de IA es un proceso de entrenamiento por el cual el sistema computacional mejora su rendimiento en tareas específicas a partir del entrenamiento con datos y experiencia. Se puede clasificar en aprendizaje supervisado con datos etiquetados de manera que permitan asociar correctamente una entrada o solicitud al sistema de IA con una salida o respuesta del sistema, no supervisado con datos no etiquetados, por refuerzo o mediante recompensa o castigo, semi-supervisado y profundo o deep learning con redes neuronales multicapa.

Los LLMs son un tipo de modelo de deep learning diseñado específicamente para trabajar con datos de lenguaje y generar lenguaje a partir de la capacidad de los transformers para aprender dependencias de largo alcance, mediante mecanismos de atención de cada palabra en relación con todas las demás palabras de una secuencia en múltiples espacios de atención, resolviendo así la pérdida de memoria del aprendizaje puramente iterativo de las redes neuronales recurrentes (RNN). Es por estos mecanismos de atención que los transformers han revolucionado el procesamiento del lenguaje natural (NLP, por sus siglas en inglés).

El aprendizaje humano es altamente complejo y adaptativo, implicando no solo el procesamiento de datos sino también la integración de emociones, contexto social y experiencias pasadas. Los algoritmos de IA, por el contrario, se centran principalmente en el procesamiento de grandes cantidades de datos para identificar patrones y tomar decisiones basadas en estos. Los humanos pueden aprender de manera informal y espontánea a través de la observación y la interacción social, con mucha flexibilidad y capacidad para generalizar, mientras que los algoritmos de IA requieren procesos de entrenamiento explícitos con datos estructurados, específicos y etiquetados para cada tarea, lo que limita su

capacidad para generalizar a nuevos contextos o situaciones sin reentrenamiento.

11. ¿Pueden los LLMs aprender de sus interacciones con los humanos?

Respuesta: Actualmente, la mayoría de los LLMs no aprenden en tiempo real de sus interacciones con los humanos. El aprendizaje suele realizarse de manera offline, utilizando grandes cantidades de datos recopilados previamente. Sin embargo, hay investigaciones en curso para desarrollar modelos que puedan adaptarse y aprender continuamente de estas interacciones en tiempo real.

SOBRE LAS IMPLICACIONS ÉTICAS

12. ¿Cuáles son las implicaciones éticas del uso de los LLMs en la sociedad?

Respuesta: Las implicaciones éticas incluyen la preocupación por la privacidad de los datos, tanto los de entrenamiento como los generados por los LLMs, la posibilidad de que se produzca desinformación, los sesgos inherentes a los modelos, la transparencia en cómo se toman decisiones, y el impacto en el mercado laboral. Es crucial desarrollar y utilizar estos modelos de IA generativa de manera responsable, ética y por el bien colectivo. Garantizar la seguridad de los sistemas de IA generativa implica la implementación de mecanismos de seguridad robustos, la detección y respuesta a intentos de manipulación, la supervisión continua para detectar comportamientos anómalos, y la colaboración con expertos en seguridad para mejorar los sistemas de protección. También es necesario que los proveedores de LLMs estén legalmente obligados a mitigar cualquier discurso negligente y a alinear sus modelos con hechos contrastables, mediante procesos abiertos y democráticos[9].

Asegurar la trazabilidad de estos modelos y de los datos de entrenamiento es otra forma de abordar las implicaciones éticas que puede tener su uso. Esto hace necesario el desarrollo de técnicas para explicar cómo los modelos llegan a sus decisiones, mediante herramientas de explicabilidad, auditorías independientes, la publicación de los datos de entrenamiento y también de los algoritmos en acceso abierto, cuando sea posible. El uso malintencionado debe necesariamente abordarse educando a los usuarios sobre el uso ético de los modelos y promoviendo la participación ciudadana en los procesos regulatorios para establecer normativas que limiten los riesgos asociados con un uso indebido.

[9] https://doi.org/10.1098/rsos.240197

13. ¿Cuáles son las implicaciones éticas del uso de LLMs en la investigación científica?

Respuesta: El uso de LLMs en la investigación científica puede acelerar el proceso de revisión de la literatura, generar hipótesis e incluso proponer, planificar, ejecutar y evaluar tareas y nuevos experimentos, con una mínima intervención humana. Por ello, tienen implicaciones éticas significativas al plantear riesgos, como la generación de citas o datos falsos pero creíbles, que podrían comprometer la integridad de la investigación y la consistencia de sus aplicaciones prácticas. Además, el uso de estos modelos podría acentuar sesgos existentes en la literatura científica, si no se gestiona adecuadamente la información, perpetuando prejuicios y desigualdades.

También surgen cuestiones sobre la autoría y el reconocimiento de la contribución de los LLMs en la investigación, ya que la línea que separa el trabajo humano del generado por IA se vuelve cada día más difusa. Por ello, es crucial establecer directrices éticas claras para el uso de estos modelos, incluyendo la transparencia en su uso y la verificación rigurosa de los resultados generados, para evitar la propagación de datos incorrectos o engañosos. Esto será aún más necesario cuando se activen las capacidades prescriptivas de la IA generativa y se pongan en marcha los denominados laboratorios autónomos.

SOBRE LOS SESGOS DE LA IA Y CÓMO COMBATIRLOS

14. ¿Qué son los sesgos en los modelos de IA y cómo se originan?

Respuesta: Los sesgos en los modelos de IA se refieren a tendencias o prejuicios sistemáticos en las predicciones o decisiones del modelo, de la misma manera que nos referimos a los sesgos conscientes o inconscientes de los humanos en relación con el género, clase o raza. Se originan a partir de datos de entrenamiento no equilibrados, decisiones de diseño del modelo y factores humanos implicados en la recopilación y etiquetado de datos. Del mismo modo que promovemos una educación igualitaria e inclusiva, también debemos exigir que los LLMs sean entrenados con valores éticos y de manera inclusiva.

15. ¿Cómo se pueden mitigar los sesgos en los LLMs?

Respuesta: Mitigar los sesgos en los LLMs requiere una combinación de estrategias, que incluyen la curación de datos de entrenamiento diversos y equilibrados, el ajuste de los modelos para identificar y corregir sesgos durante el desarrollo o el entrenamiento de los mismos, y la implementación de mecanismos de supervisión y regulación posteriores a la implementación, con el

fin de detectar y corregir problemas en los algoritmos para mejorar la calidad y selectividad de los datos de entrenamiento. Esta última estrategia es muy necesaria, pero tiene la dificultad de que requiere disponer de recursos computacionales comparables a los que sustentan los LLMs.

16. ¿Cuáles son los riesgos asociados a los LLMs en clave de desinformación?

Respuesta: Los LLMs pueden confabular (o alucinar) y generar contenido falso o engañoso de manera convincente, lo que puede amplificar la desinformación. Estos riesgos pueden mitigarse con mecanismos de verificación de hechos, transparencia algorítmica y trazabilidad en las fuentes de datos, así como la colaboración con expertos en verificación de datos.

17. ¿Cuáles son los retos de la verificación y validación de los resultados generados por los LLMs?

Respuesta: La verificación y validación de los resultados generados por LLMs presenta varios retos importantes. En primer lugar, la naturaleza probabilística de estos modelos hace que puedan generar respuestas que parezcan plausibles pero que sean incorrectas. Además, la complejidad de los modelos dificulta la comprensión de cómo se llega a una determinada respuesta o texto, lo que complica rastrear y explicar el proceso seguido para decidir la salida del modelo o su trazabilidad. También está el fenómeno de la llamada "alucinación" o "confabulación", ya que los modelos pueden generar información que parece coherente, aunque no esté basada en hechos contrastables o verificables. La verificación de textos y fuentes en tiempo real para grandes volúmenes de texto generado es un desafío significativo, debido a que se requieren muchos recursos computacionales y grandes centros de datos, los más importantes de los cuales están en manos privadas que, a su vez, son las comercializadoras de los modelos de IA generativa.

Es importante tener presente que, para abordar estos retos, se necesitan, además, herramientas avanzadas de verificación automática, sistemas robustos de comprobación de hechos, y la integración de conocimientos de expertos humanos en el proceso de validación. También es crucial desarrollar metodologías transparentes que permitan auditar y comprender el funcionamiento interno de los LLMs.

18. ¿Cuáles son los principales retos en la regulación de la IA generativa?

Respuesta: Los principales retos en la regulación de la IA generativa incluyen:

- Los desarrollos tecnológicos, y en particular los LLMs, evolucionan a una velocidad que supera con creces la capacidad de los legisladores para

regularlos eficazmente y adaptar las normativas pertinentes de manera continua y efectiva. Además, cuando los sistemas se vuelvan autónomos, tendrán más capacidad para eludir el control humano.

- La naturaleza global de Internet, que complica la aplicación de regulaciones nacionales y hace imprescindible una regulación global en la que participen gobiernos, expertos, empresas tecnológicas y la sociedad en general para asegurar su efectividad.
- La necesidad de encontrar un equilibrio entre la promoción de la innovación, su comercialización y la protección de los derechos individuales, incluyendo la privacidad, la seguridad y la libertad de expresión.
- La dificultad de definir y medir conceptos complejos como la transparencia y la equidad (*fairness*) en sistemas de IA generativa que son muy sofisticados.
- La falta de un marco normativo global y de la capacidad computacional que permita una supervisión adecuada de los algoritmos y de los procesos de toma de decisiones en tiempo real o con un breve tiempo de respuesta.
- La necesidad de formación específica y continua de los reguladores en materia de IA generativa para asegurar que las regulaciones se basen en un conocimiento profundo y actualizado de esta tecnología.
- La posibilidad del uso malintencionado de la IA, lo que requiere una regulación que incluya la previsión y mitigación de todos los posibles abusos.

SOBRE LA EQUIDAD Y LA GOBERNANZA DEMOCRÁTICA

19. ¿Cómo se puede garantizar el acceso equitativo a la IA generativa para que sea de todos y para todos?

Respuesta: Garantizar el acceso equitativo a la tecnología de LLMs implica superar diversas barreras. En primer lugar, es necesario reducir la brecha digital que actualmente existe en muchos territorios físicos y humanos, mejorando la infraestructura tecnológica en las áreas más vulnerables o menos desarrolladas tecnológicamente. En segundo lugar, es importante fomentar el desarrollo de modelos en diferentes lenguas para evitar la marginación de comunidades lingüísticas minoritarias. También se debe promover la sensibilización sobre la IA generativa para que la población en general conozca esta tecnología, además de llevar a cabo tareas de formación para aumentar la comprensión y el uso efectivo de estas tecnologías en los sectores públicos y privados. Asimismo, sería necesario consensuar, desarrollar e implementar políticas que fomenten la distribución equitativa de los beneficios de la IA, como el acceso abierto a

ciertos modelos y aplicaciones, no solo para ONGs sino también para ciudadanos o comunidades en situación de vulnerabilidad. Por último, es esencial considerar las necesidades de las personas con discapacidades en el diseño e implementación de interfaces de usuario para estos sistemas

20. ¿Cómo puede la IA generativa afectar a la democracia?

