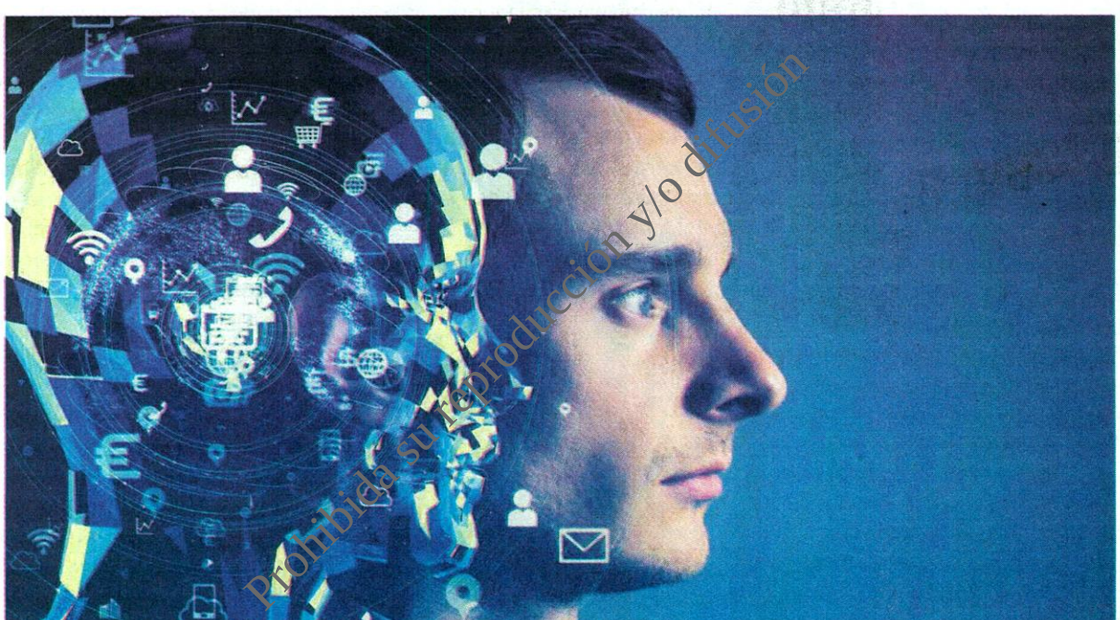
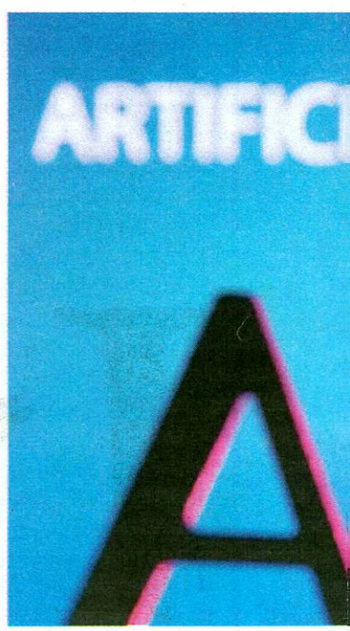
GERMÁN ALARCO

Profesor de la Universidad del pacífico

Capacidades y riesgos Artificial (IA) de fronte

GRAN BRETAÑA CONVOCÓ, la primera semana de noviembre, a una Cumbre Internacional de gobiernos, organizaciones internacionales, Universidades y la industria. El objetivo era discutir las capacidades y riesgos de la IA, especialmente de frontera o de vanguardia.





El Departamento de Ciencia, tecnología e Innovación británico preparó como base para las discusiones, entre otros documentos, uno relativo a la IA que se comenta en esta nota. En la reunión participaron algunas universidades de la región, pero a nivel gubernamental de América Latina, solo Brasil y Chile. Desafortunadamente, como siempre, en el Perú seguimos mirándonos al ombligo, ignorando las grandes transformaciones internacionales.

