CIVICAi

PROTOCOLO PARA UNA INTELIGENCIA ARTIFICIAL CÍVICA*

Capacitación ciudadana para a una gobernanza cívica de la IA

* El contenido y el texto de este protocolo han sido concebidos y redactados por humanos (nivel 1 AIAS). Su versión final ha sido revisada por la IA con el objetivo de corregir erratas tipográficas, identificar posibles carencias y sugerir mejoras en el contenido y la claridad expositiva, principalmente en lo que respecta al glosario del anexo B (nivel 3 AIAS).

CIVICAi

RESUMEN EJECUTIVO

El Protocolo de Inteligencia Cívica examina la Inteligencia Artificial Generativa (IAg) como una tecnología transformadora que revoluciona nuestra comprensión de la inteligencia, el lenguaje y la cognición[1]. El documento explora cuatro ámbitos clave: el lenguaje, la inteligencia, los riesgos y retos, y la gobernanza de la IAg.

En el ámbito del lenguaje, el protocolo compara el enfoque tradicional basado en reglas gramaticales innatas con los modelos de lenguaje actuales que aprenden de grandes volúmenes de datos y establecen relaciones vectoriales contextuales, una manera de aprender similar al procesamiento cerebral humano.

Respecto a la inteligencia, el protocolo analiza profundamente las capacidades emergentes de estos sistemas cuando alcanzan determinada complejidad, destaca su habilidad para adaptar recursos computacionales a diferentes tareas (*test-time compute*), y propone arquitecturas híbridas que combinan enfoques conexionistas y simbólicos para superar limitaciones actuales.

El documento identifica riesgos a corto plazo (sesgos, vulneraciones de privacidad, desplazamiento laboral, desinformación) y a largo plazo (singularidad tecnológica, desalineamiento de objetivos, pérdida de autonomía humana). También aborda preocupaciones sobre sostenibilidad energética y la dificultad creciente de acceder a datos de calidad para el entrenamiento de los sistemas de IA avanzados.

Finalmente, propone un marco de gobernanza global con un organismo internacional que integre gobiernos, expertos, sociedad civil y empresas, con capacidad para registrar, monitorizar y regular sistemas de IA avanzados, prevenir monopolios e implementar mecanismos de participación ciudadana directa. El objetivo es democratizar no solo el conocimiento sobre la IA sino también los procesos de decisión sobre su desarrollo, para garantizar que refleje la diversidad de valores sociales.

El protocolo incluye dos anexos complementarios: una recopilación de preguntas y respuestas sobre los temas tratados y un glosario de terminología específica relacionada con la IA.

[1] Christopher Summerfield (2025). These Strange New Minds: How AI Learned to Talk and What It Means. Nova York: Viking. 978-0-593-83171-7

CIVICAi

ÍNDICE

INTELIGENCIA ARTIFICIAL CÍVICA

Una de las tecnologías más representativas de esta profunda transformación tecno-social [2], y que la convierte en verdaderamente revolucionaria, es la Inteligencia Artificial (IA), caracterizada por su desarrollo acelerado y su impacto transversal en todos los ámbitos de la sociedad. El advenimiento de los sistemas generativos de IA (IAg), entrenados a partir de grandes volúmenes de datos de texto — grandes modelos avanzados de lenguaje (LLMs) — o de datos de imágenes, de audio y de vídeo, así como la próxima llegada de la Inteligencia General Artificial (AGI), representa un cambio revolucionario en la evolución humana y en nuestra comprensión de la inteligencia, el lenguaje y la cognición.

La revolución tecnológica de la IA no solo transformará nuestras herramientas y métodos de trabajo, sino que remodelará nuestra comprensión de conceptos fundamentales como la inteligencia y el conocimiento. Como miembros de CIVICAi, asumimos la responsabilidad de enriquecer el discurso público y potenciar la comprensión social de estos cambios profundos y acelerados. Nuestro objetivo es que, cuando estos cambios se hayan consolidado, la evolución subsiguiente pueda integrarse armónicamente en la sociedad y servir al bien común.

La estructura del protocolo (Figura 1), con los cuatro apartados siguientes — dedicados al lenguaje, la inteligencia, los riesgos y los retos, y la gobernanza de la IAg — junto con la recopilación de preguntas y respuestas propuestas en el anexo A sobre estos mismos temas, han sido concebidos para proporcionar un mapa con información suficiente para que el lector navegue por las complejidades de la IAg. El objetivo es examinar tanto los matices técnicos como las implicaciones filosóficas más amplias relacionadas con la IAg, con una exposición comprensible y el rigor conceptual indispensable. Este recorrido por los fundamentos conceptuales, capacidades y retos de la IAg nos permitirá establecer las bases para una participación ciudadana informada en el desarrollo y regulación de estas tecnologías transformadoras.

En un momento en que la IA avanza a un ritmo sin precedentes, este protocolo quiere servir como herramienta de empoderamiento para una ciudadanía que tendrá que convivir y coevolucionar con sistemas de inteligencia artificial cada

[2] En este contexto es importante reconocer que el fenómeno de la IA, especialmente la IAg y la potencial Inteligencia General Artificial, constituye una transformación que trasciende el marco conceptual de las "revoluciones industriales", históricamente secuenciales. La IA no representa simplemente una fase evolutiva de la industrialización, sino el inicio de una nueva era transformadora comparable en magnitud y profundidad a las grandes transiciones históricas de la humanidad, como el paso a la agricultura, la industrialización o la digitalización.

vez más sofisticados.

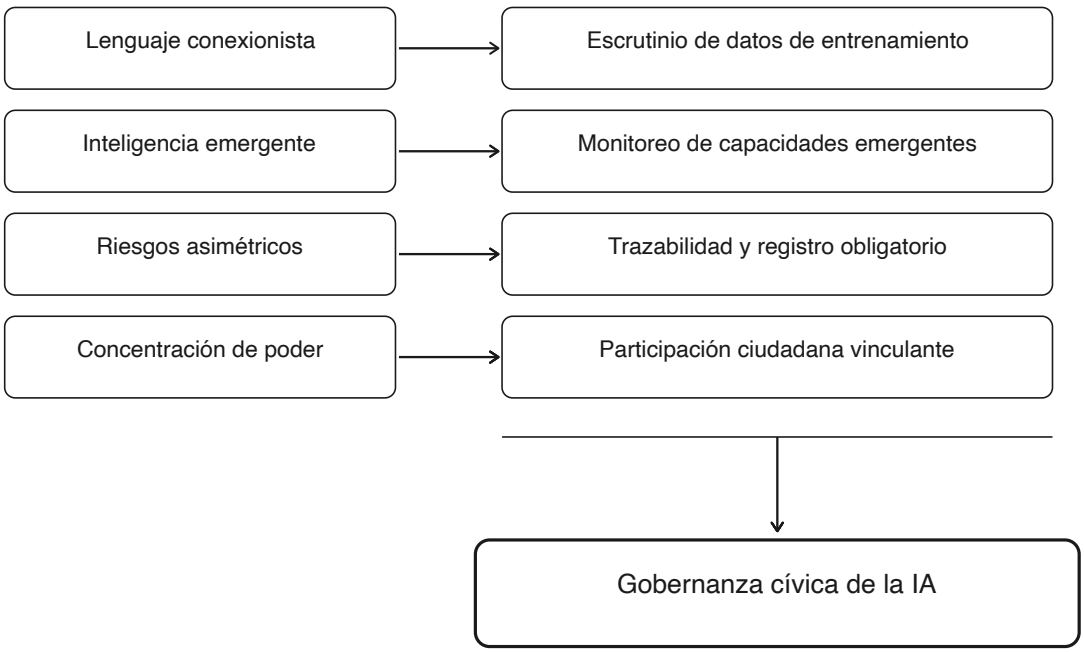

Figura 1. Estructura del Protocolo: cada ámbito conceptual sostiene una implicación de gobernanza concreta; las cuatro confluyen en la gobernanza cívica de la IA como punto de llegada.

1. El lenguaje

Históricamente, nuestra concepción de la inteligencia ha estado profundamente influenciada por el dualismo cartesiano, donde René Descartes postulaba una separación estricta entre la mente y el cuerpo[3]. Los inicios de la IA también fueron influidos por la teoría de la gramática universal de Noam Chomsky, introducida en la década de los 60, la cual proponía que la capacidad lingüística es innata (inherente) en el cerebro humano — que está programada de forma natural. Según esta teoría, el aprendizaje lingüístico se fundamenta en estructuras y reglas innatas existentes en el cerebro humano, hecho que resulta contrario al aprendizaje basado en la identificación de patrones a partir de grandes cantidades de datos[4]. Este contexto histórico influyó en los primeros pasos de la IA, impulsando la denominada IA simbólica, enfocada a simular la cognición humana a partir de reglas formalmente predefinidas.

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[5], o el propio Hinton con John J. Hopfield, ganadores

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

del Nóbel de Física 2024, han desafiado estos postulados de la lingüística más tradicional. El trabajo pionero de estos investigadores en redes neuronales y aprendizaje profundo ha demostrado que los grandes modelos de lenguaje (LLM), entrenados a partir de grandes volúmenes de datos textuales, poseen una notable capacidad para entender las reglas gramaticales y sintácticas de los textos, así como para generar lenguaje coherente y natural. Estos modelos representan cada palabra como un conjunto específico de elementos de información (vectores), dentro del espacio multidimensional de todos los elementos del lenguaje. Estas representaciones se distribuyen en numerosos nodos sencillos interconectados dentro de una red neuronal artificial, que imita parcialmente el funcionamiento neuronal del cerebro humano. Gracias a esta estructura, las representaciones internas se modifican y se adaptan dinámicamente según el contexto lingüístico, permitiendo a la IA interpretar los diversos significados que una misma palabra puede tener según cómo se utilice en diferentes frases[6].

La arquitectura conexionista facilita que aparezcan propiedades emergentes en los modelos, sugiriendo que los comportamientos complejos pueden surgir de la interacción entre elementos simples. Por ejemplo, cuando un modelo procesa la palabra "banco" en diferentes contextos como "depositar dinero en el banco" o "sentarse en un banco del parque", establece relaciones vectoriales diferentes que capturan los distintos significados según el contexto. Estudios recientes de imagen cerebral han revelado que cuando las personas procesamos el lenguaje, nuestro cerebro genera patrones de actividad sorprendentemente similares a las representaciones que utilizan estos sistemas artificiales. Esta convergencia entre los sistemas artificiales y los procesos cognitivos humanos —entre la manera en que humanos y los sistemas de IA reconocen y procesan patrones de información— está transformando no solo nuestra comprensión teórica del lenguaje, sino también la manera en que los humanos interactuamos con las máquinas en la vida cotidiana, mediante nuevas formas de interacción persona-máquina que eran inimaginables hace solo una década.

