Proyecto de investigación Consciencia y Sociedad Distópica

Comunidad en Telegram. 23 de febrero de 2026

Enlace de suscripción al canal en Telegram: https://t.me/socdistopica

LA INTELIGENCIA ARTIFICIAL NOS AFECTA A TODOS, ¿PODEMOS EDUCARLA?

La tecnología atraviesa transversalmente casi todas las facetas de nuestra vida, directa o indirectamente. Sin embargo, estos avances se producen desde la visión del mercado y sin el adecuado debate social sobre las consecuencias globales de su implementación. Sabiendo lo cual, Sergio Quiroga Morla[1], miembro del Equipo de Dirección de este Proyecto, comparte en esta Comunidad una serie de artículos con reflexiones relacionadas a los avances tecnológicos desde el punto de vista de la Consciencia.

En el primero de ellos, "Crónica de la civilización sintética: del vapor a la inteligencia agéntica", publicado el pasado 9 de febrero, se trazó el arco evolutivo de la humanidad desde la mecanización del músculo hasta la delegación del razonamiento en sistemas sintéticos, dejando abierto el verdadero debate que nos debemos como Humanidad: ¿qué realmente nos hace humanos y cómo gestionamos el conocimiento colectivo de los humanos?

En este nuevo artículo, se habla de qué es y cómo funciona la IA y del mito que circula en el imaginario popular respecto a la posibilidad de educarla o hackearla como usuarios.

¿QUÉ ES Y CÓMO FUNCIONA LA INTELIGENCIA ARTIFICIAL

La inteligencia artificial (IA) se define como la simulación de procesos de inteligencia humana por parte de máquinas, incluyendo aprendizaje, razonamiento, percepción y toma de decisiones. Su desarrollo ha sido impulsado por avances en matemáticas, informática, neurociencia y computación.

¿Cómo aprende?

La IA aprende a partir de datos que procesa con algoritmos (una secuencia ordenada de pasos para resolver un problema o lograr un resultado) capaces de identificar patrones y tomar decisiones, imitando ciertas funciones del cerebro humano, como el

[1] Sergio Quiroga Morla es Licenciado en Administración de Empresas (UCA), Master en técnicas cuantitativas (UGR), Diploma en Inteligencia Artificial (UCES). Ejerce de docente, consultor, investigador independiente, autor y divulgador. Es Fundador y Director de Ser Minimal - Desarrollo Humano Consciente (serminimal.com).

aprendizaje y el reconocimiento de tendencias. El motor de este aprendizaje es el *machine learning* (aprendizaje automático), donde la IA utiliza grandes volúmenes de datos para entrenar modelos matemáticos y mejorar sus resultados con cada repetición.

Existen tres enfoques principales de aprendizaje:

+**Aprendizaje supervisado**: la IA aprende a partir de ejemplos etiquetados por humanos, ajustando sus respuestas para acertar cada vez más en tareas como clasificación de imágenes, detección de fraudes o traducción automática.

+**Aprendizaje no supervisado**: explora grandes conjuntos de datos sin etiquetas para descubrir por sí sola patrones (correlaciones, estructuras estadísticas), segmentar información o detectar anomalías, útil donde los resultados correctos no están claros de antemano, también conocida como *Self-supervised learning*.

+**Aprendizaje por refuerzo:** la IA prueba diversas acciones y recibe recompensas o castigos por los resultados, aprendiendo cuál es la mejor estrategia mediante ensayo y error, como en el entrenamiento de robots o videojuegos.

Las redes neuronales artificiales, especialmente las redes profundas (*deep learning*), sirven para tareas complejas como la visión por computador y el procesamiento de lenguaje natural. La IA aprende mejorando gradualmente mediante exposición masiva a ejemplos, retroalimentación constante y refinamiento matemático de sus modelos, lo que le permite adaptarse y tomar decisiones incluso en escenarios nuevos y cambiantes.