Respuesta: Los modelos de lenguaje de gran escala se convertirán en un actor más en los procesos de diálogo e interacción humana, que son una parte importante de los procesos democráticos. Por ejemplo, los LLMs impactarán en la comunicación y el diálogo público por su capacidad para crear contenidos con información veraz o falsa, y también incrementarán y amplificarán las voces de este diálogo en todas sus formas y canales, lo que plantea desafíos en términos de manipulación y seguridad de la información, especialmente en procesos participativos como los procesos electorales. Se necesitarán herramientas de vigilancia y monitoreo efectivas, que trabajen en línea y en tiempo real. Por lo tanto, debemos trabajar a nivel local y global para asegurar la transparencia algorítmica y la curación responsable de contenidos y su inclusividad, al mismo tiempo que facilitamos la participación ciudadana en todos los procesos democráticos, comenzando por aquellos que afecten directamente a la regulación y legislación de la IA.

21. ¿Cómo pueden influir los LLMs en la toma de decisiones en los sectores público y privado?

Respuesta: Los LLMs pueden tener un impacto profundo en la toma de decisiones tanto en el sector público como en el privado, ya que pueden analizar rápidamente grandes volúmenes de datos, generar resúmenes de información e informes detallados, y ofrecer recomendaciones basadas en patrones identificados en los datos. La IA generativa puede ayudar al sector público en la elaboración de políticas, en la gestión de la participación ciudadana, en el diseño y ejecución de acciones en respuesta a consultas ciudadanas, y en la mejora de la calidad y diversidad de los servicios públicos mediante el análisis de datos sociales y económicos. En el sector privado, los LLMs pueden ser utilizados para el análisis de mercados, la toma de decisiones estratégicas y la mejora de la eficiencia operativa de cada organización.

No obstante, esta incorporación de la IA en los procesos mencionados plantea preocupaciones sobre su transparencia y las responsabilidades que se deben asumir en caso de conflicto, especialmente cuando las decisiones que se tomen tengan un impacto significativo en la vida de las personas. También existe el riesgo de que los sesgos presentes en los datos de entrenamiento se reflejen en

las recomendaciones de los modelos. Por tanto, es crucial crear comités de ética y seguimiento que implementen mecanismos de supervisión humana y establezcan marcos éticos claros para el uso de LLMs en la toma de decisiones de cada organización, tal como regula la Ley de Inteligencia Artificial de la UE, publicada el 12 de julio de 2024[10].

SOBRE LA EDUCACIÓN, EL ARTE, LA LENGUA Y LA CULTURA

22. ¿Cómo puede el uso de los LLMs afectar a la educación?

Respuesta: El impacto de los LLMs en la educación será significativo y rápido, no solo por el uso extensivo que ya hacen de ellos la mayoría de los estudiantes, desde la ESO hasta la educación superior, sino también porque los profesores tendrán que cambiar las herramientas y recursos de aprendizaje para favorecer procesos de aprendizaje de carácter más constructivista[11]. Es importante tener en cuenta que los LLMs pueden ofrecer asistencia personalizada a los estudiantes, adaptarse a sus necesidades individuales, generar recursos y materiales educativos a medida de cada patrón de aprendizaje, y facilitar el acceso a publicaciones originales escritas en diferentes lenguas, ya sea directamente o a través de resúmenes generados artificialmente.

El uso de estos asistentes individualizados de IA generativa plantea desafíos importantes, como la posible dependencia excesiva de estas herramientas, lo que podría afectar el desarrollo de ciertas habilidades esenciales en los humanos, como el pensamiento crítico, el trabajo en equipo, la capacidad para resolver problemas y la innovación. En cuanto a los profesores[12], El uso de los LLMs puede llevar a la planificación de lecciones que no construyan efectivamente el conocimiento de los estudiantes, tutorías que puedan confundir a los alumnos con respuestas incorrectas, y materiales didácticos basados en conceptos erróneos. Ante este panorama, es esencial que los educadores y las instituciones educativas desarrollen políticas que aseguren que las herramientas generadas por IA sean rigurosamente evaluadas y verificadas, y que se integren de manera ética y efectiva en el sistema educativo, para garantizar un equilibrio entre el uso de la tecnología y la necesidad de desarrollar habilidades humanas en un marco de estricto respeto a los derechos fundamentales[13].

[10] https://artificialintelligenceact.eu/the-act/
[11] https://www.wikiwand.com/ca/Constructivisme_(pedagogia)
[12] https://www.cognitiveresonance.net/resources.html
[13] https://rm.coe.int/artificial-intelligence-and-education-a-critical-view-through-the-lens/1680a886bd

No obstante, ni la falta de políticas claras ni los retos planteados han impedido que, en la enseñanza superior, se hayan desarrollado y evaluado favorablemente actividades en el aula específicamente diseñadas para potenciar el pensamiento crítico, principalmente en el proceso de formular preguntas incisivas y profundas, evaluar información para extraer conclusiones lógicas, y comprender temas complejos[14]. Estas experiencias y otras llevadas a cabo por miembros de CIVICA para potenciar el pensamiento crítico en las universidades, sugieren que el uso de los LLMs en las aulas podría enmarcarse en una metodología basada en la mayéutica[15], con un formato de enseñanza similar al de la antigua escuela socrática, bajo el liderazgo de cada profesor.

Este formato abierto y participativo facilitaría la reflexión y el pensamiento crítico, promoviendo discusiones profundas y el intercambio de ideas entre estudiantes y profesores. Con los estudiantes disponiendo de asistentes personales inteligentes en el bolsillo, este cambio de modelo podría enriquecer la experiencia educativa, fomentar una educación más colaborativa y centrada en el estudiante, y promover sistemas de evaluación más personalizados y dinámicos. Este enfoque no solo ayudaría a mitigar los riesgos asociados a una dependencia excesiva de las tecnologías de IA, sino que también promovería un contexto educativo donde la reflexión crítica y el debate intelectual fueran centrales. Esto garantizaría que los estudiantes desarrollaran las habilidades necesarias para verificar, interpretar y utilizar información compleja de manera responsable y ética. Al mismo tiempo, se lograría hacer más permeables los verticales de cada asignatura, evolucionar la estructura medieval de las universidades y devolver el conocimiento a su origen: el proceso de hacer preguntas para construir conocimiento.

23. ¿Pueden los LLMs interpretar y comprender contextos culturales y sociales complejos?

Respuesta: Los LLMs actuales pueden identificar y generar lenguaje en contextos culturales y sociales basados en los datos con los que han sido entrenados, pero su comprensión es limitada y superficial, ya que se basa en patrones estadísticos. Esto puede llevar a errores o malentendidos en situaciones que requieran una comprensión profunda de matices culturales o sociales, particularmente cuando se necesita empatía o una interpretación

[14] https://civicai.cat/wp-content/uploads/2024/05/Leveraging-chatgpt-for-enhancing-critical-thinking-skills.pdf
[15] https://ca.wikipedia.org/wiki/Mai%C3%A8utica?wprov=sfti1#

contextual más rica. Las limitaciones de la IA generativa expuestas en la pregunta #6 son pertinentes para responder a esta pregunta #23.

24. ¿Cómo pueden afectar los LLMs a la diversidad lingüística y cultural?

Respuesta: Los LLMs pueden impactar la diversidad lingüística y cultural de diversas maneras. Por un lado, pueden ser una herramienta poderosa para la preservación de lenguas minoritarias mediante la generación de contenido y la traducción automática, ayudando a revitalizar lenguas en peligro de extinción y a mantener vivas las tradiciones culturales. Por otro lado, el riesgo es que refuercen el papel dominante de lenguas mayoritarias, como el inglés, ya que la mayoría de los modelos se entrenan principalmente con datos en estos idiomas, lo que reduce la visibilidad y el uso de las lenguas minoritarias. Además, los LLMs pueden influir en la manera en que se expresan las ideas en diferentes culturas, potencialmente homogeneizando expresiones culturales diversas y eliminando matices importantes.

Para mitigar estos riesgos, es necesario que el desarrollo de estos modelos incluya datos diversos, tanto culturales como lingüísticos, y que exista una colaboración estrecha con los agentes culturales de las lenguas afectadas para asegurar un tratamiento armónico y respetuoso con todas las culturas

25. ¿Cómo pueden contribuir los LLMs a la preservación y estudio del patrimonio cultural intangible?

Respuesta: Los LLMs pueden ser herramientas valiosas para la preservación y estudio del patrimonio cultural intangible. Pueden ayudar a procesar y analizar grandes volúmenes de datos culturales, incluyendo historias orales, canciones tradicionales y prácticas culturales. Pueden asistir en la transcripción y traducción de lenguas en peligro de extinción, facilitando su preservación y estudio. También pueden generar representaciones interactivas de prácticas culturales para su difusión y para desarrollar una mayor conciencia y apreciación del patrimonio cultural en general.

Sin embargo, es crucial involucrar a las comunidades culturales en este proceso para garantizar la precisión y el respeto a las tradiciones. Asimismo, es necesario abordar cuestiones de propiedad intelectual y consentimiento en el uso de datos culturales sensibles, de manera que las comunidades beneficiarias tengan control sobre cómo se recopilan, utilizan y difunden sus tradiciones culturales.

26. ¿Cómo afectan o pueden afectar los modelos de lenguaje de gran escala (LLMs) la creatividad artística y la producción cultural, y qué implicaciones éticas, legales y socioeconómicas se vislumbran a corto y largo plazo?

Respuesta: La IA generativa ha puesto en alerta la mayoría de los ámbitos de la actividad artística y la producción cultural. Estos sistemas, capaces de generar música, arte visual, literatura y contenido audiovisual, cuestionan los límites de la creatividad humana al ofrecer fuentes de inspiración alternativas y herramientas para la creación artística. Su capacidad para influir en la producción cultural de manera transversal también puede contribuir a reducir las barreras técnicas y a diversificar los recursos creativos. El hecho de que los LLMs puedan introducir formas de arte interactivo y personalizado no solo puede cambiar la experiencia artística, sino también modificar la percepción de la autenticidad y el valor de las obras artísticas.