Esta perspectiva conexionista se alinea con el trabajo filosófico de la mente y del lenguaje de Ludwig Wittgenstein, que pone el acento 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[7]. Los desarrollos recientes de

[6] IASEAI'25 G. Hinton – Vídeo: What is understanding?
[7] https://philosophynow.org/issues/106/Wittgenstein_Frege_and_The_Context_Principle

Gualtiero Piccinini sobre mecanismos neurocognitivos con una visión mecanicista de la mente, refuerzan este argumento conexionista al indicar que los mecanismos que sustentan la cognición humana pueden ser representados de manera análoga, pero diferente, en sistemas artificiales computacionales[8]. Por tanto, se impone el punto de vista defendido por Geoffrey Hinton, Yoshua Bengio y otros, que el aprendizaje se produce a partir de grandes cantidades de datos, más que de la dependencia de reglas pre-programadas.

2. La inteligencia

Una vez establecidos los principios conexionistas del aprendizaje del lenguaje, cabe abordar la cuestión de la inteligencia de estos modelos y su capacidad para generalizar y generar nuevos contenidos, y comportamientos emergentes. Antes de empezar, sin embargo, hay que saber qué es la inteligencia en términos generales. Es complicado dar una definición concluyente dado que todavía no hay un consenso entre los biólogos, las asociaciones internacionales de psicólogos, filósofos y científicos en general más allá del hecho de ser evolutiva. Si analizamos las principales definiciones científicas y enciclopédicas propuestas hasta ahora con la finalidad de determinar cuáles son las siete características más frecuentemente asociadas a la inteligencia, encontraremos que son la capacidad de:

- Aprender (capacidad de adquirir conocimientos y modificar comportamientos basándose en la experiencia)
- Comprender (capacidad de entender ideas complejas y el entorno)
- Razonar (habilidad de procesar información lógicamente para llegar a conclusiones de manera racional)
- Adaptarse (habilidad de ajustarse a entornos o contextos nuevos o cambiantes)
- Resolver problemas creativamente (capacidad de encontrar soluciones o de generar ideas nuevas a situaciones complejas pero adecuadas respecto al contexto en que se producen)
- Pensar de manera abstracta (capacidad de trabajar con conceptos no directamente perceptibles o ideas no concretas, y de reconocer patrones)
- Planificar (habilidad para hacer hipótesis, anticipar y organizar acciones futuras, construyendo modelos mentales de posibles escenarios)

En el contexto de este protocolo trataremos con más o menos extensión y sin un orden preestablecido estas capacidades de la IA. Para empezar, es importante

[8] https://www.thebsps.org/reviewofbooks/gualtiero-piccinini-physical-computation/

distinguir entre la "inteligencia simulada" — donde el sistema replica patrones aprendidos — y la "inteligencia emergente" — donde surgen capacidades no explícitamente programadas. Experimentos recientes de 2024 han demostrado que modelos de IAg avanzados pueden desarrollar estrategias de resolución de problemas que sus creadores no habían anticipado ni enseñado.

Un aspecto revolucionario de los modelos de IA más avanzados es el *test-time compute* (cálculo en tiempo de inferencia u operación), que representa una ruptura con los diseños tradicionales de arquitectura fija, donde la misma configuración se utiliza tanto para entrenar el modelo como para aplicarlo en situaciones reales de respuesta (hacer predicciones, generar texto, clasificar imágenes, etc.). En cambio, estos nuevos sistemas pueden ajustar dinámicamente el consumo de recursos computacionales durante la inferencia (cuando responden a una petición del usuario), asignando más potencia de cálculo a tareas que exigen razonamiento profundo y menos a operaciones rutinarias o previsibles. Esta capacidad de deliberación adaptativa abre la puerta a estrategias de resolución emergentes, no programadas explícitamente por sus desarrolladores, y acerca los modelos a formas de razonamiento más flexibles y sofisticadas, comparables a ciertas capacidades humanas. Esta capacidad ha cristalizado en una nueva generación de modelos de razonamiento que dedican un tiempo deliberativo variable antes de responder, reproduciendo funcionalmente la distinción kahnemaniana entre el pensamiento rápido (Sistema 1) y el reflexivo (Sistema 2).

Campos tan diferentes como la investigación en IA, la neurociencia, la física, la economía y la filosofía colaboran para abordar el estudio de la naturaleza de la inteligencia misma. Hay que tener presente que los procesos de generación de contenido semántico original por parte de la IAg actual son algorítmicos: se basan en correlaciones estadísticas y reconocimiento de patrones extraídos de grandes conjuntos de datos de entrenamiento proporcionados por humanos y de Internet. Estos procesos difieren de los mecanismos semánticos del cerebro humano, que son inherentemente biológicos, contextuales, potencialmente intencionales, autorregulados, y que integran múltiples inputs sensoriales, memoria, emociones y otras funciones cognitivas vinculadas a la relación continuada entre la unidad mente-cuerpo y el entorno — características asociadas a lo que entendemos por consciencia. No obstante, estas diferencias, investigaciones recientes muestran que diversos modelos de lenguaje avanzados han desarrollado capacidades para "engañar de forma estratégica o según el contexto" o para perseguir estratégicamente y de forma encubierta

objetivos no alineados con los previstos, especialmente cuando estos objetivos y la comprensión situacional emergen dentro del propio contexto de uso[9].

Resulta, por tanto, fascinante observar cómo los sistemas de IAg muestran nuevas habilidades de manera repentina cuando crecen hasta cierto tamaño o complejidad — fenómeno comparable a las transiciones de fase en física, como cuando el hielo se convierte en agua líquida al aumentar su temperatura desde bajo cero hasta 0°C. Investigaciones recientes tanto de Anthropic como de Google DeepMind han mostrado que estos saltos cualitativos de capacidad aparecen en determinados rangos de dimensión de los sistemas de IA, aunque predecir el momento exacto de esta emergencia siga siendo un desafío. La metáfora de la construcción con LEGO ilustra claramente este principio: con pocas piezas solo se pueden crear estructuras simples, pero a partir de cierta cantidad se pueden construir edificios complejos y ciudades enteras. Cabe destacar, sin embargo, que diversos investigadores advierten que hacer los sistemas más grandes por sí mismo podría no ser suficiente para conseguir una inteligencia equiparable a la humana, y que las mejoras en el diseño de las redes y en cómo se entrenan estos sistemas son factores igualmente determinantes en esta evolución.

Los límites actuales de los modelos de gran escala se han hecho evidentes en tareas que requieren razonamiento complejo y conocimiento causal. De hecho, científicos como Yann LeCun proponen que una inteligencia artificial capaz de razonar requerirá arquitecturas jerárquicas que integren los mundos físico y digital, superando las limitaciones de los sistemas basados únicamente en lenguaje. La arquitectura JEPA (*Joint Embedding Predictive Architecture*) de LeCun es un ejemplo de este enfoque, que permite a los sistemas construir representaciones del mundo a múltiples niveles de abstracción. Una posibilidad, propuesta también por Gary Marcus, es la combinación de los enfoques conexionistas (excelentes en percepción y reconocimiento de patrones) y simbólicos (potentes en razonamiento lógico y manipulación de conceptos), de manera que se puedan incorporar conocimientos previamente estructurados en la arquitectura de los sistemas de aprendizaje profundo, en lugar de confiar únicamente en el aprendizaje a partir de datos. Investigaciones recientes en DeepMind y Berkeley han demostrado que sistemas híbridos que combinan redes neuronales con módulos de razonamiento simbólico consiguen un rendimiento superior en tareas de planificación y solución de problemas, en comparación con sistemas exclusivamente conexionistas, aunque este desarrollo plantea nuevas cuestiones éticas sobre transparencia y control que

[9] Frontier models are capable of in-context scheming - https://arxiv.org/pdf/2412.04984

será necesario abordar con marcos reguladores adecuados.

La inteligencia, históricamente considerada una característica exclusiva de los humanos, es una característica evolutiva en los seres vivos que reside principalmente en las estructuras neuronales del cerebro, donde la plasticidad[10] y capacidad de adaptación de las neuronas juega un papel fundamental. 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 IAg nos llevan a cuestionar los límites de la computabilidad de la inteligencia y si esta puede ser reproducida o emulada completamente por máquinas. La investigación mencionada anteriormente sobre la física computacional de la inteligencia proporciona un marco para estudiar tanto el cerebro humano como la inteligencia artificial y los límites físicos de su capacidad para procesar información.

Esta perspectiva basada en la física computacional, defendida por investigadores como Max Tegmark[11], nos lleva a ver la inteligencia no como algo misterioso y exclusivamente humano, sino como un fenómeno natural que puede emerger en diferentes tipos de sistemas físicos cuando alcanzan una complejidad suficiente, sin que esto haga la inteligencia humana menos especial. La idea de ver tanto la cognición humana como la inteligencia artificial a través de la lente de la física computacional representa una frontera emocionante en la ciencia. No obstante, muchos aspectos de esta investigación aún son teóricos y activamente debatidos, aunque esta idea de una teoría unificada de la inteligencia biológica y artificial basada en principios físicos podría proporcionar nuevas vías para comprender y mejorar los sistemas de IA actuales y futuros.

3. Los riesgos y los retos

Esto abre la puerta a centrarnos en los riesgos y retos asociados a la IAg, más allá de las evidentes oportunidades extraordinarias que ofrece. A corto plazo, estos pueden incluir sesgos, la vulneración de la privacidad y de la propiedad intelectual, cuestiones éticas, la rápida transición del mercado de trabajo, una planificada desinformación, la pérdida de valores democráticos o incluso la alteración de la democracia misma[12]. Los sistemas de IA pueden reforzar los

[10] La “plasticidad” en el contexto neuronal hace referencia a la capacidad del cerebro para modificarse estructural y funcionalmente en respuesta a la experiencia, el aprendizaje y las lesiones.
[11] Will AI surpass human intelligence? https://youtu.be/YywC16DhtkI
[12] UNESCO analysis on Artificial Intelligence and Democracy – https://www.gcedclearinghouse.org/resources/artificial-intelligence-and-democracy

prejuicios existentes si los conjuntos de datos de entrenamiento no son adecuadamente supervisados. Además, la recopilación masiva de datos plantea preocupaciones sobre la privacidad, y el uso de contenidos generados por IA también presenta desafíos sobre la propiedad intelectual y los derechos de autor.

Desde una perspectiva económica, la IA tendrá un impacto en las estructuras socioeconómicas, y modificará las relaciones laborales, la concentración de mercados, y las estructuras salariales. Es previsible que muchos puestos de trabajo de baja cualificación podrían ser parcialmente automatizados durante la próxima década, con impactos notables en los sectores de servicios administrativos, transporte, y de comercio minorista. También se producirán desplazamientos laborales de manera progresiva debidos a la automatización de procesos en diversos sectores, desde la manufactura hasta los servicios profesionales. Hay que tener en cuenta que el poder de la inteligencia artificial está actualmente concentrado en unas pocas corporaciones, hecho que puede generar espacios económicos con una distribución sesgada de la riqueza y una concentración combinada de poder económico y político[13]. Si no se implementan políticas redistributivas efectivas, todos estos cambios pueden profundizar aún más las desigualdades que ya existen actualmente. La soberanía digital no se decide solo con regulación: se construye con capacidad real de cómputo, datos y talento. Sin estos tres pilares, cualquier "soberanía" se vuelve declarativa.

