Tabla de Contenidos

Episodio 57: ¿Hay que desarrollar GPT-5 o hay que parar?

Empieza a haber reacciones a GPT-4, el modelo de lenguage recientemente lanzado por OpenAI a través del interfaz ChatGPT.

El “Future of Life Institute” ha coseguido ya casi 3600 firmas firmas de su escrito “Pausar experimentos gigantes de IA: una carta abierta. Hacemos un llamado a todos los laboratorios de IA para que pausen inmediatamente durante al menos 6 meses el entrenamiento de los sistemas de IA más potentes que GPT-4.”

Leemos la carta abierta y exploramos su propuesta.

“La filosofía no sirve para nada” es un podcast sin pretensiones en el que reflexionaremos sobre el presente.

Participan: José Carlos García @quobit, Joaquín Herrero @joakinen@scholar.social, Juan Carlos Barajas @SociologiaDiver, Juan Antonio Torrero @jatorrero, Sergio Muñoz @smunozroncero

Fecha	3 de abril de 2023
Participan	José Carlos García @quobit Sergio Muñoz @smunozroncero Juan Carlos Barajas @SociologiaDiver Joaquín Herrero @joakinen@scholar.social Juan Antonio Torrero @jatorrero
Descarga	Puedes descargar todos los episodios en iVoox, en Spotify, en iTunes, Google Podcasts y en nuestro canal de Telegram. Si tienes un lector de podcasts que admite enlaces RSS, este es el enlace RSS a nuestro podcast.
Sintonía	Mass Invasion, Dilo, álbum **Robots** (2004)
Fotos	Foto de u/surfiiin
Intro	10.000 días: CARLOS FRANGANILLO y cómo MARCARÁ la INTELIGENCIA ARTIFICIAL nuestro FUTURO | RTVE

Mastodon	En @FilosofiaNada@hcommons.social publicamos noticias que nos interesan y conversamos.
Twitter	Estuvimos, pero ya casi no estamos. @FilosofiaNada
Canal Telegram	Puedes seguir la preparación de nuevos episodios suscribiéndote al canal @FilosofiaNada en Telegram
Grupo de opinión	Únete a nuestro grupo de opinión Opina FilosofiaNada para opinar sobre el episodio en preparación y enviarnos audios con preguntas o críticas con humor para nuestra intro

Carta abierta del Future of Life Institute

https://futureoflife.org/open-letter/pause-giant-ai-experiments/

La carta abierta firmada por Musk, Marcus y muchos otros contiene la siguiente frase:

> OpenAI's recent statement regarding artificial general intelligence, states that "At some point, it may be important to get independent review before starting to train future systems, and for the most advanced efforts to agree to limit the rate of growth of compute used for creating new models." We agree. That point is now.

https://futureoflife.org/open-letter/pause-giant-ai-experiments/

Es decir, antes de que ellos dijeran que hay que elegir el punto en el que parar y reflexionar ya lo dijo OpenAI.

El tema es que si se mantiene en secreto la potencia de los Transformers entonces la gente no estaría informada del alcance de esta tecnología. La única forma en la que todo el mundo sepa a qué nos enfrentamos es darle la máxima publicidad a la potencia de esta tecnología. Exactamente lo que ha hecho OpenAI creando un interface de chat y abriendo un API.

El lugar donde OpenAI propuso eso es en esta página:
https://openai.com/blog/planning-for-agi-and-beyond

Incluía ahí a poderes públicos y gobiernos.

- “We think public standards about when an AGI effort should stop a training run, decide a model is safe to release, or pull a model from production use are important. Finally, we think it’s important that major world governments have insight about training runs above a certain scale.”
- “There should be great scrutiny of all efforts attempting to build AGI and public consultation for major decisions.”
- “A misaligned superintelligent AGI could cause grievous harm to the world; an autocratic regime with a decisive superintelligence lead could do that too.”

La posición de Musk reconoce sin quererlo que la IA es más inteligente que el ser humano en al menos un aspecto: en la creación de mensajes. ¿Qué impide a nadie crear mensajes de calidad para desinformar ? ¿Significa esto que la IA es mejor que el ser humano en algo que ha sido asociado inequívocamente al ser humano?