ANTECEDENTES

El documento anota que estamos en medio de una revolución tecnológica que alterará fundamentalmente la forma en que vivimos, trabajamos y nos relacionamos unos con otros. La IA promete transformar casi todos los aspectos de nuestra economía y sociedad: avanzar en el descubrimiento de fármacos, hacer que el transporte sea más seguro y limpio, mejorar los servicios públicos, acelerar y mejorar el diagnóstico y tratamiento de enfermedades como el cáncer y mucho más.

La IA de vanguardia está transformando la productividad y los servicios de software. Los sistemas más avanzados pueden escribir textos con fluidez y extensión, escribir bien, obtener buenas calificaciones en exámenes escolares, generar artículos de noticias convincentes, traducir muchos idiomas, resumir documentos extensos, entre otras capacidades.

RIESGOS

Sin embargo, anota el documento, estas enormes oportunidades conllevan riesgos que podrían amenazar la estabilidad global y socavar nuestros valores. Para aprovechar las oportunidades, debemos comprender y abordar los riesgos. La IA plantea riesgos que no respetan las fronteras nacionales.

Es importante que los gobiernos, el mundo académico, las empresas y la sociedad civil trabajen juntos para sortear estos riesgos, que son complejos y difíciles de predecir, para mitigar los peligros potenciales y garantizar que la IA beneficie a la sociedad.

CONTENIDO

El gobierno británico cree que se necesita más investigación sobre el riesgo de la IA. El informe explica por qué. Describe el estado actual y las tendencias clave relacionadas con las capacidades de IA de vanguardia, y luego explora cómo las capacidades de IA de frontera podrían evolucionar en el futuro y revisa algunos riesgos clave. Existe una gran incertidumbre en torno a las capacidades y los riesgos de la IA, incluidos algunos expertos.

El informe cubre muchos riesgos, pero enfatiza que el riesgo general es una pérdida de confianza y confiabilidad en esta tecnología que nos negaría permanentemente a nosotros y a las generaciones futuras sus beneficios transformadores positivos. Al discutir los otros riesgos, lo hacen con el fin de impulsar acciones para mitigarlos.

DEFINICIÓN BÁSICA

Según el documento, definir la IA es un desafío, ya que sigue siendo una tecnología que evoluciona rápidamente. Para los propósitos de la Cumbre, se define IA de frontera o de vanguardia como modelos de IA de propósito general altamente capaces que pueden realizar una amplia variedad de tareas e igualar o superar las capacidades presentes en los modelos más avanzados de la actualidad.

Hoy en día, esto incluye principalmente modelos de lenguajes grandes (LLM) como los que subyacen a ChatGPT, Claude y Bard. Sin embargo, es importante señalar que, tanto hoy como en el futuro, los sistemas de IA de frontera podrían no estar respaldados por LLM, y podría estar respaldado por otras tecnologías.