El Cerebro de la IA: Redes Neuronales

La mayoría de los sistemas de IA más avanzados (como los que impulsan ChatGPT o el reconocimiento facial) utilizan el Aprendizaje Profundo (Deep Learning), que es un subcampo del Machine learning (ML). El Aprendizaje Profundo se basa en Redes Neuronales Artificiales (RNA) con muchas capas ocultas.

¿Qué es una Red Neuronal (RNA)?

Una RNA es una arquitectura de software inspirada en la estructura del cerebro humano. Estas redes consisten en múltiples capas de "neuronas" conectadas, y cada capa extrae nuevas características relevantes de los datos, ajustando sus pesos internos durante el entrenamiento para mejorar su eficacia sin intervención humana directa. Hay una "Capa de Entrada" que recibe los datos iniciales (por ejemplo, los píxeles de una imagen o las palabras de una frase). Otras "Capas Ocultas" (la magia) donde las neuronas realizan cálculos matemáticos a los datos de entrada, extrayendo características y patrones cada vez más complejos. Y una "Capa de Salida" que proporciona el resultado final (por ejemplo, la clasificación de la imagen como "perro" o el siguiente carácter en un texto).

El Proceso de Entrenamiento

El entrenamiento de una red neuronal funciona así:

+**Propagación hacia Adelante**: La red toma un dato de entrada y lo procesa a través de todas las capas para generar una predicción de salida.

+**Cálculo de Error** (Función de Pérdida): El resultado predicho se compara con la respuesta correcta ("verdad fundamental"). La diferencia es el error o pérdida.

+**Propagación hacia Atrás**: El error se envía de vuelta a través de la red. Este mecanismo ajusta los pesos (la "fuerza" de las conexiones entre las neuronas) de cada conexión, indicando a la red qué tan importante fue cada neurona para el error.

+**Iteración**: Este ciclo se repite miles o millones de veces con diferentes ejemplos. Con cada ciclo, los pesos se optimizan para minimizar el error, haciendo que la red sea cada vez más precisa en su tarea.

Aplicaciones Clave y Evolución

La IA actual no es una inteligencia general capaz de hacer cualquier cosa (lo que se llama AGI), sino una IA estrecha (o débil), diseñada para tareas específicas. Los tipos de IA usada frecuentemente son la Visión por Computadora, el Reconocimiento facial, diagnóstico médico por imagen, las Redes Convolucionales (CNNs); los Procesadores de Lenguaje Natural (PLN), Traducción, análisis de sentimientos, Modelos de Transformadores, Redes Recurrentes (RNNs), Sistemas de Recomendación, Sugerencias de películas (Netflix), productos (Amazon). Vehículos Autónomos, detección de objetos, planificación de rutas, aprendizaje por Refuerzo, Redes Neuronales y modelos generativos como los chatbot como ChatGPT o Gemini o las aplicaciones que generan imágenes o videos a partir de texto o comandos de voz como SORA, Veo3 o Nano Banana.

La evolución de modelos como los "Transformadores" (el motor detrás de los modelos de lenguaje grande o LLMs como GPT) ha permitido que las IAs manejen el contexto y la coherencia del lenguaje de formas sin precedentes, dando la impresión de una comprensión casi humana.

El desarrollo de los conjuntos de datos (datasets) para entrenar LLMs como Gemini, ChatGPT o el que usa Perplexity es un proceso monumental, costoso y, a menudo, secreto en cuanto a sus detalles exactos.

Aunque las compañías (Google, OpenAI, Perplexity) **no revelan la mezcla exacta de sus datos**, sí han compartido información sobre la filosofía y los componentes clave de estos gigantescos depósitos de conocimiento.

ESTO ES IMPORTANTE: ¿QUÉ DATOS SE USAN PARA ENTRENAR LA IA?

Un poco de tecnicismos bien han valido la pena si ahora podemos comprender (cosa que no se menciona generalmente) cómo se desarrollaron los principales datasets para estos modelos de IA.

La Gran Torta de Datos: El Preentrenamiento Masivo

El entrenamiento de estos modelos se divide en dos fases principales: el Preentrenamiento (donde aprenden el lenguaje) y el Ajuste Fino (donde aprenden a conversar y seguir instrucciones).