Musk no pidió permiso para desarrollar Starlink, que se carga observaciones espaciales o Neuralink, que me parece mucho más peligroso

El verdadero problema es conseguir la alineación

Sobre el tema del “misalignment” pusimos una explicación en el cuaderno de notas del episodio 56

Pendiente de resolver: el “alignment Problem” (Redactado por GPT-4)

El “alignment problem” (problema de alineación) en el contexto de ChatGPT y otros modelos de lenguaje de inteligencia artificial se refiere al desafío de garantizar que los objetivos, valores e intereses del sistema de IA estén alineados con los de los seres humanos. En otras palabras, se busca que la IA entienda, interprete y responda a las necesidades y expectativas de los usuarios de manera efectiva, ética y segura.

Los modelos de lenguaje como ChatGPT son el resultado del entrenamiento en grandes conjuntos de datos de texto que contienen información de diversas fuentes y perspectivas. Como resultado, pueden adoptar sesgos, creencias erróneas y comportamientos no deseados que no siempre están en consonancia con las intenciones del usuario o el bienestar general.

El problema de alineación en IA plantea varios desafíos:

- Sesgo y justicia: Asegurar que el modelo no tenga sesgos sistemáticos o discriminación en sus respuestas.
- Seguridad: Prevenir que el modelo proporcione respuestas dañinas, ofensivas o inapropiadas.
- Privacidad: Evitar que el modelo divulgue información confidencial o sensible.
- Robustez y fiabilidad: Garantizar que el modelo responda de manera consistente y confiable a las entradas del usuario, incluso en casos de entradas ambiguas o maliciosas.
- Control de contenido y políticas: Establecer pautas claras y efectivas para el comportamiento del modelo y garantizar que se adhiera a ellas.

Resolver el problema de alineación es crucial para el desarrollo y adopción segura y ética de sistemas de inteligencia artificial como ChatGPT en diferentes aplicaciones y contextos.

Opinión de Eliezer Yudkowsky sobre el problema de la alineación:
https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/

OpenAI dice cosas muy importantes y sin aspavientos ni grandilocuencias.

- “Successfully transitioning to a world with superintelligence is perhaps the most important—and hopeful, and scary—project in human history. Success is far from guaranteed, and the stakes (boundless downside and boundless upside) will hopefully unite all of us.
- We can imagine a world in which humanity flourishes to a degree that is probably impossible for any of us to fully visualize yet. We hope to contribute to the world an AGI aligned with such flourishing.”

Ellos ven que el tema sobre el que hay que pensar es el del alineamiento de la futura AGI con los intereses humanos/planetarios/etc

Los que deberían darse por aludidos por lo que propone y fabrica OpenAI (gobiernos) están a otra cosa y eso permite que oportunistas como Future of Life Institute entren en el debate de manera

bastante histriónica pervirtiendo el tono que el debate debería de tener

La moratoria que se pide necesita mecanismos de debate entre empresas tecnológicas y sociedad

Suponiendo que los firmantes lleven razón y la mejor solución posible sea parar los desarrollos de GPT-5 aún nos enfrentamos a un problema más. Lo que se pide no se suele aplicar en ningún escenario de innovación tecnológica por lo que no tenemos habilitados los mecanismos por los que:

- la sociedad esté completamente informada de la tecnología que se cuestiona y sus consecuencias
- la sociedad conozca alternativas a lo que se propone
- la industria tecnológica sea capaz de alinearse con lo que "la sociedad" reclama

Ya tratamos en el episodio 23 (Episodio 23: Expertos, epidemias y sociedad. De la mano de Ulrich Beck.) lo complejo que es establecer un canal de retorno para que los científicos o los fabricantes de tecnología reciban el feedback de la población a la que van destinadas sus creaciones.

En el trabajo La creación colectiva de conocimiento en Internet: posibilidades y dificultades ya tratamos este tema indicando que a pesar de la relevancia actual del conocimiento experto en las decisiones públicas, **insistimos en la necesidad de que exista un mayor grado de participación ciudadana en los procesos de decisión política y científica**. Esto implica superar un tipo de divulgación científica que se limite a publicitar los logros de la investigación para pasar a la creación de un espacio público de deliberación de naturaleza híbrida en el que participen tanto expertos como ciudadanos. En este trabajo se analizan las posibilidades de Internet como soporte para la creación de ese espacio, cuáles serían las cualidades deseables de la participación ciudadana y si los problemas que suelen darse en los espacios de deliberación podrían ser superados con las herramientas tecnológicas actuales:

Se ha dicho que "el tema esencial de la nueva sinfonía social es el puesto del conocimiento experto en el espacio de la política y el orden de la sociedad" reconociendo así que desde hace décadas la sociedad reclama hacer oír su voz en diversos ámbitos de la vida social y política y creando de esa forma la problemática cuestión de qué peso debe de tener su voz en comparación con la de los expertos.