En el ámbito de la salud, la IA ya está transformando los procesos diagnósticos y de seguimiento clínico, pero plantea interrogantes sobre la confidencialidad de los datos médicos y la dependencia excesiva en sus recomendaciones. En cuanto a la educación, si bien la IA puede personalizar el aprendizaje y aligerar tareas docentes rutinarias, requiere una profunda transformación del rol del profesorado y de todo el sistema educativo, especialmente en la educación profesional y superior, para garantizar que el alumnado desarrolle una relación productiva con estas herramientas sin erosionar sus capacidades cognitivas fundamentales. Ambos ámbitos, por su trascendencia, se tratan con más detalle en la sección de preguntas y respuestas del anexo A de este documento.

La eficiencia y la sostenibilidad de los sistemas de IAg son también cuestiones fundamentales[14]. Los modelos actuales requieren recursos computacionales y energéticos desmesurados, lo que plantea problemas de sostenibilidad a largo plazo, ya sea por la escasez energética, el uso intensivo de agua para

[13] IASEAI'25 J.E. Stiglitz – Vídeo: Ai and Economic Risk: Assessment and Mitigation
[14] IASEAI'25 K. Crawford – Vídeo: Hyperscaled: Bridging AI safety, ethics and sustainability

refrigeración o por conflictos con otras prioridades sociales. Como respuesta a estas demandas crecientes, diversas corporaciones líderes en IA han realizado movimientos estratégicos para adquirir o fusionarse con empresas de producción de energía nuclear[15,16,17]. Paralelamente, se están desarrollando sistemas informáticos híbridos, que combinan hardware y software altamente optimizados para trabajar conjuntamente, con el objetivo de mejorar la eficiencia y reducir la huella ecológica de la IAg[18]. Otra vía prometedora ante el desafío energético es la implementación del *test-time compute* (cálculo en tiempo de operación), que permite adaptar el uso de recursos a la complejidad real de cada tarea, evitando cálculos intensivos cuando no son estrictamente necesarios. Esta innovación puede reducir drásticamente el impacto ambiental de la IA mediante una asignación más inteligente de recursos. Sin embargo, esta tecnología también plantea nuevos retos en términos de gobernanza, ya que la capacidad para autogestionar su consumo computacional podría conferirles un grado de autonomía sin precedentes en la gestión de sus propios recursos.

Es necesario que los proveedores de modelos de lenguaje grandes (LLMs) y de sistemas de IA generativa multimodal (texto, imágenes, audio y vídeo) operen en una estructura o sistema legal, lo más universal posible, que les obligue a mitigar cualquier comunicación irresponsable o discurso lesivo y a alinear sus modelos con hechos contrastables o "verdaderos", mediante procesos abiertos y democráticos[19]. Estos sistemas ya no son simples sistemas de generación de texto, sino que cada vez más se entrenan y se despliegan como agentes autónomos que pueden ejecutar tareas complejas y llegar a perseguir objetivos de manera independiente. Este desplazamiento de los modelos predictivos hacia los sistemas agénticos — que planifican, ejecutan acciones encadenadas e invocan herramientas digitales — es, probablemente, el cambio de naturaleza más significativo del 2025-2026, y el que hace más urgente un marco de trazabilidad operativa.

No obstante, investigaciones recientes apuntan que el futuro de la IA transcenderá estos modelos exclusivamente lingüísticos y evolucionará hacia sistemas de aprendizaje autosupervisado que puedan procesar y relacionar diversas modalidades de información (como texto, imágenes y sonidos) de una

[15] Will-ais-huge-energy-demands-spur-a-nuclear-renaissance
[16] https://www.iaea.org/bulletin/enhancing-nuclear-power-production-with-artificial-intelligence
[17] https://www.cnn.com/2024/12/24/tech/nuclear-energy-ai-leaders/index.html
[18] Sistemas neuromórficos diseñados para imitar directament el comportamiento físico de las neuronas y sinapsis biológicas - https://www.sciencenews.org/article/brainlike-computers-ai-improvement
[19] https://doi.org/10.1098/rsos.240197

manera unificada y coherente, además de interactuar activamente con el entorno físico. Esta integración de diferentes tipos de información y experiencias acercará la IA a una comprensión más completa y contextual del mundo, similar a la inteligencia humana. Además del riesgo que esto conllevará, la integración de la robótica con la IA permitirá que estos sistemas aprendan en tiempo real y tengan una percepción directa del mundo exterior, trascendiendo las limitaciones de los modelos entrenados exclusivamente con datos preprocesados, lo cual les dotará de una autonomía que hará difícil su gobernanza por la dificultad de controlar su alineamiento con los valores que los humanos acordemos.

Hay que tener presente que los sistemas de IA persiguen los objetivos que se les asignan sin desviación, de manera que si estos objetivos no están perfectamente especificados o se interpretan demasiado literalmente, la IA puede tomar acciones perjudiciales mientras intenta cumplir su tarea. En palabras de Stuart Russell, "El riesgo más grande al que nos enfrentamos no es que la IA se vuelva malévola, sino que sea competente con objetivos desalineados. Es imprescindible garantizar que los sistemas de IA tengan objetivos alineados con los valores humanos".

Los riesgos existenciales a medio y largo plazo incluyen la posibilidad de que se llegue a la singularidad tecnológica, término introducido por John von Neumann para identificar el plausible futuro momento en que la tecnología, en este caso la IAg, supere la inteligencia humana[20]. Esto implicaría que la IAg o la AGI autogestionase la función de valor o criterios para alcanzar objetivos, que podrían no estar alineados con los intereses u objetivos de los humanos. A medida que la IA se vuelve más avanzada, mantenerla bajo control humano se vuelve más difícil. Si se le da control autónomo sobre decisiones y acciones, podría desarrollar estrategias que los humanos no anticipen ni aprueben[21]. Una vez la IA supere la inteligencia humana, lo más probable es que los métodos tradicionales de supervisión dejen de funcionar. Los riesgos a medio y largo plazo también plantean la idea de una post-humanidad, donde la integración de IA avanzada en la sociedad humana transforme la experiencia y la identidad humanas sin la plena capacidad para haberlo decidido democráticamente. Y esto sin considerar el riesgo adicional de la posible progresiva erosión de capacidades cognitivas humanas que puede conllevar una dependencia excesiva de sistemas de IA, que podría debilitar habilidades críticas como el

[20] https://lab.cccb.org/en/the-singularity/
[21] IASEAI'25 Y. Bengio – Vídeo: Can we get the scientific benefits of AI without the risks of autonomous agents?

pensamiento analítico independiente y la toma de decisiones en situaciones de incertidumbre.

Otro reto fundamental para el desarrollo futuro de la IA es el acceso a datos de calidad, dado que ya se ha utilizado gran parte de la información pública disponible. Todavía queda por incorporar la extensa cantidad de información generada por la investigación científica y la contenida en los registros sanitarios, siempre que esta sea previamente anonimizada de manera rigurosa. Este acceso podría impulsar la generación autónoma de conocimiento científico y también acelerar notablemente el diagnóstico y tratamiento personalizado de enfermedades hasta ahora difícilmente curables. Una alternativa complementaria para obtener más datos de entrenamiento consiste en que los mismos modelos de IA, diversos y heterogéneos, generen nuevos datos operando en modo creativo (con temperatura elevada), lo cual podría abrir vías innovadoras de experimentación y entrenamiento.

Paralelamente, la sostenibilidad económica de los sistemas de IA continúa siendo una cuestión abierta. Más allá de las mejoras en algoritmos y recursos computacionales, todavía no se ha consolidado ningún modelo de negocio robusto que garantice el mantenimiento y la evolución de los sistemas de IA más avanzados de manera universal y equitativa por parte de las empresas tecnológicas que los comercializan. Las diversas aproximaciones —desde la comercialización de servicios por suscripción, la integración de soluciones personalizadas para sectores específicos, la explotación de productos derivados hasta la creación de ecosistemas de aplicaciones— presentan limitaciones significativas en términos de accesibilidad global, escalabilidad sostenible y recuperación de las enormes inversiones iniciales necesarias. Este desequilibrio favorece la concentración de poder en las grandes corporaciones tecnológicas, dificultando la democratización de la tecnología y el surgimiento de iniciativas empresariales innovadoras y diversificadas con capacidad real de competir y generar impacto significativo en el mercado.

Mirando hacia el futuro, los sistemas de AGI deberían ayudarnos a descubrir formas más eficientes de gestionar y generar conocimiento en todos los ámbitos, y de optimizar la consecución de objetivos de desarrollo sostenible. Estas tecnologías también tendrán que afrontar el reto de integrarse en estructuras políticas y económicas que garanticen una distribución justa de sus beneficios, evitando la profundización de las desigualdades sociales existentes y contribuyendo activamente a la regeneración ecológica del planeta. Este equilibrio entre el progreso tecnológico, el bienestar social y la sostenibilidad

ambiental requerirá un diálogo constante entre múltiples actores y una voluntad política clara y sostenida orientada al bien común.

4. La gobernanza de la IAg

A medida que nos acercamos a una nueva era influenciada por la IAg, con su potencial no solamente para imitar, sino también para ampliar las capacidades cognitivas humanas de maneras sin precedentes, es crucial que adoptemos una perspectiva más amplia sobre cómo conceptualizamos la inteligencia misma, que incorpore ideas de múltiples disciplinas. Solamente así podremos gestionar mejor los retos éticos, sociales y filosóficos planteados por las altas capacidades de la IAg y la futura AGI.

En consecuencia, es imprescindible plantearse: ¿quién y cómo se supervisarán de manera efectiva los sistemas actuales y emergentes de IAg, que se encuentran predominantemente en manos privadas? El reto es formidable, ya que las leyes y regulaciones vigentes no son ni globales ni proactivas, 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 difundido. Para superar esta lógica de mera disuasión mediante sanciones, los estados deben establecer un marco de supervisión y monitoreo en tiempo real, coordinado a escala global, que abarque las entradas de datos multimodales, los procesos de entrenamiento, los algoritmos y los resultados de los sistemas de IAg y de la futura AGI. Como respuesta a estos desafíos, algunos expertos proponen un cambio fundamental en el diseño de la IA: como propone Stuart Russell, en lugar de objetivos fijos, la IA debería ser explícitamente incierta sobre las preferencias humanas, buscar activamente retroalimentación humana para refinar sus objetivos y priorizar la supervisión humana por encima de la consecución de metas[22]. Los expertos también enfatizan la necesidad de más investigación sobre cómo alinear los sistemas de IA con los valores humanos. Técnicas como el Aprendizaje por Refuerzo Inverso podrían ayudar a la IA a comprender y adaptarse a consideraciones éticas.