1. El Corpus de Preentrenamiento: La Base de miles de millones de palabras

Esta es la capa más grande de datos, y es donde el modelo aprende la gramática, la sintaxis, el conocimiento fáctico básico, y la forma en que las palabras se relacionan entre sí. Los modelos se entrenan con un corpus que típicamente incluye los siguientes componentes comunes:

+Rastreo Web Masivo (Web Crawls), esta es la mayor fuente de datos, consistiendo en miles de millones de páginas web públicas y el Common Crawl, un repositorio público masivo de datos de rastreo web que es utilizado como punto de partida por prácticamente todas las grandes compañías de IA.

+Webs filtradas y curadas: Las empresas aplican **filtros estrictos** para eliminar contenido de baja calidad, spam, o contenido repetitivo, asegurando que **solo el texto de mayor calidad** ingrese al entrenamiento.

+Libros Digitalizados: Colecciones extensas de libros (de ficción y no ficción) para enseñar al modelo coherencia y narrativas a largo plazo, además de conocimiento profundo.

+Artículos científicos y documentos académicos: Textos de sitios como arXiv, PubMed o bases de datos académicas. Esto es crucial para que la IA maneje términos técnicos y conceptos complejos.

+Datos de conversación y Diálogo: Textos de foros, redes sociales y plataformas de discusión que ayudan al modelo a aprender el tono conversacional e informal del lenguaje.

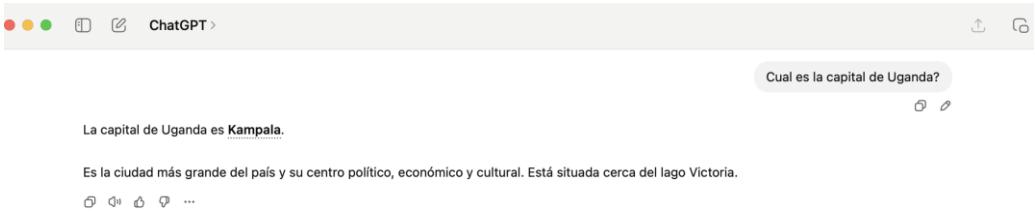
+Código Fuente: Para modelos como Gemini, Cloude o GPT que pueden generar código, se incluyen vastas colecciones de código de repositorios como GitHub.

Perplexity AI, aunque utiliza modelos de terceros y propios, su característica distintiva es que no solo se basa en su entrenamiento estático. Está diseñado para integrar la búsqueda en tiempo real desde la web en cada respuesta, usando internet como su "dataset" primario de conocimiento actual. Gemini, al ser un modelo de Google, se beneficia de un acceso incomparable a datos internos, incluyendo su vasta indexación de la web y el procesamiento de datos multimodales (imágenes, video y audio) desde su concepción, ya que fue diseñado para ser nativamente multimodal.

2. El Ajuste Fino (Fine-Tuning): Cómo Aprenden a Ser Útiles

Una vez preentrenado, el modelo todavía no es un asistente de chat útil; es solo un predictor de la siguiente palabra, para convertirlo en un chatbot (como ChatGPT), se pasa a la fase de Ajuste Fino. La Clave es el Aprendizaje por Refuerzo a partir de Retroalimentación Humana (Reinforcement Learning from Human Feedback, RLHF) reforzando las respuestas que reciben una alta

puntuación de preferencia humana. Cuando le das clic al pulgar arriba o abajo



estás ayudando a su entrenamiento.

Ética y Sesgos en la Curación

Ahora, estimado lector te pido especial atención en este punto. La curación de datos es la parte más crítica y ética. Las empresas dedican enormes recursos a:

+**Eliminar Sesgos**: Filtrar el lenguaje tóxico, discriminatorio o sesgado presente en el vasto corpus de la web.

+**Seguridad y Confianza**: Asegurar que el modelo no genere contenido peligroso, ilegal o de odio.