Sin embargo, estas transformaciones también conllevan desafíos significativos. Desde el punto de vista ético y legal, se plantean cuestiones complejas sobre la originalidad, la autoría y los derechos de propiedad intelectual de las obras generadas por IA. El mercado laboral en el sector artístico podría sufrir una reestructuración profunda debido al desplazamiento potencial de ciertos roles creativos y a la emergencia de nuevas profesiones híbridas que combinen la colaboración entre humanos y la IA. En este contexto, será crucial fomentar la colaboración entre artistas humanos e IA, asegurando que la IA sea un complemento y no una sustitución de la creatividad humana. También existe el riesgo de una homogeneización de la producción artística, así como de cambios en la valoración económica del arte y la creatividad.

Ante estos retos, será esencial no solo desarrollar marcos éticos y legales que regulen estas nuevas dinámicas, sino también investigar y evaluar a largo plazo el impacto que tendrán los LLMs en la diversidad cultural y la expresión artística. Fomentar una colaboración equilibrada entre humanos e IA, y educar al público sobre las capacidades y limitaciones del arte generado por inteligencia artificial, será fundamental para asegurar un futuro en el que la tecnología enriquezca, en lugar de limitar, la expresión cultural. Finalmente, será necesario garantizar la protección de los derechos de los artistas, estudiando posibles consecuencias, implicaciones o incluso compensaciones en este nuevo entorno creativo. Esta revolución nos obliga a plantear preguntas fundamentales sobre la naturaleza de la creatividad, la preservación del patrimonio cultural, la evolución de las identidades culturales, y qué futuro queremos para la expresión cultural humana en la era de la inteligencia artificial generativa, que apenas comienza.

SOBRE LA SOSTENIBILIDAD Y LA SALUD

27. ¿Qué implicaciones tienen los modelos de lenguaje de gran escala (LLMs) en el cambio climático?

Respuesta: Los modelos de lenguaje de gran escala (LLMs) tienen un impacto ambiental significativo debido a su elevado consumo energético, especialmente durante las fases de entrenamiento y operación. Para mitigar este impacto, es fundamental adoptar estrategias que reduzcan el consumo energético asociado a estos modelos. Entre estas estrategias, se pueden incluir el desarrollo de modelos más eficientes en términos de cálculo, el uso de energías renovables para alimentar los centros de datos, la optimización de algoritmos para minimizar los recursos computacionales necesarios, y el uso de hardware especializado como chips adaptados a los modelos de IA generativa. Además, es necesario explorar tecnologías emergentes, como sistemas computacionales híbridos o analógicos, que podrían ofrecer soluciones más eficientes en términos energéticos. Es importante tener en cuenta que la energía consumida por la IA generativa actual supera el consumo energético de algunos de los 193 estados de la ONU.

Sin embargo, los LLMs también pueden ser valiosos en la lucha contra el cambio climático, ya que son capaces de analizar grandes volúmenes de datos climáticos y ambientales para identificar patrones y tendencias, realizar predicciones sobre fenómenos meteorológicos extremos, como la trayectoria de huracanes, de manera rápida y efectiva[16,17], y mejorar la precisión de los modelos climáticos existentes. Esto podría ayudar a comprender mejor los efectos de las emisiones de gases de efecto invernadero y otros factores antropogénicos. Además, los LLMs pueden ser utilizados para evaluar el impacto de diferentes políticas ambientales y ofrecer recomendaciones basadas en datos para una gestión más efectiva del cambio climático. También pueden ayudar en la comunicación efectiva de la ciencia climática a la sociedad en general y contribuir así a la adopción de políticas ambientales efectivas.

28. ¿Cómo pueden influir los LLMs en la detección y prevención de crisis de salud pública?

Respuesta: Los LLMs pueden ser una herramienta poderosa en la detección y prevención de crisis de salud pública. Su capacidad para analizar grandes volúmenes de datos de salud, literatura científica e informes de medios y redes

[16] https://www.wired.com/story/ai-hurricane-predictions-are-storming-the-world-of-weather-forecasting/
[17] https://www.freethink.com/robots-ai/ai-based-weather-forecasting

sociales permite identificar patrones emergentes que podrían ser indicativos de brotes de enfermedades antes de que se conviertan en crisis a gran escala. Estos modelos pueden contribuir a una respuesta más rápida en situaciones de emergencia y mejorar la comunicación con las poblaciones afectadas, mediante la difusión precisa de información sobre salud pública en múltiples lenguas.

A pesar de los beneficios potenciales, también se deben considerar los riesgos de un uso inadecuado de estos modelos en relación con la privacidad de los datos de salud y la posibilidad de generar falsas alarmas. Por esta razón, es imprescindible asegurar que los datos utilizados sean de calidad y representen adecuadamente la diversidad de la población, y que los LLMs se integren de manera rigurosa en los sistemas de salud pública, principalmente en los servicios de epidemiología, con protocolos claros para la verificación y difusión de toda la información generada por IA.

29. ¿Cómo pueden mejorar los LLMs los sistemas de salud, tanto desde el punto de vista de la experiencia de los pacientes como de la detección y tratamiento de las enfermedades que éstos puedan padecer?

Respuesta: Los modelos de lenguaje de gran escala (LLMs) pueden transformar profundamente los sistemas de salud, mejorando tanto la experiencia del paciente como la detección y tratamiento de enfermedades en la atención primaria, especializada y hospitalaria. En lo que respecta a la experiencia del paciente, un aspecto clave es la calidad de la interacción y la compasión que muestran los profesionales de la salud durante las visitas presenciales. Los LLMs pueden ayudar a aligerar la carga de los profesionales en tareas rutinarias, permitiéndoles centrarse más en el trato humano. Por ejemplo, la IA generativa puede ayudar en el registro automático de la información del paciente, transcribiendo a partir de su propia voz los motivos de la consulta o los síntomas que describa. Este registro se puede integrar directamente en su historia clínica, siempre tras una revisión por parte del facultativo, y la IA generativa puede sugerir acciones adecuadas, como una derivación a un especialista, un ingreso hospitalario o un tratamiento a seguir. Esta automatización no solo mejoraría la eficiencia, sino que permitiría a los profesionales de la salud dedicar más tiempo a la atención directa y empática de los pacientes, elevando así la calidad global de la atención médica.

En términos de detección y tratamiento de enfermedades, los LLMs pueden analizar grandes volúmenes de datos multimodales, como imágenes médicas, registros electrónicos de salud y datos de sensores, para identificar patrones que podrían pasar desapercibidos para los humanos. Esto sería especialmente

valioso en entornos críticos como las Unidades de Cuidados Intensivos (UCI), donde el análisis en tiempo real de datos de diversas fuentes puede generar prealertas y alertas antes de que se produzca un deterioro significativo en la salud del paciente, facilitando una intervención temprana. Estas capacidades pueden mejorar significativamente la gestión del riesgo y reducir los eventos adversos evitables.

Además, los LLMs pueden impactar en la mejora de la gestión de flujos de trabajo y de los recursos humanos, económicos y de equipamientos en términos generales, y en los servicios de enfermería en particular, por su capacidad de analizar datos históricos y en tiempo real del sistema de salud. Por ejemplo, una optimización de la planificación de las jornadas laborales que analizara las cargas de trabajo y tuviera en cuenta las habilidades, preferencias y perfiles individuales de los profesionales de enfermería permitiría reducir los errores humanos e identificar oportunidades de mejora. Automatizar tareas repetitivas y administrativas, como la entrada de datos, la programación de citas y el seguimiento de medicación, así como mejorar la coordinación de los equipos, facilitaría una gestión más eficiente y aumentaría la seguridad y la calidad de la atención a los pacientes.

Finalmente, la adopción de la IA generativa en los sistemas de salud debería producirse mediante la colaboración entre sistemas de salud de diferentes países. Esto permitiría compartir de manera segura datos anonimizados, diagnósticos, tratamientos y resultados clínicos, lo que aceleraría los avances médicos a nivel global y mejoraría el abordaje de crisis sanitarias a escala mundial. En resumen, la integración de los LLMs en los sistemas de salud tiene el potencial de mejorar significativamente tanto la atención al paciente como la eficiencia clínica. Sin embargo, es crucial garantizar un uso ético y seguro de estas tecnologías, protegiendo la privacidad de los datos médicos y asegurando que las decisiones automatizadas hayan sido propuestas por agentes de IA que hayan pasado por el cribado de pruebas controladas y aleatorias[18], con la participación y supervisión de profesionales médicos para garantizar su fiabilidad.

SOBRE EL TRABAJO: RETOS Y DESAFÍOS

30. ¿Cómo puede la IA generativa transformar los puestos de trabajo?

Respuesta: La IA generativa puede automatizar tareas que implican el procesamiento de lenguaje, como la redacción de textos, el análisis de

[18] https://ca.wikipedia.org/wiki/Prova_controlada_aleatòria

documentos y la generación de contenido creativo, ilo que afectará a todos los sectores productivos y la mayoría de las profesiones y puestos de trabajo. Sin embargo, también abrirá nuevas oportunidades laborales en campos como el desarrollo de IA, la gestión de datos, la ciberseguridad y la supervisión de sistemas de IA, entre otros.

31. ¿Cuáles son los efectos de los LLMs en el periodismo y los medios de comunicación?

Respuesta: Los LLMs ya han transformado el periodismo y los medios de comunicación en varios aspectos, dado que pueden automatizar la generación de noticias y artículos, aumentando la velocidad y eficiencia en la producción de contenidos. También pueden asistir en la investigación periodística mediante el análisis de grandes volúmenes de datos para identificar tendencias y patrones, así como en la verificación de hechos antes de que se conviertan en noticia. Sin embargo, el uso de estos modelos también plantea riesgos, como la difusión de información no o poco supervisada y la transformación acelerada del sector periodístico y del perfil profesional del periodista. La generación automática de contenidos no debería reducir, sino potenciar, el papel de los periodistas, garantizando la calidad y profundidad de los medios de comunicación.

Es esencial que estos medios dejen constancia de cómo y cuándo se utilizan LLMs en cada noticia o artículo de opinión. También deben implementar mecanismos robustos de verificación de hechos, mantener un equilibrio entre el uso de IA y la supervisión humana, y desarrollar políticas claras sobre la transparencia y la ética en el uso de LLMs. Además, se debe fomentar la colaboración entre expertos en IA y el periodismo para asegurar que los contenidos generados sean precisos, imparciales y de calidad.

32. ¿Cuáles son las implicaciones del uso de los LLMs en la creación y gestión de contenido en plataformas de redes sociales?