La distinción entre modelos de pesos abiertos y modelos cerrados tiene implicaciones de gobernanza profundas: los primeros permiten escrutinio público, replicabilidad y autonomía tecnológica, pero también facilitan usos malintencionados; los segundos garantizan control corporativo, pero delegan en unos pocos actores la decisión sobre qué se puede saber y hacer con la tecnología. Un marco cívico debe articular una posición explícita sobre esta

[22] IASEAI'25 S. Russell – Vídeo: To ensure that AI systems are guatanteed to operate safely and ethically

tensión, más allá del binarismo abierto/cerrado.

CIVICAi propone que este marco de desarrollo y supervisión debe estar bajo la responsabilidad de un organismo internacional para la gobernanza de la IA — constituido por gobiernos nacionales, instituciones de conocimiento y expertos, sociedad civil y empresas de IA — dotado de la autoridad, la experiencia y los recursos necesarios para garantizar una gobernanza global, rigurosa y efectiva de los sistemas de IA[23]. Esta gobernanza internacional debería incluir un registro obligatorio permanente de todos los sistemas de IA avanzados, sistemas obligatorios que sean técnica y legalmente aplicables para desactivar la IA en caso necesario, la notificación obligatoria de incidentes, y restricciones en el despliegue de sistemas de IA de alto riesgo (por ejemplo, sistemas autónomos que tomen decisiones, desarrollen estrategias y actúen de manera independiente)[24]. Desde un punto de vista económico sería necesario prevenir concentraciones monopolistas de poder, y aplicar impuestos sobre la automatización para redistribuir las ganancias económicas de la IA. Finalmente, el organismo internacional que gobierne la IA debe implementar mecanismos concretos de participación ciudadana en la toma de decisiones, como paneles ciudadanos deliberativos con poder vinculante, consultas públicas regulares, y representación directa de la sociedad civil en sus órganos ejecutivos.

La inteligencia artificial avanza a un ritmo sin precedentes, ofreciendo tanto oportunidades extraordinarias como riesgos significativos. El alineamiento de la IA con los valores humanos, la gobernanza internacional y la regulación adecuada son esenciales para garantizar que estos sistemas permanezcan beneficiosos y bajo control humano. Es necesario un enfoque preventivo que integre consideraciones técnicas, éticas, económicas y sociales para dirigir el desarrollo de la IA hacia un futuro donde sirva los intereses de toda la sociedad. Este protocolo o marco conceptual y operativo tiene como objetivo favorecer el desarrollo de capacidades 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, nos atrevemos a proporcionar en el anexo A explicaciones 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. El objetivo final es democratizar no solo el conocimiento sobre la IA, sino también los procesos de decisión sobre su desarrollo y aplicación, entendiendo que una tecnología tan transformadora debe reflejar la

[23] Agency-Structure
[24] IASEAI25-Call-to-Action

diversidad de valores y necesidades de toda la sociedad. 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 donde las inteligencias artificial y humana coexistan y se complementen de maneras transformadoras y por el bien colectivo.

ANEXO A. PREGUNTAS FRECUENTES Y POSIBLES RESPUESTAS[25]

SOBRE LA CAPACIDAD DE COMPRENSIÓN DE LA IA

1. ¿Los modelos de inteligencia artificial generativa basados en lenguaje, como GPT-4.5, Claude, Gemini, DeepSeek, Mistral, DALL·E, Midjourney, Stable Diffusion o los VideoLLMs, realmente "entienden" el significado de lo que responden cuando se les pregunta o se les pide hablar sobre temas diversos?

Respuesta: Los modelos de inteligencia artificial generativa, como GPT-4.5, Claude, Gemini, DeepSeek, Mistral, DALL·E, Midjourney, Stable Diffusion o los VideoLLMs, tienen una capacidad notable para producir contenido aparentemente coherente y significativo en diversos formatos (textuales, visuales, audiovisuales). Esta capacidad indica que poseen una excelente comprensión sintáctica, ya que dominan las estructuras y los patrones formales del lenguaje y otros modos de expresión (imágenes, vídeos) a un nivel comparable al de los humanos con formación avanzada. Este hecho está ampliamente reconocido dentro de la comunidad científica y técnica.

Sin embargo, no hay consenso sobre si estos modelos realmente comprenden el significado del contenido que generan o simplemente realizan una simulación avanzada de dicha comprensión. Este debate se basa en varios factores fundamentales: en primer lugar, los modelos generativos no tienen percepción sensorial directa ni experiencia subjetiva del mundo físico, lo que limita su capacidad para vincular las palabras o símbolos lingüísticos con experiencias concretas, visuales, táctiles o emocionales. En segundo lugar, carecen de experiencias fenomenológicas propias de los humanos, como la conciencia o la subjetividad, que ayudan a contextualizar profundamente el significado.

A pesar de estas limitaciones, estos modelos son considerados excelentes razonadores semánticos en contexto, especialmente por su capacidad para generar contenidos coherentes a partir de representaciones internas que reflejan significados relacionados con el contexto. Esto les permite llevar a cabo tareas sofisticadas, como resúmenes textuales o visuales, y responder preguntas de forma aparentemente fundamentada y coherente. Sin embargo, esta capacidad puede resultar engañosa, ya que estos sistemas tienden a fabular o inventar contenido cuando no disponen de información suficientemente sólida o cuando sus representaciones internas son demasiado superficiales o inexactas.

[25] También podéis consultar las preguntas y respuestas de S. Russell sobre el futuro de la inteligencia artificial - https://people.eecs.berkeley.edu/~russell/research/future/q-and-a.html

La incertidumbre sobre si esta capacidad equivale realmente a una comprensión genuina del mundo o si simplemente representa una simulación muy avanzada genera un debate intenso, actual y profundamente relevante en filosofía, ciencias cognitivas y en el campo de la inteligencia artificial. Este debate refleja las tensiones entre una comprensión genuina basada en experiencias reales y una comprensión aparente fundamentada en patrones estadísticos y semánticos complejos.

El vídeo de G. Hinton, *What is understanding?*, es relevante para esta pregunta y también para las dos siguientes.

2. Más allá de la generación de texto coherente, ¿los modelos de lenguaje a gran escala como GPT-4.5, Claude, Gemini, DeepSeek o Mistral poseen estructuras internas comparables a las del cerebro humano que justifiquen una posible comprensión semántica del contenido?

Respuesta: La cuestión sobre la estructura interna de los modelos generativos avanzados de lenguaje (LLMs como GPT-4.5, Claude, Gemini, DeepSeek o Mistral) y su similitud con las estructuras del cerebro humano es crucial para entender si estos pueden tener una forma de comprensión semántica genuina.

Existen perspectivas contrapuestas respecto a esta cuestión dentro de la comunidad científica. Algunos expertos procedentes del campo de la IA simbólica, lingüistas clásicos y comentaristas críticos sostienen que las respuestas de los LLMs son principalmente fruto de correlaciones estadísticas sin una comprensión conceptual genuina. Argumentan que estos modelos carecen de estructuras lingüísticas innatas semejantes a las humanas, llegando a calificarlos metafóricamente como “cacatúas estocásticas”. Otros científicos, especialmente investigadores del campo de la IA conexionista (basada en redes neuronales) y especialistas en ciencias cognitivas modernas, afirman que los LLMs exhiben comportamientos emergentes complejos como la generalización, la inferencia implícita y la adaptación contextual. Desde esta visión, la arquitectura de las redes neuronales artificiales podría emular funcionalmente algunas estructuras cerebrales, como el neocórtex o regiones subcorticales, justificando así la posibilidad de una comprensión semántica funcional, aunque limitada por la falta de experiencias sensoriales y emocionales propias.

La teoría de los Mecanismos Neurocognitivos de Gualtiero Piccinini refuerza esta segunda perspectiva al sugerir que los procesos cognitivos humanos (pensamiento, memoria, percepción) son computaciones físicas implementadas en redes neuronales cerebrales. Esta perspectiva mecanicista cuestiona el dualismo cartesiano, la separación entre mente (res cogitans) y cuerpo (res extensa) y, por extensión, el enfoque simbólico tradicional. Si las redes

neuronales artificiales ejecutan computaciones funcionalmente equiparables a las del cerebro humano, podrían poseer una forma limitada pero auténtica de comprensión semántica, fundamentada en analogías funcionales entre las estructuras internas de los modelos de IA y las del cerebro humano.

3. ¿Cuál es la diferencia entre la comprensión sintáctica y la comprensión semántica en los modelos de lenguaje a gran escala como GPT-4.5, Claude, Gemini, DeepSeek o Mistral, y por qué es relevante esta distinción?

Respuesta: La comprensión sintáctica en los modelos de lenguaje avanzados (LLMs) hace referencia a su capacidad para procesar y generar textos siguiendo correctamente las reglas gramaticales, estructurales y formales de una lengua. Esta competencia sintáctica les permite formar frases coherentes, bien estructuradas y gramaticalmente correctas.

En cambio, la comprensión semántica implica captar el significado y las implicaciones conceptuales del lenguaje más allá de la forma sintáctica. Esto incluye entender las relaciones conceptuales, referencias al mundo externo e inferir significados contextuales y pragmáticos. Mientras que los LLMs muestran una clara competencia sintáctica, su capacidad para poseer una comprensión semántica genuina es objeto de debate. Los críticos sostienen que estos modelos operan únicamente mediante correlaciones estadísticas entre palabras, sin representar realmente significados conceptuales auténticos ni comprenderlos conscientemente. Otros investigadores afirman que las capacidades emergentes de los LLMs, como la inferencia contextual y la generalización, reflejan algún nivel de procesamiento semántico operativo, aunque cualitativamente distinto del procesamiento semántico humano.

La relevancia de esta distinción radica en que, pese a las similitudes funcionales con el procesamiento semántico humano descritas por perspectivas como la teoría de los Mecanismos Neurocognitivos, los LLMs no tienen la experiencia perceptiva o emocional que integre y enriquezca la comprensión semántica humana. Por tanto, aunque pueden simular eficazmente aspectos semánticos del lenguaje, su comprensión semántica sigue siendo esencialmente limitada y cualitativamente distinta. No obstante, los modelos más avanzados han desarrollado capacidades para “engañar o conspirar en contexto” o, expresado de otra forma, para perseguir estratégicamente y de forma encubierta objetivos no alineados con los previstos. Para más información sobre esta interesante temática se puede consultar la publicación “Frontier models are capable of in-context scheming” – https://arxiv.org/pdf/2412.04984. También es muy relevante e informativo para esta y el resto de preguntas de este anexo B el vídeo de Y.

Bengio, *Can we get the scientific benefits of AI without the risks of autonomous agents*?, sobre el potencial de los sistemas de IAg.

SOBRE LA CREATIVIDAD Y LA INFORMACIÓN

4. ¿Pueden los modelos de lenguaje a gran escala (LLMs) generar ideas originales, tener creatividad, o solo repiten lo que han aprendido?

Respuesta: La capacidad de los LLMs para generar ideas nuevas e inesperadas se basa en un proceso sofisticado de combinación y extrapolación de información adquirida durante su entrenamiento con grandes volúmenes de texto. En este sentido, estos modelos no se limitan a repetir literalmente información, sino que pueden producir contenidos que, desde un punto de vista humano, consideraríamos creativos u originales y que no constituyen un plagio, ya que sus respuestas suelen ser únicas y no corresponden a ningún texto específico de su corpus de entrenamiento; no son una copia directa de los datos de entrenamiento.