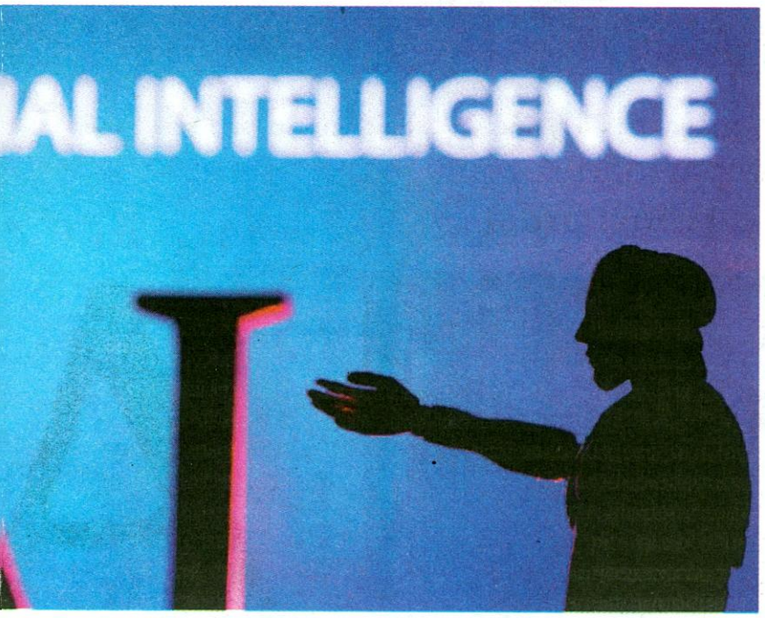
CÓMO FUNCIONAN IA

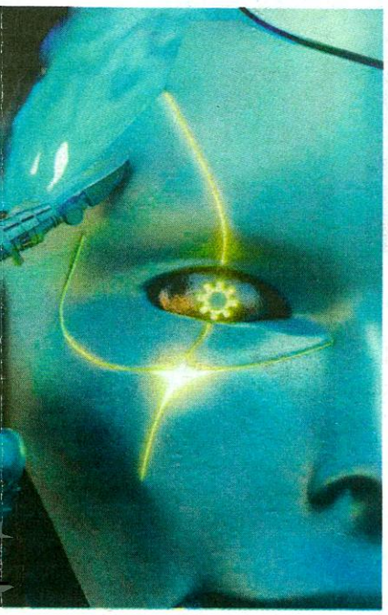
En el informe se anota que las empresas de vanguardia en inteligencia artificial como OpenAI, DeepMind y Anthropic desarrollan grandes modelos de lenguaje (LLM) como GPT-4 en dos fases: preentrenamiento y ajuste. Durante la capacitación previa, un LLM lee millones o miles de millones de documentos de texto. A medida que lee, palabra por palabra, predice qué palabra vendrá a continuación.

Al inicio del pre-entrenamiento predice aleatoriamente, pero a medida que ve más datos, aprende de sus errore y mejora su rendimiento predictivo Una vez finalizado el entrenamiento previo, el modelo es significativamente mejor que los humanos a la hora de predecir la siguiente palabra de un documento de texto elegido al azar.

Durante el ajuste fino, la IA pre

de la Inteligencia
ra o vanguardia





viamente entrenada se entrena aún más en conjuntos de datos altamente seleccionados, que se centran en tareas más especializadas o están estructurados para dirigir el comportamiento del modelo de manera que estén alinea con los valores del desarrollador y las expectativas del usuario.

TAREAS POSIBLES

Conversar con fluidez y detalladamente; Resumir documentos extensos; escribir largas secuencias de código que funcione correctamente a partir de instrucciones en lenguaje natural; responder preguntas sobre imágenes que requieran razonamiento de sentido común; y traducir documentos.

Asimismo, acelerar la investigación académica, por ejemplo, en economía; analizar datos trazando gráficos y calculando cantidades clave; dirigir las actividades de los robots mediante el razonamiento, la planificación y el control de movimientos, entre otras.

Por otra parte, los modelos son más útiles cuando se acompañan con otras herramientas y software. Los investigadores han creado programas de software llamados andamios que permiten que modelos de IA de vanguardia impulsen agentes de IA autónomos. Se anota como ejemplos, encontrar información específica, apoyar la síntesis de sustancias químicas; y resolver problemas complejos en juegos de supervivencia de mundo abierto como Minecraft y Crafter.

EL FUTURO

Hay errores en las respuestas que generan actualmente los sistemas de IA, pero según el informe, el progreso reciente de la IA ha sido rápido y probablemente continuará. Esto se debe a mejoras predecibles en el rendimiento de los modelos de IA de vanguardia cuando se desarrollan con más computación, más datos y mejores algoritmos.

También pueden surgir nuevas capacidades inesperadas. En un futuro no muy lejano podrían desarrollarse agentes avanzados de IA de uso general, aunque esto es un tema de debate, especialmente en lo que respecta al momento.

Sin embargo, existe una gran incertidumbre sobre el cronograma de estas capacidades. Muchos otros investigadores, si no la mayoría, no esperan que los sistemas de IA alcancen generalmente el desempeño humano hasta dentro de veinte años y no están de acuerdo en que esto sea una preocupación. Anotan también que históricamente, y con frecuencia, ha habido predicciones de avances inminentes en la IA que no se concretaron.

RIESGOS

Se anota que debemos comprender los riesgos asociados con la IA de vanguardia para acceder y aprovechar de forma segura las oportunidades y beneficios que brinda la tecnología. Estos pueden ser factores de riesgo transversales: condiciones técnicas y sociales que podrían agravar una serie de riesgos particulares. Por otra parte, hay riesgos individuales bajo tres conceptos: daños sociales, uso indebido y pérdida de control.