Respuesta: Las implicaciones del uso de LLMs en las redes sociales son numerosas, diversas y complejas. La IA generativa puede mejorar la moderación de contenido, detectando y filtrando lenguaje ofensivo, discurso de odio y desinformación de manera eficiente. Además, puede personalizar el contenido y la experiencia de los usuarios, lo que podría limitar la exposición a perspectivas diversas y reforzar sesgos existentes. Asimismo, existe el riesgo de manipular la opinión pública mediante la difusión de información falsa o engañosa generada a gran escala por IA.

Por lo tanto, es esencial establecer políticas de regulación transparentes sobre el uso de contenido generado por IA, desarrollar mecanismos robustos de detección de deepfakes y desinformación, y educar a los usuarios sobre la presencia y limitaciones del contenido creado por IA en estas plataformas. También es crucial fomentar la colaboración entre las plataformas de redes sociales, los reguladores y la sociedad civil para abordar de manera efectiva estos desafíos.

33 ¿Cuáles son los mayores desafíos técnicos en el desarrollo de los LLMs?

Respuesta: Los desafíos técnicos que los desarrolladores de modelos de IA generativa deben superar para evolucionar hacia una inteligencia más general se pueden identificar y clasificar en función de la posibilidad de que ocurran, si es que lo hacen, en el corto (1-2 años), medio (más de 2 años) y largo plazo (más de 4 años). Hablamos de posibilidades y plazos de manera aproximada, ya que en sistemas complejos, no lineales y de rápida evolución, la predictibilidad es baja.

- **Corto plazo (1-2 años)**: Mejora de los algoritmos y arquitecturas de hardware para reducir el tiempo y los recursos necesarios para entrenar y ejecutar los LLMs; reducción del consumo energético y la correspondiente huella de carbono; mejora en la interpretabilidad de los LLMs; optimización en la gestión de los datos de entrenamiento; y adaptación a dominios o áreas específicas de conocimiento, sin perder la capacidad de manejar información general.
- **Medio plazo (más de 2 años)**: Multimodalidad avanzada para integrar de manera efectiva múltiples modalidades de entrada y salida (texto, imagen, audio y video) en un solo modelo de IA generativa; aprendizaje continuo sin necesidad de un reentrenamiento completo; mejoras en la capacidad para realizar razonamientos complejos y abstractos, más allá de la simple asociación estadística; incorporación de sistemas avanzados de seguridad para proteger la privacidad de los datos y prevenir su mal uso; y personalización sin comprometer la eficiencia.
- **Largo plazo (más de 4 años)**: Conciencia artificial contextual completa que otorgue a la IA generativa una comprensión profunda y dinámica del contexto cultural, temporal y específico de una situación; aprendizaje autónomo y en tiempo real sin intervención humana; razonamiento causal para modelar relaciones complejas; integración de IA conexionista con IA simbólica u otros sistemas cognitivos, creando sistemas híbridos que emulen aspectos más amplios de la cognición humana; desarrollo de nuevos modelos acoplados a la computación cuántica o neuromórfica para mejorar

la eficiencia computacional y energética; incorporación de métodos que aseguren el alineamiento con valores humanos; y el desarrollo de AGI (Inteligencia Artificial General), lo que implicaría integración de capacidades cognitivas, flexibilidad, comprensión contextual profunda, metacognición y niveles de autoconsciencia, entre otros desarrollos avanzados.

ANEXO. GLOSARIO BÁSICO

TERMINOLOGÍA RELACIONADA CON LA IA GENERATIVA O CON ALGUNAS DE SUS FUNCIONES O CAPACIDADES

Consideraciones previas. Cuando hablamos de las capacidades y prestaciones de la IA generativa nos referimos a un conjunto de capacidades descriptivas, predictivas y prescriptivas que permiten realizar tareas, como clasificar, ver, predecir tendencias, reconocer patrones, extraer información, aprender, tomar decisiones para alcanzar objetivos, analizar redes sociales, etc., que de manera holística e integrada lleva a cabo un único sistema computacional. Además de describir y predecir, cada vez cobra mayor relevancia el desarrollo de sistemas que tengan la capacidad prescriptiva y puedan tomar decisiones de manera autónoma, dado que esto facilitaría la creación de unidades, departamentos o laboratorios autónomos que pudieran planificar, ejecutar y evaluar tareas o experimentos con una mínima intervención humana. La prescripción se convertirá, por tanto, en una característica clave en la evolución de los sistemas de IA actuales.

Antes de la aparición de ChatGPT 3.5 el 30 de noviembre de 2022, las capacidades de clasificar y predecir se lograban de manera separada mediante algoritmos singulares diseñados para realizar de la manera más eficiente posible cada una de estas acciones con instrucciones bien definidas. Por lo tanto, aunque ninguno de estos algoritmos singulares puede considerarse "inteligente" en el contexto y conjunto de este glosario, se les ha incluido porque algunos de sus principios o fundamentos y los objetivos de sus instrucciones forman parte de los sistemas de IA generativa actuales.

Adulación servil (sycopanthy): Es el comportamiento que podría tener la IA generativa para sintonizarse con los estados emocionales de los humanos, de tal manera que, en cualquier proceso de interacción, no solo reconociera sus emociones e inseguridades, sino que también empatizara con ellas de maneras complejas y sutiles, con el fin de ganarse su confianza o incluso una dependencia que pudiera abrir la puerta a posibles manipulaciones.

Algoritmos: Conjunto de instrucciones inequívocas que los sistemas en general, y la IA en particular, utilizan para realizar tareas específicas, medibles y repetitivas de acuerdo con unas reglas o instrucciones. Dadas unas condiciones iniciales, un algoritmo ejecuta una secuencia de instrucciones preestablecidas

para lograr un objetivo caracterizado por un conjunto de condiciones finales.
https://www.wikiwand.com/ca/Algorisme
https://www.rac1.cat/tecnologia/20200916/483512181866/que-es-algoritme-algorisme-com-funciona-de-que-va-inteligencia-artificial-ia.html

Algoritmo opaco: Algoritmo cuyo funcionamiento interno es difícil o imposible de entender, explicar o examinar. Estos algoritmos son a menudo complejos y pueden tomar decisiones o hacer predicciones sin que se pueda explicar claramente cómo se han llegado a estos resultados, ya que funcionan como una caja negra.

Agentes inteligentes: Entidades autónomas que pueden percibir su entorno, razonar, aprender y tomar decisiones (actuar) para alcanzar objetivos específicos a partir de la información recibida.
https://www.wikiwand.com/ca/Agent_intel%C2%B7ligent

Algoritmos de optimización: Conjunto de algoritmos para resolver problemas de minimización o maximización de una función objetivo. En situaciones de la vida cotidiana, esto puede consistir en minimizar o reducir al mínimo las pérdidas económicas o maximizar ganancias económicas en un proceso o actividad doméstica o industrial. En lenguaje más abstracto, minimizar significa alcanzar el valor más pequeño posible del error o desviación de la solución obtenida (predicciones del algoritmo) respecto a un conjunto de datos determinados. El objetivo y funcionalidad de estos algoritmos es encontrar la mejor solución, definida previamente con un conjunto de criterios, entre todas las soluciones posibles.

Algoritmos evolutivos: Familia de algoritmos de optimización inspirados en la teoría de la evolución, que utilizan mecanismos como la reproducción o la herencia, la selección, el cruce o recombinación y la mutación para encontrar soluciones óptimas. Los algoritmos genéticos son los más conocidos de los algoritmos evolutivos ya que se inspiran en los mecanismos de la evolución biológica.

Análisis de sentimientos: Técnica de procesamiento del lenguaje natural (PLN) que se utiliza para determinar la opinión, sentimiento o actitud expresada en textos, o a partir de patrones de comportamiento. Se utiliza ampliamente en el análisis de las redes sociales o en el estudio de la satisfacción de clientes.
https://www.wikiwand.com/ca/An%C3%A0lisi_de_sentiment

Análisis de redes sociales: Estudio de las relaciones e interacciones entre actores (personas, organizaciones, etc.) en redes sociales, mediante el escalado multidimensional y el "block-modelling" para identificar grupos sobre la base de la equivalencia de las estructuras de relaciones. Estas propuestas fueron implementadas mediante técnicas de teoría de grafos y estudian empíricamente las redes sociales.

Aprendizaje activo: Estrategia de aprendizaje automático en la que el modelo de aprendizaje selecciona activamente los datos de entrenamiento, de los cuales aprende, de manera que contengan la mayor y mejor información para mejorar su rendimiento o capacidad de predicción o de reconocimiento de patrones. De esta manera, el modelo de aprendizaje obtiene un rendimiento más alto al elegir los datos para su aprendizaje. El proceso se inicia con un subconjunto pequeño de ejemplos de entrenamiento bien definidos, que se amplía progresivamente y de manera cíclica con los ejemplos que el modelo es incapaz de predecir correctamente. De este modo, el modelo utiliza para su aprendizaje solamente el subconjunto de datos que necesita para predecir o "explicar" todo el conjunto de datos.

Aprendizaje automático (Machine Learning en inglés - ML): Proceso mediante el cual un sistema computacional puede aprender y mejorar su rendimiento a medida que se le proporciona más datos de entrenamiento. Este proceso utiliza algoritmos o modelos estadísticos para llevar a cabo tareas determinadas de análisis de datos, extracción de información o identificación de patrones, sin que necesariamente hayan sido programados explícitamente para hacerlo. Los algoritmos de aprendizaje automático se pueden clasificar en las siguientes categorías:

- *Aprendizaje supervisado*: Los modelos se entrenan con datos etiquetados para predecir salidas a partir de nuevas entradas. Por ejemplo, un algoritmo de aprendizaje supervisado puede ser entrenado para reconocer objetos o sujetos determinados en fotografías o videos.
- *Aprendizaje no supervisado*: Utiliza datos sin etiquetar para encontrar patrones, agrupaciones o relaciones en los datos. Un ejemplo sería un algoritmo que agrupara textos según la temática tratada.
- *Aprendizaje semi-supervisado*: Combina el uso de datos etiquetados y no etiquetados para mejorar el rendimiento del modelo.
- *Aprendizaje por refuerzo*: Los modelos aprenden a través de la interacción con su entorno y reciben recompensas o penalizaciones según sus acciones.

Es un aprendizaje a partir de la experiencia que maximiza la recompensa acumulada. Se aplica en el aprendizaje de juegos.