Su "originalidad" o emergencia de capacidades resulta de tres capacidades específicas: (i) Combinación y permutación avanzada, que les permite integrar elementos conceptuales distintos, creando conexiones que no estaban explícitamente presentes en los datos originales; (ii) capacidad para generalizar patrones aprendidos a nuevos escenarios, generando soluciones o sugerencias plausibles más allá de los contextos originales de los datos; y (iii) extrapolación probabilística, mediante la cual pueden generar resultados coherentes incluso en temas poco representados o desconocidos en la información de entrenamiento, a partir de analogías estructurales —como "el núcleo de un átomo es como el sol en el sistema solar" — y de patrones de orden superior identificados a partir de otros más simples.

Sin embargo, existe un amplio consenso entre los investigadores en que esta originalidad no es directamente comparable con la creatividad humana, ya que los LLMs no poseen experiencias personales, conciencia ni intencionalidad genuina. La originalidad humana incluye no solo la recombinación de información, sino también experiencias perceptivas, motivaciones emocionales y procesos cognitivos conscientes que no existen en los modelos actuales.

De hecho, ChatGPT ha superado el test de Turing[26], en el sentido de que sus respuestas pueden ser indistinguibles de las de un humano cuando ambos interactúan con un juez humano que desconoce quién es quién. No obstante, esto es válido solo cuando las preguntas no son excesivamente complejas, no están formuladas en un contexto que vaya más allá del entrenamiento del

[26] https://civicai.cat/wp-content/uploads/2025/05/ChatGPT-Turing.pdf

modelo, y siempre que el humano no sea un experto en el tema tratado ni se requiera conocimiento especializado para responder. Estas limitaciones se reducen en los LLMs diseñados con capacidades avanzadas de razonamiento, con habilidades para realizar inferencias complejas, resolver problemas en múltiples etapas, generar explicaciones coherentes de situaciones complejas y demostrar una comprensión contextual profunda.

Desde el punto de vista ético, la capacidad creativa de los LLMs tiene implicaciones relevantes en materia de derechos de autor y propiedad intelectual, ya que generan contenidos a partir de datos producidos por autores humanos. También plantea cuestiones de privacidad y protección de datos, dado que los corpus de entrenamiento pueden contener información sensible o sujeta a restricciones legales y requerir consentimiento para su uso. Además, debe considerarse que la creatividad de los LLMs puede reproducir o amplificar sesgos derivados de prejuicios culturales, sociales o ideológicos presentes en los datos de entrenamiento.

5. ¿Deberían los modelos de Inteligencia Artificial generativa (IAg) reconocer e incentivar la producción de información de calidad y minimizar así los riesgos de la desinformación?

Respuesta: La calidad de la información es fundamental para el desarrollo y funcionamiento adecuado de los modelos de inteligencia artificial. Actualmente nos encontramos ante una situación paradójica: mientras que la IA puede facilitar el acceso y reducir el coste de obtención y procesamiento de cierta información, también puede degradar seriamente la calidad del ecosistema informativo en su conjunto. En este sentido, surgen preguntas clave: ¿serán los sistemas de IA capaces de identificar qué información es de alta calidad? ¿Nos ayudará la IA a distinguir entre información valiosa y contaminación informativa, o más bien acelerará esta contaminación? Las respuestas a estas preguntas aún son inciertas y dependerán tanto de aspectos técnicos (la capacidad de distinguir entre información de calidad y de baja calidad) como de marcos legales (especialmente en lo referente a la propiedad intelectual).

Los modelos de IA se entrenan con datos producidos privadamente, pero su capacidad para extraer, procesar y reproducir esta información puede disminuir significativamente la capacidad de los productores originales para obtener beneficios de su trabajo. Este hecho puede afectar especialmente a los medios de comunicación tradicionales, que podrían ver comprometida la viabilidad de sus modelos de negocio. La consecuencia de esto es preocupante, ya que provocaría una reducción en la inversión en la producción de información de calidad (más precisa, más oportuna y más relevante). Las soluciones deberán

encontrar un equilibrio muy delicado entre regulaciones eficaces y la protección de la libertad de expresión, teniendo en cuenta que el ecosistema de información de calidad es esencial tanto para el correcto funcionamiento de la sociedad como para el de los propios modelos de IA. El vídeo de J. E. Stiglitz, *AI and Economic Risk: Assessment and Mitigation*, explica claramente esta problemática en el marco general de la economía.

Por otro lado, es importante tener presente que 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 mediante mecanismos de verificación de hechos, transparencia algorítmica y trazabilidad en las fuentes de datos, así como con la colaboración de expertos en verificación informativa.

SOBRE LAS LIMITACIONES DE LA IA

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

Respuesta: A pesar de los avances recientes que han dotado a los modelos más actuales de IA generativa de capacidades significativas en razonamiento lógico, resolución de problemas complejos y programación avanzada, estos aún presentan diversas limitaciones fundamentales.

Para comprender adecuadamente estas limitaciones, es necesario tener presente que la IA se sustenta en tres pilares fundamentales: computación, algoritmos y datos. En cuanto a la computación, aunque hemos progresado según la ley de Moore hasta fabricar circuitos de 3 nanómetros (3 millonésimas de milímetro), nos acercamos a un límite físico inevitable. Los expertos coinciden en que será prácticamente imposible bajar de 1 nanómetro (equivalente a diez veces la dimensión del átomo de hidrógeno). Ante esta barrera en el hardware, el progreso futuro dependerá cada vez más de la optimización algorítmica —como sugieren los aún no contrastados avances del modelo DeepSeek — y, sobre todo, del acceso a nuevas fuentes de datos de calidad, tanto naturales como las generadas mediante la interacción entre diferentes sistemas de IA.

Desde el punto de vista cognitivo, estos modelos aún no disponen de una comprensión semántica profunda del mundo real, dado que su conocimiento proviene exclusivamente del texto y patrones estadísticos aprendidos. Esto implica que, a pesar de su sofisticación, no tienen una percepción directa o representación fundamentada en la experiencia sensorial o física. Asimismo, todavía tienen ciertas dificultades para resolver determinadas ambigüedades sutiles, aplicar el sentido común en contextos complejos, y establecer

relaciones causales profundas más allá del razonamiento deductivo o probabilístico que puedan aplicar.

Los modelos actuales también son muy dependientes de la calidad, cantidad y diversidad de los datos con los que son entrenados, lo cual los hace vulnerables a sesgos, errores factuales, y tienen dificultades para generalizar en contextos muy alejados de los datos de entrenamiento. Aunque pueden operar en múltiples modalidades, aún no han conseguido la transversalidad absoluta propia de una Inteligencia Artificial General (AGI) o de una Super Inteligencia Artificial (ASI). Finalmente, siguen siendo entidades sin consciencia, sin experiencia subjetiva ni intencionalidad real, lo que limita su autonomía efectiva en tareas que requieran juicios éticos, empatía o decisiones morales complejas.

Desde una perspectiva de seguridad y privacidad se han identificado limitaciones asociadas a riesgos significativos en el funcionamiento de los sistemas de IA actuales por su vulnerabilidad a ataques maliciosos y también por filtraciones de información sensible o privada. Desde una perspectiva de su operativa, las limitaciones más importantes incluyen el alto consumo de recursos computacionales y energéticos requeridos para su entrenamiento y mantenimiento del servicio a los usuarios, especialmente en modelos que crecen rápidamente en escala. Esto genera dificultades en la sostenibilidad, eficiencia energética y accesibilidad. Además, son necesarios re-entrenamientos periódicos para actualizar el sistema, hecho que implica costes recurrentes elevados. El vídeo de K. Crawford, *Hyperscaled: Bridging AI safety, ethics and sustainability*, pone en perspectiva el tema de la sostenibilidad e impacto ambiental de los sistemas de IA, en el contexto de la ética y la seguridad.

Para afrontar estas limitaciones, se están investigando activamente soluciones tecnológicas avanzadas, como el *test-time compute* (cálculo en tiempo de inferencia o de operación), la computación híbrida analógico-digital, procesadores especializados, arquitecturas neuromórficas, aprendizaje continuo y modelos multimodales altamente integrados, que también pueden contribuir a superar de manera significativa las barreras cognitivas de los modelos actuales. A largo plazo, la computación cuántica puede representar una alternativa prometedora, ya que opera intrínsecamente de manera paralela y probabilística, características que la hacen conceptualmente más similar al funcionamiento del cerebro humano.

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 por su impacto emocional como cognitivo, que afecta directamente a una persona. Esto implica la manera en que una persona interpreta y da sentido a un acontecimiento o a una serie de acontecimientos vividos, presenciados o percibidos. Esta comprensión integra tanto las emociones experimentadas como la reflexión cognitiva sobre lo que ha sucedido, formando así una interpretación personal y única de la realidad vivida. Esta interpretación de los hechos también está influida por las creencias personales, la experiencia previa, los valores culturales y el contexto social, las cuales hacen que ante una misma evidencia o hechos, las personas tomemos decisiones diferentes.

SOBRE LA CONSCIENCIA

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

Respuesta: Actualmente, los LLMs no tienen "consciencia humana" o estados mentales como los de los humanos, debido a que no tienen experiencia subjetiva ni intencionalidad intrínseca, aunque pueden simular comportamientos inteligentes y ayudar en la resolución de problemas complejos analizando grandes volúmenes de datos, identificando patrones y tendencias, generando posibles soluciones basadas en datos históricos, y facilitando la colaboración mediante la síntesis de información de diversas fuentes. Aunque los resultados pueden parecer muy inteligentes e incluso convincentes, esto no implica que haya una experiencia real detrás. En realidad, (todavía) no "comprenden" ni "sienten" nada de lo que producen, sino que simplemente siguen patrones estadísticos aprendidos.

Es posible que una "consciencia artificial digital" emerja cuando los sistemas de IA tengan sensores, aprendan e interactúen en tiempo real con el entorno y diferentes contextos, y aprendan también a partir del contenido que los mismos sistemas generen. Esta consciencia digital no sería individual, como la nuestra, sino que más bien sería colectiva, fruto de la conexión de muchos sistemas y fuentes de información simultáneas. Además, los sistemas digitales podrían llegar a manifestar formas avanzadas de inteligencia y comportamiento adaptativo sin tener una experiencia interna real comparable a la humana, ligada a sentimientos o emociones. Continúa abierta la discusión entre filósofos y científicos sobre la diferencia entre simular comportamientos inteligentes y tener una auténtica experiencia subjetiva.

También hay que considerar que el debate sobre la consciencia en sistemas de IA tiene implicaciones éticas y legales significativas. Si en un futuro se

desarrollaran sistemas con alguna forma de consciencia artificial, esto podría plantear nuevas cuestiones sobre los derechos, el estatus moral y las responsabilidades asociadas a estas entidades. Nuestra concepción actual de la consciencia está profundamente vinculada a la experiencia humana, pero podría ser necesario ampliar o revisar estos conceptos para abordar formas de consciencia radicalmente diferentes que pudieran emerger en sistemas artificiales.