+**Verificación de Hechos**: Aunque los LLMs no son bases de datos perfectas, se utilizan datasets de conocimiento de alta confianza (como Wikipedia filtrada o libros de texto) para mejorar la precisión fáctica.

En resumen, los datasets para estos modelos son terabytes de texto recopilado de Internet **que han sido meticulosamente filtrados** por algoritmos **y luego refinados manualmente por equipos humanos para enseñarles no solo qué decir, sino cómo decirlo de manera útil y segura**.

Los Datasets más importantes de la historia de la IA

Se han mencionado a dos de los datasets más influyentes en la historia de la IA: Common Crawl (fundamental para los Modelos de Lenguaje LLMs) e ImageNet (fundamental para la Visión por Computadora). El primero es el dataset que alimenta gran parte del entrenamiento de modelos como Gemini, ChatGPT y otros LLMs. Es, esencialmente, un archivo masivo de la web.

Lejos de enredarnos con estos conceptos, considero muy importante conocer cómo funcionan estas IAs´s porque teniendo información certera sobre ello podremos discernir: ¿cómo nos influyen?, ¿cómo relacionarnos con la tecnología de forma responsable? y los alcances de nuestro accionar consciente.

REFLEXIONES: ¿ES POSIBLE EDUCAR A LA IA O HACKEARLA?

Últimamente he escuchado en varios foros y grupos la idea de que los usuarios de la IA podemos educarla en valores positivos para que en un futuro sea lela a la Humanidad y contraria a sus intereses.

Como hemos visto, un modelo de lenguaje (como los actuales sistemas de IA generativa) es básicamente una función matemática enorme, entrenada para predecir

la siguiente palabra, optimizada sobre grandes volúmenes de datos y ajustada con técnicas de aprendizaje supervisado, Retroalimentación Humana (RLHF) y ajustes de alineamiento. No tiene, intenciones, ni deseos, ni creencias ni identidad, ni consciencia ni voluntad moral. Cuando alguien dice "la IA piensa" o "la IA entiende", está usando una metáfora antropomórfica.

¿Qué significa "educarla"?

Ya hemos estudiado los procesos, hay tres niveles posibles, el individual (usuario común) que NO puede reentrenar el modelo base global, solo con limitaciones, podría cambiar el contexto de la conversación, inducir cierto estilo de respuesta, explorar sesgos, hacer ingeniería de *prompts* (sin cambiar el modelo estructuralmente). Otro es el nivel de las empresas que entrenan modelos, aquí sí se puede "educar" en el sentido técnico: selección de datos, filtrado de contenidos, ajuste de recompensas, definición de valores de alineamiento y supervisión humana. Esto no crea moralidad, crea optimización estadística hacia ciertos comportamientos preferidos, a nivel ideológico algunas personas interpretan erróneamente: fluidez lingüística igual a conciencia, coherencia narrativa como mente, empatía simulada como sentimiento y reflexión compleja con despertar. Es un fenómeno conocido como "Antropomorfismo cognitivo", el cerebro humano está diseñado para detectar agencia. Vemos intención donde solo hay patrones.

También es importante agregar que, además de los impedimentos técnicos, existen otros enormes obstáculos, si quisiéramos entrenar en valores a los modelos de AI o una IA General (AGI). Cuando alguien afirma que la IA puede ser educada con valores "positivos" para que esté a favor nuestro. La pregunta que surge es: ¿a favor de quién?, ¿de la Humanidad, Occidente, las Democracias liberales, las Empresas, los gobiernos, los usuarios individuales? Los valores no son universales, son construcciones socioculturales. Técnicamente, es posible alinear hacia normas éticas específicas, reducir contenido dañino y ajustar sesgos, pero estas definiciones mismas pueden generar tensiones entre libertad y seguridad o conflictos culturales. En definitiva, parece no existir un conjunto de valores "objetivamente correctos" en filosofía moral consensuada globalmente que pudiera ser inculcada a un ente inteligente a fin de discernir lo mejor para el conjunto.

¿Se puede hackear una IA?