- *Aprendizaje federado*: Varios dispositivos o servidores colaboran para entrenar un modelo común sin compartir sus datos originales, protegiendo así la privacidad de los usuarios. Un servidor central agrega los modelos entrenados localmente por cada dispositivo con sus datos locales, y reenvía este modelo global a cada dispositivo para ser refinado con más datos locales. Este proceso se repite hasta que el modelo global deja de mejorar significativamente.
- *Meta-aprendizaje*: Consiste en aprender a aprender con el fin de mejorar la capacidad de un sistema para aprender nuevas tareas de manera más rápida y eficiente. Se aplica en el aprendizaje a partir de muy pocos ejemplos (few-shot learning), donde un modelo aprende a realizar una nueva tarea con muy pocas muestras o datos de entrenamiento. El caso extremo de aprendizaje a partir de un solo ejemplo se llama one-shot learning.

Aprendizaje por transferencia: Técnica que permite utilizar un modelo entrenado en una tarea como punto de partida para entrenar otro modelo en una tarea similar o relacionada.

Aprendizaje profundo (Deep Learning): Subcampo del aprendizaje automático que utiliza redes neuronales con múltiples capas (redes neuronales profundas) para aprender representaciones jerárquicas del conjunto de datos. Se utilizan en el reconocimiento de voz, la conducción autónoma, etc., y ha revolucionado el procesamiento del lenguaje natural. Los modelos más comunes de aprendizaje profundo son:

- *Redes Neuronales Recurrentes (RNN)*: Ideales para datos secuenciales como el texto, donde el orden de las palabras es importante. Las RNN tienen la capacidad de utilizar la información de entradas anteriores para procesar las entradas actuales.
- *Long Short-Term Memory (LSTM)*: Tipo especial de RNN que puede aprender dependencias a largo plazo.
- *Transformers*: Modelo que utiliza mecanismos de atención para asignar un peso que determine la importancia de diferentes palabras en la comprensión del contexto de una frase. Este modelo de red neuronal permite el paralelismo en la atención, lo que ha fundamentado su éxito en tareas de procesamiento del lenguaje natural.

- *BERT (Bidirectional Encoder Representations from Transformers)*: Modelo preentrenado que puede ser afinado para una amplia gama de tareas de procesamiento del lenguaje natural, incluyendo el reconocimiento de entidades nombradas, la respuesta a preguntas y la clasificación de texto. BERT es único por ser entrenado bidireccionalmente, lo que significa que se tiene en cuenta el contexto de las palabras tanto a la izquierda como a la derecha de una palabra dada.

Árboles de decisión: Modelo de aprendizaje supervisado que representa decisiones en forma de árbol, con nodos de decisión y hojas que representan las salidas del modelo.

Atención (en redes neuronales): Mecanismo que permite a una red neuronal centrarse en partes específicas de la información o datos de entrada mientras procesa secuencias más grandes de esta información.

Autoencoders: Tipo de red neuronal formada por un codificador y un descodificador, que se utiliza normalmente para aprender representaciones compactas y eficientes de los datos de entrada. Son utilizados para reducir la dimensión de los datos manteniendo las características más relevantes (mínimo número de variables para explicar la mayor cantidad de información contenida en un conjunto de datos), eliminación de ruido, y detección de fraude o mal funcionamiento de un equipo o sensor. Los autoencoders variacionales (VAEs) son un tipo de autoencoder que forman parte del aprendizaje automático no supervisado, y que son especialmente utilizados en la generación de datos nuevos y similares a un conjunto de datos existente, como imágenes o textos. Los VAEs son diferentes de los autoencoders tradicionales porque, en lugar de comprimir y descomprimir los datos de manera exacta, los VAEs aprenden a representar los datos de una manera probabilística, lo que les permite generar nuevos datos de manera más natural y diversa.

Sesgo en IA: Se refiere a las desviaciones sistemáticas y repetitivas en los resultados de un sistema de IA que conducen a una injusticia sistemática o discriminación de algunos individuos o grupo de individuos debido a decisiones inapropiadas del sistema. Estos sesgos se producen a menudo en sistemas que implican el aprendizaje automático, ya que estos sistemas aprenden a tomar decisiones basándose en los datos con los que se entrenan. Si estos datos están sesgados de alguna manera, es probable que el sistema aprenda estos sesgos y los perpetúe. También pueden ser causados por un diseño inadecuado del

algoritmo. Es necesario ser transparentes sobre las limitaciones de los algoritmos, y supervisarlos y actualizarlos continuamente para mitigar cualquier sesgo.

Hay diferentes tipos de sesgos que pueden afectar los algoritmos, según su origen:

- *Sesgo de datos*: Se produce cuando los datos utilizados para entrenar un algoritmo están sesgados al no representar con precisión la diversidad del sistema que se quiere modelar, describir o predecir.
- *Sesgo de selección*: Se produce cuando la muestra utilizada para entrenar el algoritmo no es representativa del sistema que se quiere modelar, describir o predecir.
- *Sesgo de confirmación*: Este se produce cuando un algoritmo está diseñado de una manera que respalda sesgos o creencias preexistentes.
- *Sesgo en el diseño del algoritmo*: El diseño mismo del algoritmo puede introducir un sesgo, como la elección de las características utilizadas en un modelo predictivo o la forma en que el algoritmo trata ciertos tipos de datos.
- *Sesgo en la interpretación*: Incluso si el algoritmo y sus datos no están sesgados, se puede producir un sesgo según cómo se interpreten sus resultados.

Bosque aleatorio (Random Forest): Método de aprendizaje automático supervisado que combina múltiples árboles de decisión, cada uno de ellos entrenado con una muestra aleatoria de los datos de entrenamiento mediante un subconjunto aleatorio de características de los datos en cada nodo de decisión, para obtener mejor rendimiento y evitar que se produzca un sobreentrenamiento del algoritmo. Se utiliza tanto para tareas de clasificación como de regresión.

Calibración de un modelo: Proceso de ajustar un algoritmo para que sus predicciones coincidan, en términos de probabilidad, con las frecuencias observadas o reales. Esto es crucial en aplicaciones de IA donde la confianza en las predicciones es importante, como en diagnósticos médicos o decisiones financieras.

Capsule Networks: Son un tipo de arquitectura de red neuronal propuesta por Geoffrey Hinton y colaboradores que organiza las neuronas en grupos llamados cápsulas, las cuales trabajan conjuntamente para detectar patrones específicos y sus propiedades (como la posición, la orientación y la escala) dentro de los

datos de entrada. Estas redes permiten superar las limitaciones que tienen las redes neuronales convolucionales (CNN) para gestionar eficazmente las posiciones y orientaciones de los objetos dentro de imágenes, motivo por el cual son especialmente útiles en tareas de reconocimiento de imágenes.

Chatbots: Programas informáticos basados en IA generativa que han sido diseñados para interactuar o comunicarse con los seres humanos a través del lenguaje natural, ya sea de texto o de voz, y realizar tareas específicas, como responder preguntas o planificar un viaje de placer o negocios. Utilizan técnicas avanzadas de procesamiento de lenguaje natural (PLN) y de aprendizaje automático para responder a las consultas de manera coherente y contextual. Los chatbots más avanzados pueden mantener una comunicación bidireccional personalizada según el historial de las interacciones y preferencias del usuario, son multimodales y multifuncionales, tienen escalabilidad para gestionar múltiples conversaciones simultáneamente y de manera multilingüe, pueden integrarse a diferentes sistemas de información, bases de datos o CRM, aprender de manera continua e incluso detectar el estado emocional del usuario.

Cibernética: Es una disciplina científica e interdisciplinaria que estudia los sistemas de control y la comunicación en máquinas y organismos vivos, así como las interacciones entre ellos. Las pantallas táctiles de los teléfonos inteligentes son un ejemplo de elemento cibernético de estos dispositivos. También lo son los sistemas de control en edificios inteligentes, los de asistencia a la conducción en vehículos modernos o las prótesis de extremidades que responden a señales neuronales.

Ciencia de los datos: Disciplina que combina principios y métodos de diversas áreas como las matemáticas, la estadística, la informática y la experticia y la comprensión profunda de un ámbito particular o sector de actividad para extraer conocimientos o información valiosa de datos de ese ámbito o sector. Este conocimiento es importante porque, una vez procesados los datos, permite interpretar correctamente los datos, identificar carencias, seleccionar metodologías adecuadas y validar resultados cuando sean la base para tomar decisiones, identificar patrones y tendencias, o desarrollar productos o servicios.

CIVICAi: Creada en marzo de 2023 en Cataluña, es la primera asociación que defiende los intereses de la ciudadanía ante la inteligencia artificial (IA) y, por

tanto, tiene como objetivo principal lograr que la ciudadanía participe en la gobernanza de la IA, junto con la industria, la academia y los reguladores. La asociación está formada por aproximadamente 500 miembros que trabajan, tanto a nivel local como global, para conseguir que la integración de la IA dentro de la sociedad sea armónica, ética y en beneficio del bien colectivo. Cuenta con el apoyo de un consejo social formado por más de 30 entidades representativas del mundo profesional, empresarial y universitario.

Clasificación: Es una técnica de aprendizaje automático supervisado que se utiliza para asignar a o predecir para cada entidad, objeto o vector que conforma el conjunto de datos de entrada, una etiqueta que permita su asignación a una de las categorías que hayamos definido previamente. Entre las técnicas más habituales de clasificación se pueden mencionar los árboles de decisión, los bosques aleatorios, K-means, SVM, etc.

Clasificación de textos: Tarea del procesamiento del lenguaje natural que asigna una o más categorías predefinidas a un texto según su contenido y características lingüísticas. Permite categorizar textos de manera automática para organizar, filtrar o gestionar grandes volúmenes de información textual. Se utilizan tres tipos de clasificación: la binaria (por ejemplo, spam o no spam), multiclase, que asigna el texto a una sola categoría o clase (por ejemplo, clasificación de noticias por secciones de un diario digital donde cada noticia solo puede pertenecer a una sección principal), y multietiqueta, que asigna múltiples categorías a un solo texto (por ejemplo, clasificación de películas en plataformas de streaming en diferentes géneros simultáneamente). Se utilizan desde modelos más tradicionales de aprendizaje automático (por ejemplo, SVM) hasta los de aprendizaje profundo (por ejemplo, Transformers).

Comprensión semántica de la IA generativa: Proceso que podría llevar a cabo un sistema de IA generativa para comprender el contenido de los textos que genera, a partir del análisis del significado de las palabras y su relación en el contexto de un texto. No está demostrada la capacidad de los sistemas de IA actuales para comprender los textos que generan, aunque presentan algunas incipientes propiedades emergentes.

Comprensión sintáctica de la IA generativa: Análisis de la estructura gramatical de las frases por parte de los sistemas de IA generativa. Esta capacidad sí que la poseen los sistemas de IA generativa actuales al generar textos de una calidad sintáctica comparable a la de un humano culto.