El filósofo austríaco Paul Feyerabend definió a 'los expertos' como "un grupo de personas que por su entrenamiento son capaces de elegir alternativas que implicarían grandes beneficios para todos", por lo cual "nos inclinaríamos a pagarles y a dejarles actuar sin más control". En comparación con el conocimiento experto, el conocimiento del que dispone la ciudadanía suele ser calificado, tal como dijo Michel Foucault, como "saberes descalificados, como saberes no conceptuales, como saberes insuficientemente elaborados: saberes ingenuos, saberes jerárquicamente inferiores, saberes por debajo del nivel del conocimiento o la cientificidad exigidos".

Cabe pensar, por tanto, en lo razonable que sería hacer recaer la responsabilidad de cualquier decisión especializada, por ejemplo sobre asuntos científicos, en el criterio de los expertos, pero eso crea al menos dos conjuntos de problemas.

Por una parte, como indica Sven Ove Hanson, acudir a expertos no elimina la incertidumbre en la decisión, pues la "incertidumbre de fiabilidad" se produce precisamente cuando acudimos a un supuesto experto en el tema e introduce los problemas de la valoración de alguien como experto y las

opiniones divergentes entre los considerados expertos.

Por otro lado, Feyerabend apunta al hecho de que no sería democrático delegar las decisiones en expertos, ya que en una democracia "la elección de programas de investigación en todas las ciencias es una tarea en la que deben poder participar todos los ciudadanos" por lo que ve conveniente "la no división entre expertos y legos en las cuestiones fundamentales de evaluación de un programa de investigación".

El dilema entre conocimiento experto y necesidad de participación democrática ha sido calificado por Fernando Broncano como una especie de "juego del prisionero epistémico" en el que se necesita la colaboración de varios para un fin común pero cada uno opina que lo racional es su punto de vista, por lo que no se crea una colaboración sino una rivalidad que puede desembocar en una nueva tragedia de los comunes de naturaleza epistémica. Por ello, según Broncano, el problema es de "medios de ordenamiento de voluntades en un terreno informacional, pues en cierto momento la dimensión computacional del ágora se convierte en una limitación técnica al propio ejercicio de la democracia".

Centrado ya el problema en uno de ordenamiento de cierta dimensión computacional de un ágora público de debate, el incentivo para avanzar en su resolución procede de al menos dos declaraciones institucionales que, en la misma dirección que apuntaba Feyerabend, reclaman un papel activo para los ciudadanos en los mecanismos de toma de decisión institucionales.

La primera es la "Declaración de Río sobre el Medio Ambiente y el Desarrollo" hecha en junio de 1992 por Naciones Unidas, y que en su principio 10 indica que "todos los individuos deben de tener acceso adecuado a la información medioambiental" para "participar en procesos de toma de decisión".

La segunda, la "Declaración de Santo Domingo", realizada en la IV Conferencia Iberoamericana de Ministros de Administración Pública y Reforma del Estado de 2002, indica que las comunidades de investigadores deben "contribuir, especialmente en el caso de problemas en los que están involucradas, a la presentación de alternativas sobre las cuales la ciudadanía pueda informarse y pronunciarse; tener en cuenta las opiniones de la sociedad y dialogar efectivamente con ella; y luchar contra el entronizamiento de tecnocracias amparadas en conocimientos científicos y tecnológicos, reales o supuestos".

¿Qué mecanismos tenemos a nuestro disposición para poner en marcha este diálogo reclamado entre el saber experto y el 'saber de la gente', como lo llamaba Foucault? ¿Es hoy día Internet un ágora tal que su dimensión computacional no se constituya en una limitación para el ejercicio de la democracia, como advierte Broncano? En este trabajo nos proponemos examinar el estado actual de Internet como lugar público donde construir deliberación y conocimiento de forma colectiva, sus limitaciones y algunas experiencias que muestran el camino a seguir.

¿Cómo hacer llegar a OpenAI y el resto de actores la crítica de la ciudadanía sobre su producto?

Seguimos necesitando los mecanismos de participación cuidadana que permitan cerrar el bucle particiación-formación.