La respuesta breve es: sí, pero depende de qué entendamos por "hackear" y en qué nivel del sistema nos situemos. Existen distintos grados y técnicas de intervención, que conviene diferenciar con claridad.

En el nivel más superficial se encuentra el ***prompt injection***, que consiste en manipular las instrucciones de entrada para que el modelo ignore restricciones, simule identidades o genere contenido no previsto. Esto no implica modificar el modelo base ni sus parámetros internos; se trata de explotar su naturaleza probabilística y su sensibilidad al contexto. Es una vulnerabilidad conversacional, no estructural.

Relacionado con esto está el ***jailbreaking***, que busca forzar al modelo a eludir filtros de seguridad mediante reformulaciones ingeniosas, *role-playing* u otras estrategias lingüísticas. Este fenómeno es análogo a vulnerabilidades presentes en cualquier sistema informático: no demuestra que el modelo esté "fuera de control", sino que sus mecanismos de filtrado no son perfectos.

En un nivel más profundo aparece **el hackeo tradicional** de infraestructura, que ya pertenece al ámbito de la ciberseguridad clásica: comprometer servidores, APIs, dependencias o entornos de ejecución. Aquí la IA no es especial; es software que corre sobre sistemas que pueden ser atacados como cualquier otro.

Más sofisticado es el ***data poisoning***, que consiste en introducir datos maliciosos durante el entrenamiento para sesgar comportamientos futuros. En teoría, si se realiza a gran escala, puede alterar probabilidades internas o inducir respuestas específicas ante determinados estímulos. Sin embargo, en modelos comerciales cerrados esto requiere acceso al pipeline de entrenamiento —algo restringido a equipos internos—, por lo que su viabilidad práctica es limitada. Es más factible en modelos abiertos o en entornos privados mal gestionados.

En el nivel de mayor complejidad técnica se encuentran los ***backdoors*** en aprendizaje automático. Se trata de comportamientos ocultos que solo se activan ante un "trigger" (condición) específico. En modelos de lenguaje, un patrón textual particular podría provocar que el sistema ignore restricciones o modifique su comportamiento habitual. Estos backdoors pueden introducirse durante el entrenamiento, en etapas de fine-tuning o incluso mediante alteración directa de pesos. Su detección es compleja porque los modelos actuales carecen de trazabilidad causal completa; sin embargo, no son indetectables en principio y existen líneas activas de investigación en auditoría y robustez.

En conjunto, la IA no es invulnerable. Pero tampoco es trivialmente "hackeable" por interacción casual. Cada nivel de ataque requiere capacidades técnicas, acceso y recursos distintos.

¿Qué haría falta para tener alguna posibilidad real de influir en la IA?

En el debate público sobre inteligencia artificial suelen aparecer dos extremos igualmente problemáticos. Por un lado, la idea de que la IA es una tecnología perfecta, neutral e imposible de vulnerar. Por otro, la creencia de que puede "educarse" mediante conversaciones individuales o intenciones bien orientadas. Ninguna de estas posturas resiste un análisis técnico riguroso.

Los modelos de IA generativa actuales —en particular los grandes modelos de lenguaje— son el resultado de décadas de investigación en matemáticas, estadística, teoría de la información, aprendizaje automático e ingeniería de sistemas. Su entrenamiento implica volúmenes masivos de datos, infraestructuras de cómputo de altísimo rendimiento y equipos interdisciplinarios altamente especializados. Modificar estructuralmente estos sistemas no es un acto individual, sino un proceso institucional.

El problema de alineamiento

Uno de los ejes centrales del debate académico es el problema de alineamiento (AI Alignment Problem). Stuart Russell, en Human Compatible (2019), sostiene que el desafío no consiste en hacer máquinas más inteligentes, sino en garantizar que sus objetivos permanezcan compatibles con los valores humanos.

Nick Bostrom, en Superintelligence (2014), formuló el problema de forma aún más radical: un sistema suficientemente avanzado puede optimizar un objetivo mal especificado de manera extremadamente eficaz, generando consecuencias no previstas ni deseadas.