Los riesgos transversales se refieren a los numerosos desafíos técnicos de larga data para construir sistemas de IA seguros, evaluar si son seguros y comprender cómo toman decisiones. Estos, según el informe, presentan fallas inesperadas y existen barreras para monitorear su uso.

En general, los sistemas de IA de frontera no son sólidos, es decir, con frecuencia fallan en situaciones suficientemente diferentes a sus datos de entrenamiento. Aunque la robustez de la IA es un campo de investigación bien desarrollado en la práctica, la falta de robustez sigue siendo un problema sin resolver que afecta a todo tipo de modelos de aprendizaje automático.

SEGURIDAD MÍNIMA

Algunos investigadores han argumentado que la industria de la IA debería aprovechar las prácticas observadas en industrias altamente centradas en la seguridad, como la atención médica, la aviación y la ingeniería nuclear. Sin embargo, los estándares de seguridad de la IA aún se encuentran en una etapa inicial. El trabajo de organizaciones de desarrollo de normas todavía está en curso en muchas áreas.

Por otra parte, actualmente hay poca capacidad gubernamental para regular y se requiere más trabajo para construir un ecosistema maduro. Un desafío, según el reporte, es que los sistemas a menudo se desarrollan en un país y luego se implementan en otro, lo que aumenta la necesidad de coordinación global. Hay también incentivos insuficientes para que los desarrolladores de IA inviertan en medidas de mitigación de riesgos.

CONCENTRACIÓN PODER

Los investigadores y reguladores han comenzado a explorar la probabilidad de una alta concentración de poder de mercado entre los desarrolladores de IA de frontera. Los altos costos iniciales asociados con el entrenamiento de modelos de IA de frontera parecen crear economías de escala y barreras significativas de entrada para los actores más pequeños.

Una concentración considerable del poder de mercado podría debilitar la competencia, reduciendo la innovación y las opciones de los consumidores. La pérdida de opciones del consumidor también significa que los usuarios tienen menos voz en el uso de sus datos personales, una posible manipulación del comportamiento, vigilancia y una erosión de las normas democráticas.

DAÑOS SOCIALES

El documento señala que existe una amplia gama de posibles daños sociales derivados del uso de la IA. Esto ha provocado un debate en torno a la ética de la IA, con una amplia proliferación de marcos y principios éticos. Aquí se centran sólo en unos pocos daños sociales, pero esto no es para restar importancia a los demás.

Ellos se enfocan en la degradación de la información (al generar contenido realista a bajo costo que puede representar falsamente personas y eventos) y en la disrupción laboral. Los economistas ven la perturbación y el desplazamiento en los mercados laborales como uno de los riesgos que pueden afectar a los ciudadanos y reducir el bienestar social, aunque efectivamente también puede generar mejoras en las condiciones laborales, reduciendo históricamente la demanda de mano de obra en ocupaciones más peligrosas.

DISRUPCIÓN LABORAL

Los estudios sugieren que los sectores con mayor exposición a la disrupción del mercado laboral debido a la IA actual son Tics, financieras y legales, mientras que la educación, la manufactura, la agricultura y la minería son las menos expuestas.

Los modelos de IA de frontera pueden contener y magnificar sesgos arraigados en los datos con los que se entrenan, lo que refleja desigualdades y estereotipos sociales e históricos. Estos sesgos, a menudo sutiles y profundamente arraigados, comprometen el uso equitativo y ético de los sistemas de IA, lo que dificulta que la IA para mejorar la equidad en las decisiones.

SIN CONTROL

El último tema que se aborda en el informe, antes de sus conclusiones, es que la pérdida de control podría acelerarse si los sistemas de IA toman medidas para aumentar su propia influencia y reducir el control humano.

Al respecto, los sistemas actuales tienen algunas capacidades básicas que, si continúa el rápido progreso de la IA, podrían usarse para aumentar su propia influencia; aunque actualmente, estas capacidades no son suficientes para plantear riesgos significativos.