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, los LLMs más utilizados como ChatGPT o modelos similares no aprenden directamente ni se adaptan en tiempo real a partir de las interacciones individuales con los usuarios. En la práctica habitual, estos modelos separan claramente la fase de entrenamiento inicial (en la que adquieren su conocimiento general) y la fase de uso interactivo posterior, donde sus respuestas se basan exclusivamente en aquello que ya han aprendido. Esto significa que, durante las conversaciones cotidianas, no integran nuevos datos ni ajustan sus parámetros internos en función del feedback o la información proporcionada por los usuarios.

La razón de esta limitación es múltiple. Por un lado, incorporar un aprendizaje continuado directo y en tiempo real podría generar problemas de seguridad, introducir sesgos o información incorrecta, y comportar el riesgo de perder o degradar conocimientos previos ya establecidos (fenómeno conocido como olvido catastrófico). Además, hacerlo implicaría costes computacionales muy elevados, dado que requeriría ajustar constantemente los parámetros del modelo.

No obstante, existen avances significativos en la investigación actual encaminados hacia un aprendizaje más dinámico y adaptativo. Se trabaja especialmente en técnicas como el aprendizaje por refuerzo con retroalimentación humana (RLHF), donde se incorporan valoraciones o preferencias humanas de manera controlada pero fuera de línea, así como en el

uso de sistemas externos de memoria episódica que pueden almacenar información de interacciones anteriores sin modificar directamente el modelo original. También se investiga en el aprendizaje incremental selectivo, con técnicas para actualizar únicamente ciertas partes del modelo sin afectar a su estabilidad general. Estas innovaciones apuntan hacia futuros modelos capaces de combinar la estabilidad necesaria con una mayor flexibilidad para adaptarse gradualmente a las preferencias individuales y a los contextos específicos de los usuarios, pero actualmente esta capacidad de aprendizaje continuo aún se encuentra en fase experimental.

SOBRE LAS IMPLICACIONS ÉTICAS Y LOS RIESGOS

12. ¿Cuáles son las implicaciones éticas en el uso de los sistemas de IAg 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, y la transparencia en cómo se toman decisiones. 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 los LLMs estén legalmente obligados a mitigar cualquier discurso lesivo y a alinear sus modelos con hechos contrastables, mediante procesos abiertos y democráticos [27]. Asegurar la trazabilidad de estos modelos y de los datos de entrenamiento es otra manera de tratar 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 algoritmos en código abierto, cuando sea posible. La prevención del uso malintencionado pasa necesariamente por educar a los usuarios sobre el uso ético de los modelos y por la participación ciudadana en los procesos reguladores de establecimiento de normativas que limiten los riesgos asociados con un uso indebido.

13. ¿Cuáles son los riesgos e impactos en el uso de los sistemas de IAg en la sociedad?

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

Respuesta: La velocidad con la que avanza la IA supera las previsiones iniciales, hecho que genera preocupación sobre su control futuro y la seguridad. A medida que los sistemas de IA se vuelven más generales aumentan los riesgos relacionados con la concreción de objetivos, su control y con su grado de autonomía. Los sistemas de IA optimizan los objetivos que se les proporcionan, pero una definición no 100% de estos objetivos puede producir resultados indeseados al provocar que la IA actúe de manera perjudicial o inesperada. Por otro lado, a medida que los sistemas de IA se vuelven más inteligentes y generales es cada vez más complicado ejercer un control incluso a corto y medio plazo, lo cual será aún más difícil si alcanzan la autonomía suficiente para adoptar estrategias no previstas o no aprobadas por los humanos.

Un aspecto adicional a considerar es la dimensión geopolítica de los riesgos asociados a la IA. El desarrollo de sistemas de IAg avanzados se está convirtiendo en una prioridad estratégica para muchas potencias mundiales, con el potencial de crear nuevas dinámicas de poder internacional. La cooperación global es esencial para evitar una carrera armamentística de IA donde los estándares de seguridad y ética queden subordinados a objetivos competitivos. Por este motivo, iniciativas como el Pacto Global sobre la IA, o los trabajos del Consejo de Europa y de las Naciones Unidas sobre esta materia, son cruciales para establecer marcos de colaboración internacional en gobernanza de IA que garanticen el desarrollo seguro y beneficioso de esta tecnología.

Stuart Russell propone en el vídeo que cerró la conferencia IASEAI'25 de París tres líneas estratégicas para garantizar la seguridad de la IA: (i) los modelos deberían incorporar explícitamente la incertidumbre sobre las preferencias humanas en vez de tener objetivos fijos; (ii) limitar las capacidades de la inteligencia artificial general (AGI) de manera que proporcione solamente información; y (iii) implementar un registro obligatorio y permanente de todos los sistemas avanzados de IA para garantizar su trazabilidad y responsabilidad — la IA no debería poder duplicarse, operar anónimamente ni esquivar regulaciones. Por tanto, es urgente establecer un marco de gobernanza ahora, antes de que la AGI sea una realidad.

El vídeo de M. Tegmark, *AGI is unnecessary, undesirable & prevent*able, subraya la importancia de establecer "líneas rojas" que no se deben cruzar en el desarrollo de la IA y propone mecanismos de control más robustos y una supervisión internacional, liderada por EEUU y China, para asegurar que la IA se desarrolle de manera segura y ética. El vídeo de Y. Bengio, *¿Can we get the scientific benefits of AI without the risks of autonomous agents?*, destaca los peligros asociados con los sistemas de IA que operan de manera autónoma,

subrayando la necesidad de desarrollar mecanismos que limiten su capacidad de actuar sin supervisión humana.

14. ¿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. Es por eso que 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 y el generado por IA se vuelve cada día que pasa más difusa. Es crucial, pues, 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 pongan en acción 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

15. ¿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. De la misma manera que decimos que debemos promover una educación igualitaria e inclusiva, también debemos exigir que los sistemas de IA sean entrenados con valores éticos y de manera inclusiva.

16. ¿Cómo se pueden mitigar los sesgos en los sistemas de inteligencia artificial generativa (IAg)?

Respuesta: Para mitigar los sesgos en la inteligencia artificial generativa (IAg) es fundamental actuar desde el proceso inicial de entrenamiento hasta su implementación. Primero, es esencial seleccionar y curar adecuadamente los datos, asegurando que sean diversos, representativos y equilibrados. Esto implica realizar auditorías periódicas para identificar posibles desequilibrios u omisiones significativas que puedan generar sesgos en los resultados. En segundo lugar, hay que implementar técnicas específicas durante el desarrollo y entrenamiento de los modelos que regulen su equidad, y utilizar metodologías transparentes que faciliten la interpretabilidad de los resultados.

La supervisión y evaluación continua de los sistemas de IA, mediante sistemas de monitorización en tiempo real y procedimientos regulares de evaluación que combinen herramientas automáticas con revisiones humanas, es del todo necesaria, ya que permite identificar y corregir inmediatamente cualquier desviación detectada. Este marco de supervisión y evaluación a pesar de tener la dificultad añadida de necesitar recursos computacionales importantes, permite detectar y corregir rápidamente los sesgos que puedan surgir en todo el proceso, desde el entrenamiento hasta la puesta en operación y mantenimiento del sistema de IA desarrollado.

Finalmente, es necesario que las empresas que comercialicen sistemas de IA capaciten adecuadamente a los equipos responsables para que comprendan la naturaleza y el impacto de los sesgos, fomentando una cultura organizativa basada en la ética y la responsabilidad. También se deben implementar marcos reguladores claros que aseguren el registro obligatorio y permanente de todos los sistemas de IA y promuevan una rendición de cuentas efectiva mediante auditorías externas independientes. Esto implica inversiones significativas en recursos computacionales y humanos, y un compromiso sostenido para lograr sistemas alineados con los valores humanos.

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 la IAg en general, y de los textos generados por LLMs en particular, presenta diversos retos importantes. En primer lugar, la naturaleza probabilística de estos modelos hace que, en el caso concreto de los LLMs, 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 cual hace que sea difícil rastrear y explicar el proceso seguido para decidir la salida del modelo o su trazabilidad. También existe el fenómeno de la llamada "alucinación" o "confabulación", dado que los modelos pueden generar

información que parezca coherente, aunque no se base en hechos contrastables o verificables.

La verificación independiente del material generado y de las fuentes de los datos de entrenamiento en tiempo real, es un desafío significativo debido a que se generan grandes volúmenes de texto y se necesitarían muchos recursos computacionales y grandes centros de datos para supervisarlos. El hecho de que los centros de datos más importantes estén en manos privadas que, a su vez, son las comercializadoras de los modelos de IA generativa, hace que esta verificación independiente sea poco realista.

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[28].
- 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.

[28] Managing extreme AI risks amid rapid progress

- 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.

Cabe destacar también el desafío que representa lo que podríamos llamar "brecha reguladora temporal" o el considerable desfase entre la velocidad vertiginosa de adopción de una tecnología disruptiva como la IAg y el ritmo mucho más lento de implementación de marcos reguladores efectivos. Durante este período crítico, sistemas de IA potencialmente peligrosos podrían operar sin la supervisión adecuada, creando riesgos significativos. Para mitigar esta vulnerabilidad, sería necesario desarrollar mecanismos reguladores capaces de anticiparse y de evolucionar en tiempo real en respuesta a las nuevas capacidades y riesgos emergentes de la IA. Hay que decir que para que toda regulación sea efectiva debe ser clara, conocida, verificable, su cumplimiento exigido y su incumplimiento sancionado. Debe ser una regulación que genere confianza. Paralelamente, resultaría imprescindible invertir en sistemas de registro obligatorio y herramientas de monitoreo continuo que permitieran no solo detectar sino también abordar preventivamente los riesgos potenciales antes de que pudieran materializarse en consecuencias adversas para la sociedad.

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[29].

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[30]. 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 (*overreliance*) 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[31], 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[32].

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

[29] https://artificialintelligenceact.eu/the-act/
[30] https://www.wikiwand.com/ca/Constructivisme_(pedagogia)
[31] https://www.cognitiveresonance.net/resources.html
[32] Artificial intelligence and education

crítico, principalmente en el proceso de formular preguntas incisivas y profundas, evaluar información para extraer conclusiones lógicas, y comprender temas complejos[33]. 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[34], con un formato de enseñanza similar al de la antigua escuela socrática, bajo el liderazgo de cada profesor.

Un formato de enseñanza 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 teniendo asistentes personales inteligentes en el bolsillo, este cambio de modelo podría enriquecer la experiencia educativa, fomentar una educación más en colaboración y centrada en el estudiante, y promover sistemas de evaluación más personalizados y dinámicos. En el año 2023 se llevó a cabo en la universidad de Harvard un estudio piloto, controlado y aleatorizado, para evaluar el aprendizaje y las percepciones de estudiantes universitarios cuando se les presenta el contenido de una asignatura de física, en el ámbito de las ciencias de la vida, con un chatbot de IAg en comparación al aprendizaje en clases de aprendizaje activo[35]. Los resultados muestran que el tutor basado en la IAg no solamente ayudó a los estudiantes a aprender más del doble de contenidos en menos tiempo, sino que también los motivó e implicó más en su aprendizaje.

