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

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

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.

Alineación e Hiperobjetos

Ahí, en el problema de la alineación, entran de lleno los Hiperobjetos.

- ¿Con qué intereses hay que alinear a la AGI?
- ¿Con la supervivencia del capitalismo?
- ¿Con un cambio de modelo?
- ¿Con los intereses del planeta?
- ¿Con la justicia social?

Pensar en términos de hiperobjetos nos dará claves importantes para detectar los mejores intereses con los que alinear a la AGI.

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:

“*Al 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