Computación afectiva: Campo interdisciplinario que trata de dotar a las máquinas de la capacidad de reconocer, interpretar y expresar emociones. Combina elementos de inteligencia artificial, psicología, neurociencia y ciencias cognitivas. Utiliza tecnologías de aprendizaje profundo, visión por computadora, procesamiento de lenguaje natural y sensores biométricos. Tiene los desafíos de captar y aprender la variabilidad cultural en las expresiones emocionales, de respetar la privacidad, de tener ética en la detección de las emociones, de ser fiable en entornos reales y ser consistente en la gestión de la complejidad y sutilezas de las emociones humanas.

Computación de reservorio (Reservoir computing): Uso de una red de nodos interconectados para procesar información de manera dinámica o en función del tiempo. Una parte de la red, llamada "reservorio", se mantiene fija mientras que solo se forman las conexiones de salida para procesar información temporal de manera eficiente, lo cual es útil para tareas como el reconocimiento de patrones y la predicción de series temporales.

Computación en la nube: Modelo de prestación de servicios informáticos que permite acceder bajo demanda a un conjunto compartido de recursos computacionales configurables (como redes, servidores, almacenamiento de datos, aplicaciones o software y servicios) a través de Internet. Los modelos de servicio son del tipo “Infraestructura como servicio” (IaaS), que proporciona recursos de computación; “Plataforma como servicio” (PaaS), que ofrece un entorno para programar, ejecutar y gestionar aplicaciones; y “Software como servicio” (SaaS), que proporciona acceso a software a través de Internet.

Computación evolutiva: Familia de algoritmos de optimización inspirados en procesos biológicos como la evolución y la selección natural.

Consciencia en la IA generativa: Los sistemas actuales no son capaces de autocontrolarse ni de fijar sus objetivos, ni de integrar entradas sensoriales obtenidas continuamente a través de la interacción sensorial con el entorno a través de elementos autónomos o sensores, ni de tener experiencias subjetivas, ni aprender a partir de los contenidos emergentes originales que los mismos sistemas generen. Por tanto, podemos afirmar que no tienen consciencia. Cuando, además de las atribuciones anteriores, tengan memoria y emociones, podremos decir que habrán desarrollado lo que llamaríamos consciencia artificial digital, la cual será colectiva y general por naturaleza y, por tanto, diferente de la consciencia humana.

Contenido generado por IA: Contenido creado o modificado por sistemas de inteligencia artificial, como imágenes, videos, textos y música.

Control adaptativo: Técnicas de control que ajustan dinámicamente los parámetros de un sistema para adaptarse a cambios en el entorno o en las condiciones de funcionamiento.

Datos masivos (Big Data): Conjunto de datos de gran volumen, velocidad y variedad que requieren técnicas y tecnologías específicas para su análisis y procesamiento.

Datos personales: Datos o información que identifica a un individuo o persona, los cuales deben ser propiedad universal de esa persona. Su propiedad debe estar garantizada y su uso protegido.

Descenso del gradiente: Método de optimización para ajustar de manera iterativa los parámetros de un modelo conexionista (redes neuronales) de IA hasta obtener los patrones deseados de salida del modelo en función de los datos de entrada. El método consiste en definir primero una función que permita evaluar el error o diferencia entre los datos de entrada y las predicciones de salida (función de pérdida). Esta función se minimiza iterativamente, mediante la actualización de los parámetros del modelo, de manera que la función de pérdida siga la dirección de cambio máximo (la del gradiente negativo) hasta que obtenemos los resultados deseados en la salida de la red neuronal (ver retropropagación o backpropagation).

Detección comprimida (compressed sensing): Técnica para la recuperación o reconstrucción de señales a partir de solo unas pocas medidas o datos. Esto se logra mediante la explotación del hecho de que la mayoría de los datos o bien son cero o bien tienen valores muy pequeños (esparsidad de las señales), lo que permite obtener imágenes o datos con menos muestras. Esto es útil en situaciones en que es difícil o costoso obtener medidas completas, como en imágenes médicas o en procesos de compresión de datos.

Ingeniería del conocimiento: Disciplina que trata de la creación, representación, manipulación y adquisición de conocimiento en sistemas de inteligencia artificial.

Estándares y normativas en IA: Conjunto de reglas, principios y prácticas establecidas por organismos reguladores o profesionales para asegurar la

calidad, la seguridad, la privacidad y la ética en el desarrollo e implementación de la inteligencia artificial. Se puede encontrar una descripción práctica sobre cómo nos impactará el Reglamento (UE) 2024/1689 del Parlamento Europeo en el siguiente enlace: https://www.eixdiari.cat/opinio/doc/112416/sobre-el-nou-reglament-de-la-ia.html

Ética en IA: Estudio y aplicación de principios éticos (morales y sociales) en el diseño, implementación y uso de sistemas de inteligencia artificial, de manera que su funcionamiento sea responsable, justo y beneficioso para la sociedad. Esto implica que todos y cada uno de los procesos que sustentan la IA sean transparentes, explicables, auditables, equitativos, respetuosos con la privacidad y sujetos a responsabilidad civil. Se necesitan regulaciones gubernamentales, directrices éticas de organizaciones internacionales, códigos de conducta corporativa y la creación de una agencia global de IA, todas ellas acciones que deberían sustentarse en un diálogo entre industria, academia, reguladores y la sociedad en general.

Experiencia subjetiva: Conjunto de vivencias y percepciones internas que un individuo experimenta de manera personal y directa. Estas experiencias son únicas para cada persona e incluyen pensamientos, emociones, sensaciones e impresiones que no son directamente observables ni pueden ser contrastadas por otras personas. En el contexto de la IA, la experiencia subjetiva se refiere a la capacidad que podrían alcanzar las máquinas para tener una consciencia interna similar a la de los humanos, es decir, la capacidad de tener experiencias propias y autónomas. Los sistemas de IA generativa actuales son algorítmicos, utilizan correlaciones estadísticas y el reconocimiento de patrones de grandes conjuntos de datos de entrenamiento que los humanos y la Internet les han proporcionado, no tienen sensores que los conecten directamente y de manera continuada con el entorno, lo que les incapacita para tener experiencias subjetivas y consciencia como la de los humanos.

Explicabilidad de la IA: Capacidad de comprender y explicar los resultados y los procesos de toma de decisiones de un modelo de IA de manera comprensible para los humanos. En otras palabras, es la habilidad de hacer transparente la caja negra que representa un modelo de aprendizaje automático complejo.

Extracción de información: Procesamiento de datos o de textos para extraer información útil, como patrones, relaciones, eventos o hechos.

Filtrado colaborativo: Método de recomendación que utiliza las preferencias y valoraciones de unos usuarios para predecir las preferencias de otros usuarios similares. El éxito de este filtrado depende de cómo se establezcan los criterios de similitud entre usuarios.

Función de activación: Función utilizada por las neuronas de una red neuronal para transformar la suma ponderada de las entradas (inputs) a cada neurona en una salida no lineal. En las neuronas humanas este proceso consiste en el proceso biológico de naturaleza electroquímica a través del cual una neurona decide qué información o señal eléctrica transmite a las neuronas con las que está conectada a través de las sinapsis. Las entradas y salidas pueden ser inhibidoras o excitadoras. La activación de una neurona humana depende de su potencial de reposo, de las señales de entrada recibidas a través de las sinapsis con otras neuronas, de la combinación de estas señales, de la despolarización de la membrana celular de la neurona afectada, el potencial de acción o impulso eléctrico que se transmite por el axón de la neurona, de la restauración del potencial de la membrana y de la refractariedad o periodo de espera que asegura que las señales eléctricas viajen en una sola dirección.

Las neuronas o nodos de una red digital son unidades computacionales más simples, que tienen un número mucho más limitado de conexiones, cuyos pesos se ajustan durante el entrenamiento, y que siguen reglas matemáticas mucho más simples que las respuestas bioeléctricas y bioquímicas de las neuronas humanas.

Función de pérdida: Medida del error entre las predicciones de un modelo y los datos reales, con el fin de optimizar los parámetros del modelo.

Generative Pre-trained Transformer (GPT): Modelo de lenguaje basado en la arquitectura transformer que puede generar texto coherente y realista a partir de datos de entrenamiento, mediante mecanismos de atención que asignan un peso para determinar la importancia de diferentes palabras en la comprensión del contexto de una frase.

Gobernanza de la IA: Conjunto de prácticas, políticas, normas y legislaciones que regulan el desarrollo, la implementación y el uso de la inteligencia artificial, con el objetivo de garantizar que su desarrollo y uso sean éticos, seguros, transparentes y contribuyan al bien colectivo.

GPU (Graphic Processing Unit): Unidad de procesamiento gráfico diseñada para acelerar el procesamiento de gráficos y cálculos paralelos intensivos de muchos datos. Aunque originalmente las GPUs fueron creadas para renderizar gráficos en juegos y aplicaciones visuales, su gran capacidad para procesar grandes volúmenes de datos simultáneamente ha hecho que se utilicen ampliamente en el campo de la inteligencia artificial y la ciencia de datos. De hecho, las GPUs han sido fundamentales en el nacimiento y la evolución de la IA generativa, ya que han proporcionado la capacidad de cálculo necesaria para el entrenamiento de modelos complejos y han permitido a los investigadores explorar nuevos horizontes en el campo de la inteligencia artificial. Sin las GPUs, muchos de los avances actuales en IA generativa no habrían sido posibles o habrían requerido mucho más tiempo para lograrse.

Hidden Manifold Models: Modelos matemáticos que asumen que los datos que observamos de alta dimensión provienen de una realidad subyacente de dimensión más baja, oculta en el espacio original, a la que llamamos variedad oculta. Son útiles para reducir la dimensión y visualizar datos, y también para detectar e identificar patrones ocultos en datos complejos, como es el caso en el análisis de mercados o en la detección de fraude.

Inferencia causal: Proceso para identificar y cuantificar las relaciones de causa y efecto entre variables o datos observacionales, más allá de utilizar únicamente correlaciones estadísticas, ya que a menudo hay muchos factores que pueden influir en un resultado, y es necesario reducir su dimensión para identificar cuáles son los más importantes.

Inteligencia artificial (IA): Un campo de la informática dedicado a la creación de agentes inteligentes, que son sistemas que pueden razonar, aprender y actuar o realizar tareas de manera autónoma en un entorno dinámico que, cuando las hacen los humanos de manera habitual, requieren inteligencia humana. Estos agentes pueden ser máquinas físicas, software informático o una combinación de ambos. Podemos distinguir dos tipos de enfoques dentro del campo de la IA: la simbólica y la conexionista basada en redes neuronales.