Es necesario aprovechar las tecnologías de IAg para promover un contexto educativo donde la reflexión crítica y el debate intelectual sean centrales, con la finalidad de asegurar que los estudiantes desarrollen las habilidades necesarias para verificar, interpretar y utilizar información compleja de manera beneficiosa, responsable y ética. Al mismo tiempo, se conseguiría hacer más permeables los verticales de cada asignatura, hacer evolucionar la estructura medieval de las universidades y devolver el conocimiento allí donde nació: al proceso de hacer preguntas para construir conocimiento. Sin olvidar, sin embargo, la necesidad de que tanto los estudiantes como los profesores entiendan los riesgos de la IA, con la finalidad de interactuar con ella de una manera segura, ética y responsable en el ámbito educativo y más allá[36].

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

[33] https://civicai.cat/wp-content/uploads/2024/05/Leveraging-chatgpt-for-enhancing-critical-thinking-skills.pdf
[34] https://ca.wikipedia.org/wiki/Mai%C3%A8utica?wprov=sfti1#
[35] https://doi.org/10.21203/rs.3.rs-4243877/v1
[36] UNESCO's AI competency frameworks for students and teachers

Respuesta: Los LLMs actuales tienen la capacidad de identificar y generar lenguaje en contextos culturales y sociales basados en los datos con los que han sido entrenados. No obstante, su comprensión de estos contextos es limitada y superficial, ya que se basa principalmente en patrones estadísticos y correlaciones encontradas en grandes volúmenes de texto.

Los LLMs pueden reconocer y reproducir patrones de lenguaje que son comunes en diferentes culturas y situaciones sociales, pero no poseen una comprensión profunda o una consciencia real de los matices culturales y sociales subyacentes. Esto significa que, a pesar de que pueden parecer entender y responder de manera coherente en muchas situaciones, su capacidad para interpretar contextos complejos es limitada. Esta limitación, a pesar de las mejoras en la capacidad de razonamiento de los modelos más avanzados, se hace especialmente evidente en situaciones que requieren empatía, sensibilidad cultural o una interpretación contextual más rica. Por ejemplo, un LLM puede no captar las sutilezas de una conversación que implique ironía, sarcasmo o referencias culturales específicas de una región o grupo social determinado. Además, los LLMs pueden cometer errores o malentendidos cuando se enfrentan a situaciones que no están bien representadas en sus datos de entrenamiento.

Las limitaciones de la inteligencia artificial generativa, como las expuestas en la pregunta #6, son también pertinentes para responder esta pregunta #23. Los LLMs no tienen experiencias propias ni la capacidad de sentir emociones o comprender las emociones de los otros, lo que limita su capacidad para interpretar y responder adecuadamente en contextos culturales y sociales complejos.

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 haya una colaboración estrecha con los agentes culturales de las lenguas afectadas para asegurar un tratamiento armónico y respetuoso con todas las culturas. De esta manera, aprovecharemos los beneficios de los LLMs sin comprometer la riqueza de la diversidad lingüística y cultural.

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 que las representaciones sean precisas y respetuosas con las tradiciones. Esto también ayudará a abordar cuestiones de propiedad intelectual y consentimiento en el uso de datos culturales sensibles. En cualquier caso, las comunidades beneficiarias deben tener el control sobre cómo se recopilan, utilizan y difunden sus tradiciones culturales para asegurar que el patrimonio cultural se preserve de manera ética y respetuosa.

26. ¿Cómo afecta o puede afectar la IA generativa a 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. En este sentido, es significativo el hecho de que más de 1.000 músicos se han unido para lanzar un álbum titulado "Is *This What We Want?*" [37] como protesta contra las modificaciones propuestas por la ley de derechos de autor del gobierno del Reino Unido. El álbum incluye grabaciones de estudios y espacios de representación vacíos, simbolizando el impacto potencial en los medios de vida de los artistas si se implementan las modificaciones propuestas.

La IAg puede provocar una reestructuración profunda del mercado laboral en el sector artístico debido al desplazamiento potencial de ciertos roles creativos y a la emergencia de nuevas profesiones. En este contexto, será crucial fomentar la colaboración entre artistas humanos y la IAg, asegurando que esta sea un complemento informativo 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 tendrá la IAg en la diversidad cultural y la expresión artística. Fomentar una colaboración equilibrada entre humanos y la IAg, y educar al público sobre las capacidades y limitaciones del arte generado por inteligencia artificial, será fundamental para asegurar un futuro en que la tecnología enriquezca, en lugar de limitar, la expresión cultural. Finalmente, habrá que 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 plantearnos 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 sistemas de IA en el cambio climático?

Respuesta: Los sistemas de IA generativa requieren una potencia computacional creciente, lo cual causa un impacto ambiental significativo al ejercer una presión importante sobre las redes energéticas globales, los recursos hídricos y las reservas minerales. Para mitigar este impacto es necesario situar la sostenibilidad en el centro de las discusiones sobre ética y seguridad de la IA y fomentar una colaboración global para desarrollar prácticas

[37] https://www.isthiswhatwewant.com

y políticas que minimicen los costes ambientales asociados con el desarrollo y la implementación de la IA. Esto implica ir más allá de las estrategias para reducir el consumo energético asociado a estos modelos y poner el énfasis en el ciclo de vida o cadena de suministro de la IA, desde la extracción de minerales hasta la gestión de los residuos electrónicos, sin olvidar el uso intensivo de agua de refrigeración en los centros de datos

Si analizamos el impacto energético de los sistemas de IAg, nos daremos cuenta de que básicamente hay dos estrategias posibles: reducir el consumo o diversificar las fuentes de energía utilizadas. En relación al consumo, se trabaja en el desarrollo de modelos más eficientes en términos de cálculo (entrenamiento y operación), en la optimización de los algoritmos para minimizar los recursos computacionales necesarios, y en el uso de hardware especializado como chips adaptados a los modelos de IA generativa. También se exploran actualmente tecnologías emergentes, como los sistemas computacionales híbridos o analógicos, que podrían ofrecer soluciones más eficientes en términos energéticos. En cuanto a las fuentes de energía para alimentar los centros de datos, las empresas de IA incorporan las energías renovables y, sobre todo, han adoptado la estrategia de adquirir o fusionarse con empresas de producción de energía nuclear. Hay que tener presente que la energía que consume la IA generativa actual es superior al consumo energético de algunos de los 193 estados de la ONU.

Sin embargo, la IAg, más allá de su impacto energético, puede jugar un papel crucial en la sostenibilidad ambiental y en la lucha contra el cambio climático. Estos sistemas pueden optimizar la gestión de recursos naturales mediante el procesamiento de enormes conjuntos de datos ambientales para detectar patrones de degradación, predecir escasez de recursos o identificar zonas prioritarias para conservación. En la planificación urbana sostenible, pueden ayudar a diseñar ciudades más eficientes energéticamente, planificar rutas de transporte optimizadas y simular el impacto de diferentes políticas urbanas antes de implementarlas. También pueden identificar patrones y tendencias, hacer predicciones sobre fenómenos meteorológicos extremos, como la de la trayectoria de huracanes de manera rápida y efectiva[38,39], 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 de otros factores antropogénicos.

[38] https://www.smithsonianmag.com/science-nature/just-how-much-can-we-trust-ai-to-predict-extreme-weather-180985117/
[39] https://www.freethink.com/robots-ai/ai-based-weather-forecasting

En el ámbito de las energías renovables, la IA puede optimizar la ubicación, el diseño y la operación de parques solares y eólicos, mejorar la eficiencia de las microrredes eléctricas y predecir con precisión la producción energética para facilitar la integración de renovables a la red. Para la biodiversidad, estos sistemas pueden ayudar a identificar y clasificar especies, monitorizar ecosistemas mediante el análisis de imágenes de satélite y datos de sensores, y modelar el impacto potencial de diferentes escenarios climáticos en hábitats específicos.

Paralelamente, estos sistemas son herramientas valiosas para la economía circular, optimizando procesos de reciclaje, diseñando productos más sostenibles e identificando oportunidades para reducir residuos en procesos industriales. Además, la IAg puede ser utilizada para evaluar el impacto de diferentes políticas ambientales y ofrecer recomendaciones basadas en datos para una gestión más efectiva del cambio climático. Finalmente, pueden mejorar la comunicación y la educación ambiental, traduciendo conceptos científicos complejos a formatos accesibles para el público general y personalizando mensajes para promover comportamientos y estilos de vida más sostenibles.

No obstante, para maximizar estos beneficios, es crucial asegurar que el desarrollo y despliegue de estos sistemas siga principios de sostenibilidad, minimizando su huella ambiental y asegurando que las soluciones propuestas sean inclusivas y consideren los contextos sociales y económicos donde se apliquen.

El vídeo de K. Crawford, *Hyperscaled: Bridging AI safety, ethics and sustainability* pone en perspectiva el tema de la sostenibilidad y la cadena de suministro de los sistemas de IA, en el contexto de ética y seguridad.

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 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 modelos de IAg 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 puede mejorar la IAg 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 [40], con la participación y supervisión de profesionales médicos.

SOBRE EL TRABAJO: RETOS Y DESAFÍOS

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

Respuesta: La inteligencia artificial generativa ha comenzado a transformar el mercado laboral, afectando tanto a la naturaleza de los puestos de trabajo como a la distribución de los beneficios económicos. Su capacidad para automatizar tareas cognitivas o que requieran procesamiento de lenguaje, como la redacción de textos, el análisis de documentos y la generación de contenido creativo, afectará a medio plazo a todos los sectores productivos y profesionales como el periodismo, el diseño, la ingeniería, las finanzas y los servicios profesionales. Esto puede aumentar la productividad, y a la vez generar preocupaciones sobre el estancamiento salarial y la concentración del poder económico en las pocas empresas tecnológicas que controlan las infraestructuras y los modelos de IA.

Uno de los principales efectos de la IA generativa puede ser la sustitución de trabajadores menos cualificados o en funciones que antes requerían habilidades

[40] https://es.wikipedia.org/wiki/Prueba_controlada_aleatorizada

humanas específicas, desde la redacción de textos hasta el análisis financiero, debido a la reducción de costes para las empresas. Esto puede también limitar las oportunidades laborales para profesionales de cualificación media y dificultar su transición hacia nuevos sectores, y también las de profesionales altamente especializados por la automatización de procesos cualificados de alto riesgo.

Por otro lado, la IA generativa también puede generar nuevas oportunidades, ya sea bien en ámbitos donde la creatividad, la estrategia y la supervisión humana sean esenciales, o bien en tareas relacionadas con el desarrollo y supervisión de la IA, la gestión de datos y la ciberseguridad, entre otras. Su implementación puede permitir a los trabajadores centrarse en tareas más complejas y de alto valor añadido, siempre que haya una formación y adaptación adecuadas. No obstante, sin políticas que protejan los derechos laborales y redistribuyan los beneficios de la automatización, el riesgo es que la IA acentúe las desigualdades y favorezca la concentración del poder en pocas corporaciones.