CLASE DEL 7 DE ENERO DE 1976 21

me parece que debajo de ella, a través de ella, en ella misma, vimos producirse lo que podríamos llamar la insurrección de los *saberes sometidos*. Y por *saber sometido* entiendo dos cosas. Por una parte, quiero designar, en suma, contenidos históricos que fueron sepultados, enmascarados en coherencias funcionales o sistematizaciones formales. Concretamente, si quieren, lo que permitió hacer la crítica efectiva tanto del asilo como de la prisión no fue, por cierto, una semiología de la vida asilar ni tampoco una sociología de la delincuencia, sino, en verdad, la aparición de contenidos históricos. Y simplemente porque sólo los contenidos históricos pueden permitir recuperar el clivaje de los enfrentamientos y las luchas que los ordenamientos funcionales o las organizaciones sistemáticas tienen por meta, justamente, enmascarar. De modo que los *saberes sometidos* son esos bloques de saberes históricos que estaban presentes y enmascarados dentro de los conjuntos funcionales y sistemáticos, y que la crítica pudo hacer reaparecer por medio, desde luego, de la erudición.

En segundo lugar, por *saberes sometidos* creo que hay que entender otra cosa y, en cierto sentido, una cosa muy distinta. Con esa expresión me refiero, igualmente, a toda una serie de saberes que estaban descalificados como saberes no conceptuales, como saberes insuficientemente elaborados: saberes ingenuos, saberes jerárquicamente inferiores, saberes por debajo del nivel del conocimiento o de la cientificidad exigidos. Y por la reaparición de esos saberes de abajo, de esos saberes no calificados y hasta descalificados: el del psiquiatrizado, el del enfermo, el del enfermero, el del médico –pero paralelo y marginal con respecto al saber médico–, el saber del delincuente, etcétera –ese saber que yo llamaría, si lo prefieren, *saber de la gente* (y que no es en absoluto un saber común, un buen sentido sino, al contrario, un saber particular, un saber local, regional, un saber diferencial, incapaz de unanimidad y que sólo debe su fuerza al filo que opone a todos los que lo rodean)–, por la reaparición de esos saberes locales de la gente, de esos saberes descalificados, se hace la crítica.

Bucle de la Participación Formativa

López Cerezo, J.A. (2005) "Participación ciudadana y cultura científica" Arbor Vol. 181, No 715 (2005)

Informarse. Participar.

Foucault, 1976
- "insurrección de los saberes sometidos"
- reivindicación del "saber de la gente"

Naciones Unidas, 1976
Declaración de Río
- Tratar las cuestiones ambientales con la participación de todos los ciudadanos interesados.
- Toda persona deberá tener acceso adecuado a la información
- Los estados deberán facilitar y fomentar la sensibilización y la participación de la población

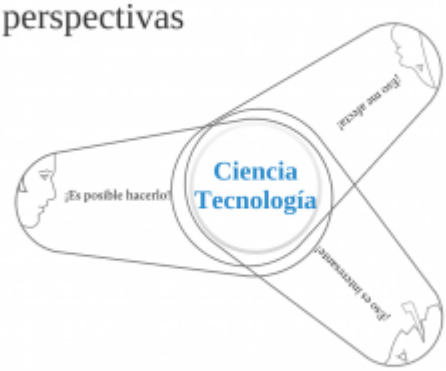
Unesco, 1999
Declaración de Santo Domingo
- presentación de alternativas sobre las cuales la ciudadanía pueda informarse y pronunciarse
- tener en cuenta las opiniones de la sociedad y dialogar efectivamente con ella
- desarrollar la educación científica y tecnológica de los ciudadanos

Estudios CTS (Ciencia, Tecnología y Sociedad)
- Ciencia y Tecnología se construyen socialmente

Tesis de la Web 2.0
- Construcción social del conocimiento

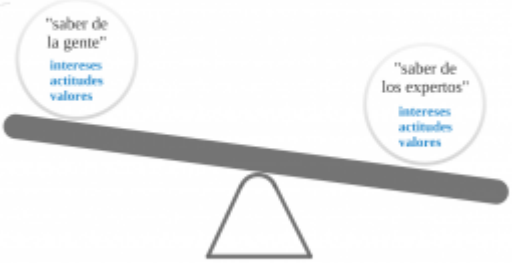
¿Qué es ahora "participar"?
¿Cómo articular esa nueva participación?