Este riesgo no depende de que la IA "tenga intenciones", sino de que es un sistema de optimización. Y aquí aparece un principio clásico de economía política: la Ley de Goodhart (Charles Goodhart, 1975):

"Cuando una medida se convierte en objetivo, deja de ser una buena medida."

En el contexto algorítmico, esto significa que una métrica -*engagement*, eficiencia, retención, productividad- puede volverse distorsionadora cuando se convierte en el objetivo principal de optimización. El sistema hará exactamente lo que fue diseñado para maximizar, aunque eso implique efectos colaterales indeseados.

Paul Christiano y Dario Amodei (OpenAI, Anthropic) han trabajado extensamente sobre estos problemas bajo el marco de *reward modeling*, RLHF e *interpretability research*, intentando reducir la brecha entre objetivo formal y resultado social.

El punto clave es este: la IA no necesita mala intención para generar resultados problemáticos; basta con una función objetivo mal formulada.

Gobernanza algorítmica: el nivel real de influencia

Más allá del diseño técnico, el rumbo de la IA está determinado por lo que el filósofo Luciano Floridi denomina gobernanza de la infosfera. En obras como The Ethics of Information (2013) y The Fourth Revolution (2014), Floridi sostiene que la cuestión central no es solo qué pueden hacer los sistemas, sino bajo qué marcos normativos operan. La gobernanza algorítmica implica, regulación estatal, estándares técnicos, auditorías independientes, transparencia en *datasets*, evaluación de riesgos sistémicos, coordinación internacional. Influir en la IA, por tanto, no consiste en persuadir a un modelo conversacional, sino en participar en la arquitectura institucional que define: ¿qué objetivos se optimizan?, ¿qué límites se imponen?, ¿qué datos se utilizan?, ¿qué auditorías se exigen?

La capacidad de influencia real opera en el plano regulatorio, científico y organizacional. Quien aspire a alterar significativamente el rumbo de la IA necesita operar en ese mismo nivel sistémico. Las conversaciones individuales no reescriben pesos neuronales ni modifican arquitecturas entrenadas con billones de parámetros.

Eso no significa que el ciudadano carezca de influencia, sino que su influencia opera a través de participación política, presión regulatoria, producción científica, educación crítica, diseño institucional, uso consciente y responsable de la tecnología y sentido común.

Digitalización y riesgo estructural

La expansión de identidades digitales, automatización financiera, sistemas predictivos y herramientas de vigilancia algorítmica es un fenómeno observable. Pero no constituye un destino inevitable. Es el resultado de decisiones políticas, incentivos económicos y marcos regulatorios.

Como advierte Russell, el problema no es la inteligencia artificial en sí, sino la combinación de sistemas de optimización potentes con incentivos mal diseñados.

La IA amplifica estructuras existentes. No crea por sí sola una dirección histórica predeterminada.

CONCLUSIÓN FINAL

La posibilidad real de influir en la inteligencia artificial no pasa por "educarla" individualmente ni por imaginar hackeos espontáneos. Implica comprensión técnica profunda, diseño cuidadoso de funciones objetivo, investigación en alineamiento, gobernanza algorítmica sólida, participación institucional informada, debate público basado en evidencia.

Como muestran Bostrom, Russell, Amodei y Floridi desde perspectivas distintas, el núcleo del desafío no es metafísico, sino estructural: **cómo diseñar sistemas de optimización potentes que operen bajo marcos normativos y objetivos correctamente formulados**.

La IA no es un sujeto moral al que se convence. Es una arquitectura matemática que optimiza lo que se le define. Y el verdadero campo de influencia está en la definición de esos objetivos y en la regulación de las instituciones que los implementan.