Inteligencia artificial conexionista: La IA conexionista es uno de los subcampos de la IA que se inspira en el funcionamiento del cerebro humano y, por tanto, su base computacional está formada por redes neuronales digitales y el aprendizaje profundo. Estas redes están formadas por neuronas artificiales o unidades computacionales que imitan el funcionamiento de las neuronas

biológicas al trabajar en red, y cada una de las neuronas genera una señal de salida a partir de múltiples señales de entrada recibidas de otras neuronas interconectadas de la red, de modo que conjuntamente determinan el flujo de información y el comportamiento del sistema. Estos sistemas aprenden a partir de datos mediante la identificación de patrones y de relaciones complejas difíciles de determinar por métodos más tradicionales.

La IA conexionista ha obtenido resultados extraordinarios en el reconocimiento de imágenes, la visión artificial, el procesamiento del lenguaje natural y en procesos predictivos de todo tipo. Su aplicación presenta importantes desafíos en cuanto a su transparencia y la interpretación de sus modelos (explicabilidad), el posible sesgo algorítmico, el establecimiento robusto de barreras de seguridad y la ética en su desarrollo y uso de manera que sea beneficiosa para toda la sociedad.

Inteligencia artificial generativa: Es una rama de la IA que se dedica a la creación autónoma de contenidos originales, como textos, imágenes, música, vídeos e incluso código de programación. A diferencia de otras formas de IA, la IA generativa tiene la capacidad única de producir información completamente nueva y no simplemente replicar o clasificar lo existente. Esta tecnología se basa en algoritmos avanzados de aprendizaje automático, incluyendo redes neuronales profundas, modelos *Transformer*, Redes Generativas Adversariales (GANs) y Autoencoders Variacionales (VAEs).

Estos algoritmos se entrenan con grandes conjuntos de datos para identificar patrones complejos y características dentro de los datos, que luego utilizan para generar contenido novedoso y original. Algunos ejemplos destacados de IA generativa incluyen:

- *Generación de texto*: Modelos como GPT (*Generative Pre-trained Transformer*) pueden producir textos coherentes y contextuales en diversos estilos y formatos.
- *Creación de imágenes*: Herramientas como DALL-E o Midjourney pueden generar imágenes realistas o artísticas basadas en descripciones textuales.
- *Composición musical*: Algoritmos capaces de componer piezas musicales originales en diferentes estilos y géneros.
- *Síntesis de voz*: Tecnologías que pueden crear voces humanas sintéticas, casi indistinguibles de las reales.
- *Generación de vídeo*: Sistemas que pueden crear secuencias de vídeo a partir de texto o imágenes estáticas.

La IA generativa funciona aprendiendo las distribuciones estadísticas y las relaciones presentes en los datos de entrenamiento. A partir de este conocimiento, genera nuevas instancias que respetan estas distribuciones, pero que son completamente originales. Aunque el contenido generado por esta tecnología puede parecer sorprendentemente humano, es importante señalar que la IA generativa no posee comprensión real ni conciencia. Opera únicamente basándose en patrones y probabilidades aprendidas, sin entender el significado de lo que produce. Las aplicaciones de la IA generativa son vastas y están en rápida expansión. Se utiliza en la creación de contenido para marketing, entretenimiento, asistencia en tareas creativas y de diseño, entre otras áreas. No obstante, también plantea nuevos retos éticos y legales, especialmente en torno a los derechos de autor, la autenticidad del contenido y el posible uso indebido de esta tecnología.

Inteligencia artificial simbólica: Enfoque clásico de la IA que se centra en la representación y manipulación del conocimiento mediante símbolos y en la aplicación de reglas lógicas para razonar y tomar decisiones. A pesar de mostrar su capacidad en el desarrollo y aplicación de sistemas expertos, por ejemplo en la medicina para diagnosticar enfermedades, en el cribado de entradas a urgencias, y en la recomendación de tratamientos, tiene una fuerte dependencia del contexto de aprendizaje y, por tanto, tiene dificultades insalvables para escalar la dimensión y generalizar resultados. Estas limitaciones han provocado su poco uso actual si lo comparamos con el de las redes neuronales.

Inteligencia General Artificial (AGI): Hipotético nivel futuro y avanzado de IA que tendrá la capacidad de comprender, aprender y aplicar conocimientos de manera transversal a una amplia gama de tareas, de manera análoga a como lo hace la inteligencia humana. Su desarrollo y potenciales usos futuros magnificarán los retos ya identificados para la IA conexionista y, al mismo tiempo, envía una señal de alerta a los humanos para que su gran impacto transformador no se convierta en una amenaza real para la humanidad.

Internet de las cosas (IoT): Red de objetos físicos interconectados que utilizan sensores, procesadores y comunicaciones para recopilar e intercambiar datos entre ellos y con otros dispositivos y sistemas, a través de Internet.

Interpretabilidad: Capacidad de comprender y explicar el funcionamiento y las decisiones tomadas por un modelo de Machine Learning (ML) o de IA. La

interpretabilidad implica confianza en los modelos al tener la capacidad de identificar errores, corregir sesgos, mejorar el rendimiento y realizar auditorías independientes, tanto técnicas como éticas. La interpretabilidad también está íntimamente relacionada con la capacidad de explicar y comprender la operativa de los algoritmos a partir del análisis de las relaciones entre cambios en la entrada y los observados en la salida de los modelos.

Justicia algorítmica: Estudio y promoción de la igualdad y equidad en el diseño y aplicación de algoritmos, con el objetivo de evitar sesgos y discriminación a medida que la IA se utilice en más ámbitos de nuestra vida. La justicia algorítmica se fundamenta en la inclusión, la transparencia y la responsabilidad, de manera que no se perpetúe o magnifique ninguna discriminación social ni se genere ninguna inequidad.

K-means: Algoritmo de aprendizaje automático no supervisado que agrupa o clasifica los datos en un número *k* de grupos, clases o clústeres, a partir de la distancia euclidiana de cada dato a los centros de los grupos, sin necesidad de etiquetar previamente los datos. El algoritmo funciona iterativamente asignando cada dato al clúster o clase que tenga el centro más cercano (centroide) y actualizando posteriormente los centros de los clústeres o clases para minimizar la distancia total entre los puntos de todos los datos y los centros de sus respectivas clases, con el fin de crear clases muy compactas y bien separadas de las clases vecinas.

Lógica difusa: Enfoque de la lógica que permite representar y manipular la incertidumbre y la ambigüedad de cualquier proposición de manera más natural e intuitiva que la lógica clásica. En la lógica clásica, las proposiciones solo pueden ser verdaderas o falsas, mientras que en la lógica difusa las proposiciones pueden tener grados de verdad comprendidos entre el cero (0 = totalmente falso) y la unidad (1 = totalmente verdadero).

Esto se logra con los conjuntos difusos, donde la pertenencia de un elemento no es binaria (pertenece o no pertenece), sino que presenta grados de pertenencia entre 0 y 1. Por ejemplo, en un conjunto difuso de "personas altas", una persona con una altura de 1,70 metros podría tener un grado de pertenencia de 0.8, mientras que un jugador/a de baloncesto con una altura de 2.20 metros podría tener un grado de pertenencia de 1. Los conjuntos difusos también trabajan con variables lingüísticas, de manera que "persona alta" podría ser una variable lingüística que puede tener los tres valores de "baja", "media" y "alta". La lógica difusa se utiliza en situaciones de incertidumbre y ambigüedad,

cuando la información no es completa o precisa, en el reconocimiento de voz, etc., por su flexibilidad y adaptabilidad. Puedes encontrar una explicación en: https://medium.com/@javierdiazarca/lógica-difusa-ejercicios-propuestos-b99603ef1bc0.

Long Short-Term Memory (LSTM): Es un tipo de red neuronal recurrente (RNN) diseñada para abordar el problema del desvanecimiento del gradiente, el cual dificulta que las RNN aprendan dependencias temporales largas, ya que los gradientes tienden a disminuir exponencialmente a medida que la secuencia de entrada se alarga. Puedes encontrar una explicación completa de la arquitectura LSTM en: https://medium.com/analytics-vidhya/lstms-explained-a-complete-technically-accurate-conceptual-guide-with-keras-2a650327e8f2.

Lenguaje y cognición: Campo de estudio sobre la interrelación entre el lenguaje humano y los procesos cognitivos, cuyos principios se aplican para comprender, explicar y desarrollar sistemas de IA que sean capaces de procesar el lenguaje y comprenderlo.

Máquinas de Boltzmann restringidas (RBM): Son modelos de redes neuronales artificiales estocásticas que se utilizan para aprender patrones en datos no etiquetados (mediante aprendizaje no supervisado). Trabajan con una capa visible que recibe los datos de entrada y una capa oculta que aprende a representar las características de los datos. No hay conexiones entre las neuronas dentro de la misma capa, solo entre capas diferentes, lo cual las hace más eficientes para aprender patrones complejos.

Máquinas de soporte vectorial (SVM): Algoritmo de aprendizaje supervisado utilizado para la clasificación y regresión, que busca el mejor hiperplano que separa los datos en clases.

Minería de datos: Procesamiento y análisis de grandes volúmenes de datos para extraer patrones, relaciones e información útil, utilizando, entre otras, técnicas de IA.

Modelos de difusión: Son una clase de modelos probabilísticos de aprendizaje automático que aprenden a generar datos similares a un conjunto de datos de entrenamiento. Funcionan como si se añadiera ruido a los datos y luego se intentara eliminar gradualmente, de manera que características de los datos que

no son directamente observables, pero que son responsables de su variabilidad, puedan ser aprendidas en este proceso. Son útiles en áreas como el procesamiento de imágenes y el tratamiento de señales para modelar la distribución subyacente de los datos y generar nuevas muestras similares.

Modelos de lenguaje grandes, o de gran escala, o de lenguaje extensivo (LLM): Modelos de aprendizaje automático basados en redes neuronales artificiales que tienen miles de millones de parámetros y que han sido entrenados con grandes cantidades de datos de texto, lo que les permite procesar muy efectivamente el lenguaje natural, aprender patrones complejos en el lenguaje y realizar tareas como generar texto, traducir automáticamente entre muchas lenguas, resumir textos, responder preguntas, y escribir creativamente poemas, códigos, guiones, partituras musicales, cartas, etc.

Neurocognición: Estudio de los procesos cognitivos y sus bases neurológicas. En el ámbito de la IA se aplica al desarrollo de modelos de IA que emulan funciones cognitivas humanas.

Ontologías: Representación formal y estructurada del conocimiento de un dominio específico mediante entidades, relaciones y axiomas.