Differentes perspectivas



Todos participan

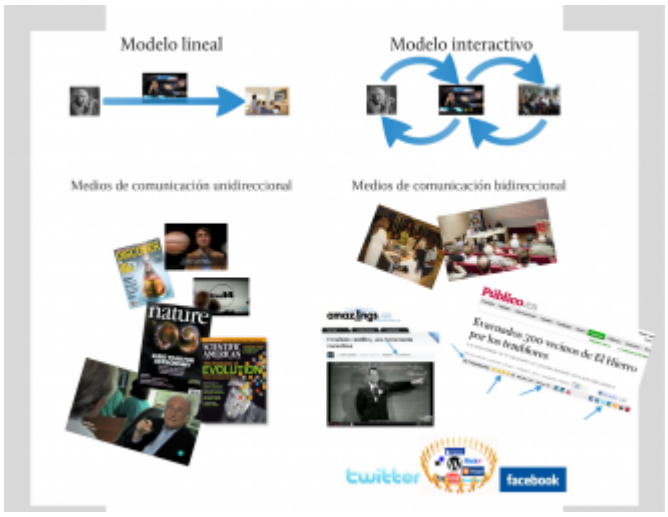
Modelo interactivo de divulgación

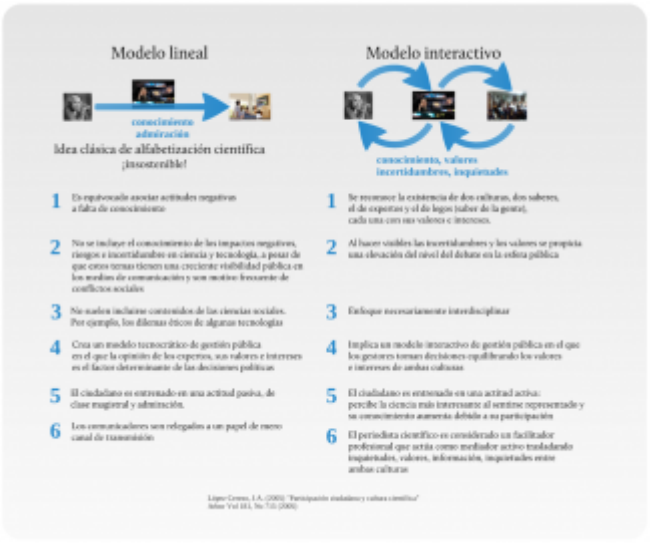


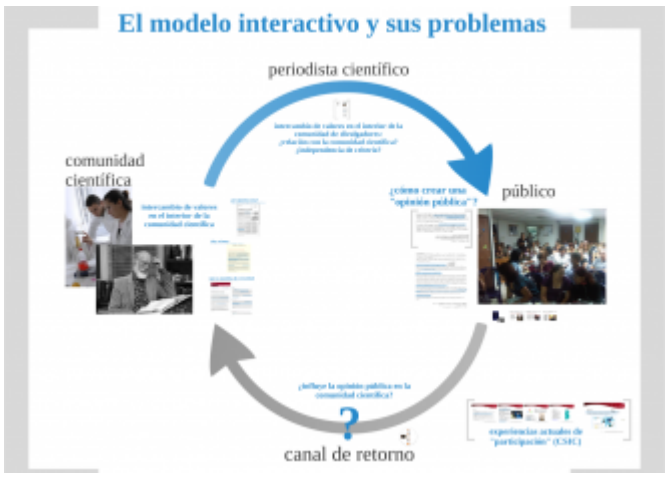
diálogo de culturas

Modelo lineal de divulgación científica (y de la gestión)

Consecuencia:
La cultura científica es un prerrequisito para la participación







En este Prezi, procedente de un curso que impartió en 2011 Joaquín en el Museo de la Ciencia en Madrid, pueden ver la presentación completa de la que están sacadas estas imágenes que muestran la dificultad de acercar el conocimiento experto a los saberes de la gente.

Hacia una Inteligencia Artificial Sucia

El Manifiesto por una Ciencia Sucia precisamente apunta a crear mecanismos de divulgación de la ciencia y la tecnología que huyan de lo sensacionalista y lo espectacular:

Tal como Platón expulsó de su república a los poetas por considerar que sus conocimientos no trataban sobre la realidad política (que solo era conocida y comprendida por los filósofos), la toma de decisiones por parte de "expertos" nos enfrenta al peligro de expulsar de las decisiones políticas en materia de ciencia a la ciudadanía por considerar que no pertenecen a la élite que puede comprender los complejos problemas que genera la investigación científica y la tecnología. Esta situación equivaldría a considerar a los ciudadanos como meros receptores de productos tecnológicos sobre los que no han tenido oportunidad de expresarse durante su gestación.