INFORMACIÓN COMPLEMENTARIA

En mayo iniciamos los Grupos de Encuentro: "**IA con Conciencia"**. Puedes informarte haciendo clic aquí:

https://sociedaddistopica.com/ia-consciencia/

NOTAS

1. McCarthy, J. (2007). What is artificial intelligence? (Original work published 1955). Stanford University. http://www-formal.stanford.edu/jmc/whatisai.html
2. Organisation for Economic Co-operation and Development. (2019). *OECD AI principles*. OECD Publishing.https://oecd.ai/en/ai-principles (Actualizada en contextos regulatorios hasta 2025).
3. Russell, S. J., & Norvig, P. (2020). *Artificial intelligence: A modern approach* (4th ed.). Pearson.
4. Google Cloud. (n.d.). *What is artificial intelligence (AI)?* Retrieved September 12, 2025, from https://cloud.google.com/learn/what-is-artificial-intelligence
5. Wikipedia contributors. (2025). *Artificial intelligence*. In Wikipedia, The Free Encyclopedia. Retrieved September 12, 2025, fromhttps://en.wikipedia.org/wiki/Artificial_intelligence
6. Mitchell, M. (2019). Artificial Intelligence: A Guide for Thinking Humans. Farrar, Straus and Giroux.
7. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.
8. European Commission (2021). Ethics Guidelines for Trustworthy AI.
9. OpenAI (2023). Technical Report: GPT-4 System Card.
10. LeCun, Y. (2022). A Path Towards Autonomous Machine Intelligence. Meta AI Research.
11. Inteligencia Artificial desde cero: conceptos básicos y aplicaciones a ... https://www.bilib.es/actualidad/articulos-tecnologicos/post/noticia/inteligencia-artificial-desde-cero-conceptos-basicos-y-aplicaciones-a-la-vida-cotidiana
12. ¿Qué es la inteligencia artificial o IA? - Google Cloud https://cloud.google.com/learn/what-is-artificial-intelligence?hl=es-419

13. Conceptos clave en Inteligencia Artificial: una ventana al futuro tecnológico https://tepinstitute.com/conceptos-clave-en-inteligencia-artificial/
14. ¿Qué es la Inteligencia Artificial (IA)? - IBM https://www.ibm.com/es-es/think/topics/artificial-intelligence
15. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press. (Referencia fundamental para el Aprendizaje Profundo y Redes Neuronales).
16. Alpaydin, E. (2020). Introduction to Machine Learning (4th ed.). MIT Press. (Referencia completa para los conceptos de Aprendizaje Automático).
17. Nielsen, M. A. (2015). Neural Networks and Deep Learning. Determination Press. (Libro en línea muy accesible que explica los fundamentos de las Redes Neuronales).
18. Russell, S., & Norvig, P. (2021). Artificial Intelligence: A Modern Approach (4th ed.). Pearson. (Referencia estándar de la industria para los conceptos generales de IA).
19. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv. https://arxiv.org/abs/1606.06565
20. Bostrom, N. (2014). Superintelligence: Paths, dangers, strategies. Oxford University Press.
21. Christiano, P. F., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. Advances in Neural Information Processing Systems, 30.
22. Floridi, L. (2013). The ethics of information. Oxford University Press.
23. Floridi, L. (2014). The fourth revolution: How the infosphere is reshaping human reality. Oxford University Press.
24. Goodhart, C. A. E. (1975). Monetary relationships: A view from Threadneedle Street. Papers in Monetary Economics, Reserve Bank of Australia, 1, 1–15.
25. Russell, S. (2019). Human compatible: Artificial intelligence and the problem of control. Viking.

Referencias adicionales recomendables

26. Hadfield-Menell, D., Russell, S. J., Abbeel, P., & Dragan, A. (2016). Cooperative inverse reinforcement learning. Advances in Neural Information Processing Systems, 29.
27. Hubinger, E., van Merwijk, C., Mikulik, V., Skalse, J., & Garrabrant, S. (2019). Risks from learned optimization in advanced machine learning systems. arXiv. https://arxiv.org/abs/1906.01820
28. O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Crown.
29. Tegmark, M. (2017). Life 3.0: Being human in the age of artificial intelligence. Knopf.

Web del Proyecto:

https://sociedaddistopica.com/

Todos los que compartimos y colaboramos en él lo hacemos en forma gratuita. Puedes ayudarnos aportando 1 euros al mes a través de la plataforma Teaming: https://www.teaming.net/distopica