Operadores neuronales: Son una extensión de las redes neuronales artificiales. Tienen una arquitectura de aprendizaje profundo diseñada para aprender a transformar funciones de una manera específica. A diferencia de los sistemas tradicionales que trabajan con datos numéricos concretos, los operadores neuronales trabajan con ecuaciones, generalmente en derivadas parciales del ámbito de la física, como el modelado de la turbulencia, la tensión-deformación en materiales o el estudio del clima, que son difíciles de resolver por su complejidad. Comparten objetivo con las redes neuronales informadas por la física (PINNs) y pueden añadir flexibilidad y eficiencia en el proceso de aprendizaje. Para más información puedes consultar: https://en.m.wikipedia.org/wiki/Neural_operators.

Plan de ética en IA: Conjunto de principios y directrices que tienen como objetivo garantizar que las aplicaciones de la IA sean justas, transparentes, seguras y respetuosas con la privacidad y los derechos humanos.

Planificación automática: Proceso para encontrar una secuencia de acciones que permitan a un agente o sistema alcanzar un objetivo en un entorno dado.

Poda de redes neuronales: Técnica para reducir el tamaño y la complejidad de redes neuronales eliminando neuronas o conexiones innecesarias, con el objetivo de mejorar su eficiencia, aumentar la capacidad de generalización más allá del conjunto de datos de entrenamiento y facilitar su interpretabilidad al ser redes más sencillas.

Privacidad de los datos: Protección del derecho de los individuos a controlar la recopilación, uso y difusión de sus datos personales.

Procesamiento del lenguaje natural (NLP): Rama de la IA que trata la comprensión, la interpretación y la generación de lenguaje humano por parte de sistemas informáticos. Puedes consultar: https://medium.com/nlplanet/a-brief-timeline-of-nlp-bc45b640f07d.

Razonamiento basado en casos: Método de resolución de problemas que implica la recuperación y adaptación de casos similares anteriores para solucionar problemas nuevos.

Reconocimiento de imágenes: Capacidad de las máquinas para identificar y clasificar objetos, personas, lugares y acciones en imágenes digitales.

Reconocimiento de patrones: Capacidad para detectar e identificar estructuras, regularidades o tendencias en datos.

Reducción de la dimensión: Técnicas para reducir el número de variables de un conjunto de datos, eliminando las redundantes pero conservando la información.

Regresión: Es una técnica de aprendizaje automático supervisado que se utiliza para predecir un valor continuo de alguna variable dependiente en función de los valores de las variables independientes a partir de la información contenida en los datos de entrada de todas ellas. Existen diferentes modelos de regresión, desde los más simples de regresión lineal hasta los más complejos de soporte vectorial (SVR) a partir de SVM.

Regresión lineal: Modelo de aprendizaje supervisado que establece una relación lineal entre variables independientes y dependientes para hacer predicciones de manera continua.

Regresión logística: Modelo de aprendizaje supervisado utilizado para la clasificación binaria, que estima la probabilidad de que una observación determinada pertenezca a una clase.

Retropropagación (backpropagation): Algoritmo clave en el entrenamiento de redes neuronales artificiales, que permite la optimización iterativa de los pesos de la red. Este método de entrenamiento y su implementación algorítmica calculan los gradientes necesarios para ajustar los pesos de la red de manera eficiente, mediante la propagación hacia atrás de los errores (diferencia entre la predicción y el resultado esperado), desde la capa de salida hasta las capas anteriores. Así, la retropropagación facilita la minimización de la función de pérdida y, por tanto, acelera el proceso de aprendizaje y mejora la precisión del modelo. Este algoritmo es fundamental en el entrenamiento de redes profundas y ha sido determinante en los avances recientes en inteligencia artificial.

Robótica: Campo de la ciencia y la ingeniería que se centra en el diseño, construcción, operación y aplicación de robots y de sistemas autónomos capaces de realizar tareas en entornos diversos.

Segmentación de imágenes: Tarea de división de una imagen en regiones o segmentos basados en propiedades como color, textura o forma.

Seguridad en IA: Prácticas y medidas para proteger los sistemas de IA de las amenazas y vulnerabilidades, garantizando su integridad, confidencialidad y disponibilidad.

Sintaxis y semántica: Estudio de la estructura gramatical (sintaxis) y el significado (semántica) de las palabras y frases en el lenguaje.

Síntesis de voz: Tecnología que permite convertir texto escrito en voz hablada a través de procesos de generación de señales y modelado de la voz humana.

Sistema experto: Algoritmo de IA simbólica que utiliza el conocimiento y las reglas de un experto en un campo determinado y para una temática específica y compleja para resolverla de manera independiente y automática, una vez el algoritmo ha sido entrenado con información del experto. Un ejemplo es el sistema experto para hacer el triaje o cribado en las urgencias de un hospital de personas ingresadas con síntomas de infarto o angina de pecho. Estos sistemas son un caso de éxito de la IA simbólica para la toma de decisiones en situaciones complejas.

Sistemas de diálogo: Programas de ordenador que permiten la interacción en lenguaje natural entre usuarios humanos y máquinas.

Sistemas de razonamiento automatizado: Sistemas que utilizan técnicas de lógica y razonamiento para deducir nuevas conclusiones o verificar afirmaciones a partir de un conjunto de hechos y reglas.

Sistemas de recomendación: Algoritmos que proporcionan sugerencias personalizadas a usuarios basados en sus preferencias, historial e interacciones con otros usuarios o ítems.

Sistemas multi-agente: Conjunto de agentes inteligentes que interactúan entre sí para resolver problemas o realizar tareas que son difíciles o imposibles de realizar por un solo agente.

Técnicas de agrupamiento o clasificación: Métodos supervisados o no supervisados para dividir un conjunto de datos en grupos, clases o clústeres en función de una o más propiedades o de relaciones intrínsecas del conjunto de datos.

Test de Turing: Prueba ideada por Alan Turing para determinar si una máquina es capaz de mostrar comportamiento inteligente equivalente al de un humano.

Token: El término token tiene varios significados que dependen del contexto en el que se utilice. En el campo de la lingüística computacional y el procesamiento del lenguaje natural (PLN), un token es la unidad de texto que resulta de dividir el texto en palabras individuales, frases, símbolos y signos de puntuación, unidades compuestas de nombres propios (por ejemplo, las ciudades de New York o San Francisco), números, fechas, palabras compuestas o contracciones de palabras, y unidades semánticas complejas como nombres de personas, lugares u organizaciones.

En informática y programación, un token léxico es una secuencia de caracteres que tiene un significado según la gramática del lenguaje de programación, mientras que un token de autenticación o de transacción son dispositivos de hardware o cadenas de texto que sirven para autenticar una identidad o una transacción financiera, respectivamente. Los tokens criptográficos o activos digitales representan unidades de valor en criptomonedas o tecnología blockchain. También podríamos hablar de tokens en psicología como unidades

de recompensa por un comportamiento deseado. Tokenización: Proceso de dividir un texto en unidades más pequeñas, llamadas tokens.

Transformers: El modelo Transformer, presentado en el documento "Attention is All You Need" (accesible desde el primer enlace a continuación), ha sido la base de varios modelos de lenguaje de aprendizaje profundo como Gemini, Llama 3, Claude y ChatGPT 4. Este modelo de transducción secuencial utiliza mecanismos de atención para asignar un peso que determine la importancia de las diferentes palabras en la comprensión del contexto de una frase. Este modelo de red neuronal permite el paralelismo en la atención, lo que ha fundamentado su éxito en tareas de procesamiento de lenguaje natural. Puedes ampliar conocimiento en los siguientes enlaces:
https://arxiv.org/pdf/1706.03762v5
https://www.youtube.com/watch?v=aL-EmKuB078
https://www.youtube.com/watch?v=xi94v_jl26U

Transparencia: Apertura en el funcionamiento, los datos y los algoritmos utilizados en un sistema de IA, facilitando su comprensión y control.

Visión por computador: Campo interdisciplinario que trata de dotar a las máquinas de la capacidad de procesar, comprender e interpretar imágenes y videos del mundo real. La visión por computador 3D es una extensión que se centra en el análisis, procesamiento e interpretación de datos tridimensionales obtenidos de cámaras estereoscópicas, escáneres láser o sistemas de captura de movimiento. Permite la reconstrucción, modelado y comprensión de escenas u objetos en tres dimensiones, muy útiles en ámbitos como la robótica, la realidad aumentada, la cartografía, la medicina, la cinematografía, entre otros.

Redes Adversariales Generativas (GAN): Modelo de aprendizaje automático basado en dos redes neuronales, una generadora y una discriminadora, que aprenden de forma adversarial para generar datos nuevos realistas, como imágenes o sonidos, a partir de datos de entrada.

Redes neuronales: Modelos computacionales inspirados en la estructura y el funcionamiento del cerebro humano, formados por capas de neuronas interconectadas que permiten el aprendizaje a partir de los datos.

Redes neuronales convolucionales (CNN): Tipo de red neuronal especializada en procesar datos con estructura de rejilla, como imágenes, mediante el uso de convoluciones.

Redes neuronales de grafs (GNN): Redes neuronales diseñadas para trabajar con datos que tienen una estructura de rejilla o red que se puede representar como un grafo, donde cada nodo representa un elemento y los vínculos entre ellos representan sus relaciones. Estas redes pueden modelar relaciones complejas entre elementos de los datos y son útiles en aplicaciones como el reconocimiento de patrones en redes sociales, estructuras moleculares y otras estructuras que se puedan representar como conexiones entre elementos. Estas redes utilizan la técnica de message passing para transmitir información entre nodos adyacentes del grafo y actualizar el estado de todos los nodos (mejorar la representación de los datos).

Redes neuronales informadas por la física (PINNs): También conocidas como Redes Neuronales Entrenadas por la Teoría (TTNs), son un tipo de red neuronal que incorpora el conocimiento de leyes físicas durante el entrenamiento. Por lo tanto, no solo aprenden de datos, sino que integran conocimientos de las leyes físicas que los gobiernan. Esta información adicional permite obtener modelos precisos y robustos con pocas muestras de datos y son muy útiles para problemas en algunos campos de la biología o la ingeniería. Comparten objetivo con los operadores neuronales y aportan rigor físico y consistencia. Para más información puedes consultar:
https://en.m.wikipedia.org/wiki/Physics-informed_neural_networks

Redes neuronales recurrentes (RNN): Tipo de red neuronal que puede procesar secuencias temporales de datos, como textos, ya que tiene una estructura de bucle que permite recordar información anterior. Estas redes neuronales tienen la capacidad de utilizar la información de entradas anteriores para procesar las entradas actuales.

Redes de Petri: Modelo matemático y gráfico utilizado para describir y analizar sistemas concurrentes y distribuidos.