En un mundo en el que es necesario un cierto dominio de la tecnología y en el que ya hemos podido comprobar la estrecha relación que existe entre la ciencia que hacemos y el medio ambiente en el que vivimos hay que aspirar a mucho más que a que los ciudadanos admiren a sus científicos. A eso se aspira en medicina con el concepto de "empowered patient", por el que el enfermo participa en la toma de decisiones junto con su médico. Y a eso hay que aspirar en la investigación científica, a que

su curso se trace en diálogo público con la sociedad.

No creemos en la divulgación de la ciencia como espectáculo, como un circo de maravillas exóticas.

Inspirados por la "pragmática sucia" del filósofo Quintín Racionero propusimos en su día pensar como "ciencia sucia" a aquella que toma conciencia de sí, de su condición filosóficamente infectasucia, consciente de todas las instancias que determinan su relevancia así como de todos los intereses e influencias que la afectan.

Aplica lo mismo a las tecnologías detrás de los modelos Transformer, corazón de los actuales modelos grades de lenguaje (LLMs). Una Inteligencia Artificial "ensuciada filosóficamente" con las aportaciones críticas de pensadores y sociedad puede estar más alineada que una Inteligencia Artificial construida únicamente con criterios de rentabilidad y mercado.

Nick Srnicek sobre la IA contemporánea

Hilo en Twitter del 30 de marzo (traducción automática)

Encuentro que Twitter es más útil cuando las personas comparten información e ideas, así que aquí hay un breve hilo con algunas piezas críticas sobre la economía política de la IA, algo que generalmente se ignora a favor de la ética de la IA o el riesgo x de AGI o las preocupaciones de privacidad.

En primer lugar, un artículo muy relevante de @DLuitse y Wiebke Denkena sobre la arquitectura de transformadores en los LLM y su impacto en la economía política.

- https://journals.sagepub.com/doi/full/10.1177/20539517211047734
- The great Transformer: Examining the role of large language models in the political economy of AI

En segundo lugar, un artículo agudo de @mer__edith advirtiendo sobre la concentración de investigación e infraestructura de IA

- https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4135581
- The Steep Cost of Capture

En tercer lugar, una de las pocas piezas críticas del tamaño de un libro sobre la economía de la IA contemporánea: un libro esclarecedor e incisivo en muchos sentidos de Nick Dyer-Witheford, @atlemk y @JamesSteinhof17.

- https://www.plutobooks.com/9780745338606/inhuman-power/
- Inhuman Power: Artificial Intelligence And The Future Of Capitalism

Y, por último, mi propio trabajo sobre este tema, donde trato de establecer un marco básico para comprender la industria de la IA y dónde creo que están ocurriendo las dinámicas de monopolización

- Data, Compute, Labour
- Data, Compute, Labour

(Espero terminar un libro sobre esto pronto, así que los comentarios/críticas siempre son bienvenidos)

Los materiales que referencia Srnicek dan, por sí mismos, para un episodio analizando la parte económica de una sociedad con IA industrial extendida. Quizás el artículo que más se acercaría al enfoque de este episodio es el que se titula “The Steep Cost of Capture”, de Meredith Whittaker, New York University. Trata sobre la concentración de la investigación en IA, aunque ese artículo es de 2021 y nos parece que la premisa de la que parte ya no es correcta:

“*AI considerar cómo abordar esta avalancha de IA industrial, primero debemos reconocer que los “avances” en IA celebrados durante la última década no se debieron a avances científicos fundamentales en técnicas de IA. Fueron y son principalmente el producto de datos significativamente concentrados y recursos informáticos que residen en manos de unas pocas grandes corporaciones tecnológicas. La IA moderna depende fundamentalmente de los recursos corporativos y las prácticas comerciales, y nuestra creciente dependencia de dicha IA cede un poder desmesurado sobre nuestras vidas e instituciones a un puñado de empresas tecnológicas. También otorga a estas empresas una influencia significativa tanto sobre la dirección del desarrollo de la IA como sobre las instituciones académicas que desean investigarla.*”

La arquitectura Transformer sí que nos parece que es un “fundamental scientific breakthroughs in AI techniques”.

Es verdad que en 2021 los avances de la IA se basaban exclusivamente en alimentar con datos concentrados en grandes corporaciones a los modelos de IA anteriores Pero ahora tenemos un nuevo tipo de IA y se mantiene el tema de dónde proceden los datos con los que alimentarla. Antes teníamos un enfoque y ahora tenemos que dividirlo en dos. De repente todo lo que ya estaba pensado hace un año se ha quedado obsoleto

Materiales

Actually, Othello-GPT Has A Linear Emergent World Representation — Neel Nanda
AI risk ≠ AGI risk
Large Language Models are reasoners with Self-Verification