El caso para FRENAR LA INTELIGENCIA ARTIFICIAL

Frenar la inteligencia artificial podría ser lo mejor que podamos hacer por la humanidad.

Primero, el año pasado obtuvimos DALL-E 2 y Stable Diffusion, que pueden convertir unas pocas palabras de texto en una imagen impresionante. Luego, OpenAI, respaldado por Microsoft, nos dio ChatGPT, que puede escribir ensayos tan convincentes que asusta a todos, desde los profesores (¿y si ayuda a los estudiantes a hacer trampa?) hasta los periodistas (¿podría reemplazarlos?) y los expertos en desinformación (¿amplificará las teorías de la conspiración?). Y en febrero, obtuvimos Bing (también conocido como Sydney), el chatbot que deleitó y al mismo tiempo molestó a los usuarios beta con interacciones espeluznantes. Ahora tenemos GPT-4, no solo el último modelo de lenguaje grande, sino uno multimodal que puede responder tanto a texto como a imágenes.

El miedo a quedarse atrás de Microsoft ha llevado a Google y Baidu a acelerar el lanzamiento de sus propios chatbots rivales. La carrera de la IA está claramente en marcha.

Pero ¿correr es una idea tan buena? Ni siquiera sabemos cómo lidiar con los problemas que plantean ChatGPT y Bing, y son muy complicados en comparación con lo que está por venir.
¿Qué pasa si los investigadores logran crear una IA que iguale o supere las capacidades humanas no solo en un dominio, como los juegos de estrategia, sino en muchos dominios? ¿Qué pasaría si ese sistema resultara peligroso para nosotros, no porque quiera eliminar activamente a la humanidad, sino simplemente porque persigue objetivos que no están alineados con nuestros valores?
Algunos expertos temen que ese sistema sea una máquina fatal, literalmente creada por nosotros mismos.

Por lo tanto, la IA amenaza con sumarse a los riesgos catastróficos existentes para la humanidad, como la guerra nuclear global o las pandemias generadas por bioingeniería. Pero hay una diferencia. Si bien no hay forma de desinventar la bomba nuclear o las herramientas de ingeniería genética que pueden exprimir a los patógenos, aún no se ha creado una IA catastrófica, lo que significa que es un tipo de fatalidad que tenemos la capacidad de detener de forma preventiva.

Sin embargo, aquí está lo extraño. Los mismos investigadores que están más preocupados por la IA no alineada son, en algunos casos, los que están desarrollando una IA cada vez más avanzada. Razonan que necesitan jugar con una IA más sofisticada para poder descubrir sus modos de falla y, en última instancia, prevenirlos mejor.

Inglés | Artículo

Fuente: Sigal, S. (2023). *Generative AI has an Intellectual Property Problem.* https://www.vox.com/the-highlight/23621198/artificial-intelligence-chatgpt-openai-existential-risk-china-ai-safety-technology