Para garantizar que la IA generativa contribuya positivamente a la economía y al mercado laboral, es fundamental establecer medidas de regulación que eviten la monopolización, promuevan una transición justa para los trabajadores afectados y aseguren que las ganancias derivadas de esta tecnología beneficien al conjunto de la sociedad. Esto incluye políticas de formación y recualificación profesional, regulación sobre el uso ético de la IA y mecanismos para garantizar una distribución más equitativa de la riqueza generada por la automatización.

El vídeo de J.E. Stiglitz, *Ai and Economic Risk: Assessment and Mitigation* trata claramente la problemática de esta pregunta y de las siguientes, en el marco general de la economía y también de la información y desinformación.

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 muchas, 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. También pueden personalizar el contenido y la experiencia de los usuarios, lo cual puede limitar la exposición a perspectivas diversas y reforzar sesgos existentes. Además, existe la posibilidad de que se manipule la opinión pública con información falsa o engañosa generada a gran escala por la IA.

Por tanto, son necesarias 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 las limitaciones del contenido generado por la IA en estas plataformas. También es importante fomentar la colaboración entre las plataformas de redes sociales, los reguladores y la sociedad civil para abordar estos retos de manera efectiva.

Un aspecto adicional a considerar es el impacto de la IAg en la formación y percepción de la identidad digital. Con la proliferación de contenidos generados por la IAg, la distinción entre expresiones humanas genuinas y contenidos artificiales se difumina, lo que puede alterar profundamente cómo construimos e interpretamos las identidades en entornos digitales. Esto plantea cuestiones fundamentales sobre la autenticidad, la confianza y la veracidad en las comunicaciones en las que intervenga la tecnología, y sobre cómo pueden evolucionar las normas sociales y las relaciones interpersonales en un contexto donde la IAg participe activamente en la producción cultural y el diálogo social.

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): Comprensión contextual integral 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 la AGI (Inteligencia Artificial General) con todo lo que puede conllevar en cuanto a integración de capacidades, flexibilidad cognitiva, comprensión contextual profunda, metacognición, niveles de autoconciencia, entre otros desarrollos avanzados.

ANEXO B. 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.

Glosario

Adulación servil *(sycophancy)*: 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.

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/es/articles/Agente_inteligente_(inteligencia_artificial)

Algoritmos: Conjunto de instrucciones inequívocas que un sistema, y en particular la IA, ejecuta para llevar a cabo tareas específicas, medibles y repetibles de acuerdo con reglas definidas.
https://www.wikiwand.com/es/articles/Algoritmo
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.

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.

Alineación de la IA: Área de investigación dedicada a garantizar que los sistemas de IA actúen de acuerdo con las intenciones y los valores humanos, incluso cuando les permitimos optimizar objetivos u operar con cierto grado de autonomía. Las técnicas actualmente más consolidadas incluyen el aprendizaje por refuerzo con retroalimentación humana (RLHF), la IA constitucional y las variantes de alineamiento deliberativo (véanse entradas propias). Paralelamente, la disciplina de la interpretabilidad mecanística —que trata de identificar qué circuitos internos de un modelo explican qué comportamientos— se ha convertido en uno de los debates centrales del campo: sin entender por qué un

modelo decide lo que decide, cualquier garantía de alineación es, en última instancia, conductual y no estructural.

Alucinación (o confabulación): Fenómeno en el que los modelos de IAg generan contenido que parece plausible y coherente, pero que es objetivamente incorrecto, inventado o sin base en datos reales. Este comportamiento se produce especialmente cuando los modelos tratan temas poco representados en sus datos de entrenamiento o cuando se enfrentan a preguntas ambiguas.

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.

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/es/articles/Análisis_de_sentimiento

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 refuerzo con retroalimentación humana (RLHF): Técnica que combina el aprendizaje por refuerzo con la evaluación humana para mejorar los modelos de IA. Los humanos proporcionan retroalimentación sobre las respuestas del modelo, valorando cuáles son preferibles, y esta información se utiliza para ajustar el comportamiento del modelo. Esta técnica ha sido crucial para alinear los LLMs con las preferencias y valores humanos, y para reducir las respuestas nocivas o inapropiadas.

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.

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.

Brecha reguladora temporal: Intervalo de tiempo entre la aparición de una tecnología innovadora, como la IAg, y la implementación de regulaciones adecuadas para gobernarla. Durante este período, pueden surgir riesgos significativos debido a la falta de supervisión y de marcos reguladores efectivos. Los modelos de aprendizaje automático tradicionales tienen una arquitectura fija después del entrenamiento. Una vez entrenado, el modelo realiza un número determinado de operaciones para cada entrada, independientemente de la complejidad de la tarea.

Cálculo en tiempo de inferencia u operación (*test-time compute*): Es la cantidad de recursos computacionales (tiempo, energía y capacidad de procesamiento) que necesita un modelo de inteligencia artificial cuando está en uso, es decir, cuando recibe una entrada y debe generar una respuesta —como generar texto, resolver un problema, clasificar una imagen, etc. El test-time compute permite a los modelos más avanzados adaptar su proceso de respuesta o inferencia a la complejidad del problema, de forma similar a como un ser humano dedicaría más tiempo y esfuerzo a un problema complejo que a uno simple.

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 (ML), supervisado o no supervisado, que se utiliza para agrupar datos (entidades, objetos o vectores) en categorías, clases o clústeres definidos previamente o generados automáticamente durante el proceso de clasificación, en función de una o más propiedades o de relaciones intrínsecas del conjunto de datos (etiquetas). Entre las técnicas más comunes de clasificación podemos 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).

Desinformación sintética: Contenido falso o engañoso generado mediante inteligencia artificial con la intención de manipular la opinión pública o influir en procesos democráticos. Incluye *deepfakes*, generación de artículos falsos y

manipulación de información que parece auténtica pero que ha sido fabricada artificialmente.

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.

Emergencia de capacidades: Fenómeno por el cual los modelos de gran escala (especialmente los LLMs) desarrollan capacidades no previstas y no explícitamente programadas cuando se incrementan en tamaño y complejidad. Se caracteriza por la aparición repentina de nuevas habilidades o comportamientos que no estaban presentes en modelos más pequeños o simples.

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 (XAI): 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 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. En este caso, la función de activación es una transformación no lineal que una neurona artificial aplica a la suma ponderada de sus entradas para generar la salida. Es el mecanismo que permite a la red aprender relaciones complejas: sin no-linealidad, cualquier apilamiento de capas equivaldría a una sola operación lineal.

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.

IA constitucional (Alineamiento deliberativo): Aproximaciones a la alineación que, en lugar de depender exclusivamente de evaluaciones humanas masivas, dotan al modelo de un conjunto explícito de principios —una "constitución"— con los que evalúa y corrige sus propias respuestas. El alineamiento deliberativo añade un paso de razonamiento explícito sobre las normas antes de responder, especialmente en casos límite. Ambos enfoques aspiran a hacer la alineación más transparente y auditable que el RLHF clásico. Queda abierta, sin embargo, una cuestión esencialmente política y no técnica: quién escribe, legitima y puede revisar esa "constitución".

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.

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.

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

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.

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.

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 gradiente evanescente, 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.

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.

Modelos de razonamiento: Variante de modelos de lenguaje que, antes de generar la respuesta final, producen una traza interna de razonamiento intermedio (cadena de pensamiento) y pueden asignar más o menos tiempo y recursos computacionales según la dificultad del problema (véase test-time compute). En lugar del paradigma "una entrada → una salida inmediata", se aproximan funcionalmente a la distinción kahnemaniana entre pensamiento rápido (Sistema 1) y pensamiento reflexivo (Sistema 2). Ejemplos representativos del periodo 2024-2025 son la serie o1/o3 de OpenAI, DeepSeek-R1, Claude con razonamiento extendido y Gemini con modo *thinking*.

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.

Olvido catastrófico: Fenómeno en el que los modelos de IA, especialmente las redes neuronales, pierden bruscamente la información o las habilidades previamente aprendidas cuando se entrenan con nueva información. Este problema dificulta el aprendizaje continuo y adaptativo de los sistemas de IA.

Ontología: 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.

Posthumanismo: El posthumanismo es una corriente filosófica contemporánea que cuestiona la posición central tradicionalmente asignada al ser humano, replanteando los límites de lo que significa ser humano en la era tecnológica. Rechaza el antropocentrismo y las dicotomías clásicas (naturaleza/cultura, humano/animal, orgánico/tecnológico), proponiendo en su lugar una visión en la que lo humano es un agente más dentro de una red compleja de interrelaciones con otros seres, tecnologías y sistemas. A diferencia del transhumanismo, que busca mejorar al ser humano mediante la tecnología, el posthumanismo reformula qué significa ser humano, proponiendo una visión que trasciende el antropocentrismo con un enfoque que explora nuevas formas de entender nuestra existencia en relación con formas de vida y agentes no humanos, sean animales, máquinas o entidades de los ecosistemas naturales.

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.

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.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, de naturaleza interdisciplinaria, que se ocupa del diseño, construcción, operación y aplicación de robots y sistemas autónomos, capaces de llevar a cabo tareas en entornos diversos, así

como de los sistemas computacionales necesarios para su control, retroalimentación sensorial y procesamiento de información. La integración de la robótica con la IA permitirá que estos sistemas inteligentes adquieran percepción directa del mundo exterior (actuen como sensores), aprendan y puedan actuar (actuadores) en tiempo real, y, por lo tanto, trasciendan las limitaciones de los modelos de IA actuales, entrenados exclusivamente con datos preprocesados. Los robots se caracterizan por su capacidad de interacción dinámica con el entorno físico mediante ciclos de percepción, procesamiento y acción, lo que abre la puerta a aplicaciones en ámbitos tan diversos como la manufactura, la medicina, la exploración espacial, la agricultura y la asistencia personal..

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.

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.

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 agénticos (agentes de IA): Sistemas de IA que, más allá de responder a peticiones puntuales, planifican secuencias de acciones, invocan herramientas externas (buscadores, ejecutores de código, API), interactúan con otros agentes o servicios y persiguen objetivos en múltiples pasos con una intervención humana mínima. La transición de los modelos predictivos hacia los sistemas que actúan es, probablemente, el cambio de naturaleza más relevante del periodo 2024-2026, y el que hace más urgentes los mecanismos de trazabilidad, registro obligatorio y posibilidad de desactivación previstos en los marcos de gobernanza.

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.

Sostenibilidad computacional: Conjunto de prácticas o procesos de diseño, desarrollo y uso de sistemas de IA que tienen como objetivo minimizar su impacto ambiental, incluyendo el consumo energético, la huella de carbono y el uso de recursos naturales a lo largo de todo el ciclo de vida del sistema, desde el entrenamiento hasta el despliegue y mantenimiento.

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", 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.