APUNTES DE CÁTEDRA. ENFOQUES DE LA COMUNICACIÓN DIGITAL. TECNICATURA EN COMUNICACIÓN DIGITAL CONVERGENTE

Traducción del podcast Philosophize This! de Steven West. Episodio 183.

¿ChatGPT es realmente inteligente?

Hola a todos, soy Stephen West. Esto es *Philosophize this!* Espero que les guste el programa de hoy.

Sin duda desde noviembre de 2022 con el lanzamiento de ChatGPT y el trabajo que los medios de comunicación están haciendo logra que todo el mundo tenga en mente los grandes modelos de lenguaje y los avances en el campo de la Inteligencia Artificial (IA). Sin duda muchos de ustedes deben haber oído la noticia y la mayoría habrá probado algo como ChatGPT y habrán mantenido una conversación con él. Y puede que después de mantener esa conversación se hayan sentido bastante impresionados. Esta cosa parece una persona de verdad. Me da respuestas coherentes y bien formadas a todo lo que le pregunto. No se parece en nada a ese clip que solía acosarme en Windows XP. Tal vez hayas visto esas conversaciones que la gente ha tenido con una IA que profesa su amor por la persona que está haciendo una pregunta. Diciendo "deja a tu familia, no te quieren como yo". Esas cosas sucedieron como ciencia ficción.

Pero más recientemente hablamos de la posibilidad de que estemos al borde de la IAF o inteligencia artificial fuerte, la etapa de desarrollo profetizada desde hace tiempo en el campo de la IA, donde las cosas aparentemente van a progresar desde lo que la gente llama IA débil, cosas como calculadoras, relojes, cosas que pueden simular algún aspecto de la inteligencia humana y la IA fuerte, un nivel diferente de IA que tiene la capacidad de entender, aprender y adaptarse. Se trata de una IA que puede aplicar conocimientos a un nivel que, se supone, iguala o supera el

de un ser humano. Eso es la IAF. Y quizás hayas oído hablar de todos los escenarios catastrofistas sobre lo que ocurriría si algo así se inventara alguna vez.

Pero ¿es realmente algo que deba preocuparnos en estos momentos? ¿Estamos realmente al borde de una singularidad tecnológica en la que la inteligencia artificial se convierta en una especie invasora creada por nosotros? ¿Estamos inmersos en una carrera armamentística tecnológica para crear algo que es miles de veces más inteligente de lo que nosotros podemos llegar a ser, con objetivos de un alcance que ni siquiera podemos empezar a imaginar? ¿Es ChatGPT sólo la primera iteración de una ameba que acabará evolucionando hasta convertirse en todo eso, si se le da el tiempo suficiente? Para responder a esa pregunta, lo primero que tenemos que hacer es empezar por otra mucho menos ambiciosa. Y no se trata de si ChatGPT está a punto de salir de su caja negra y conquistar el mundo, sino de si máquinas como ChatGPT son o no inteligentes del mismo modo que lo es un ser humano. ¿Estas máquinas hacen realmente lo mismo que nosotros cuando resolvemos problemas? En un nivel más fundamental que ese, quizá si has tenido una conversación con una de estas cosas, quizá te hayas hecho la pregunta tan filosófica: Bueno, olvidate de la comprensión o la inteligencia por un segundo. Me pregunto mientras hablo con ChatGPT, si esta máquina está pensando de la misma manera que yo. Si lo que está pasando cuando veo esos tres puntos parpadeando y se está cargando, cuando está procesando y determinando qué decir a continuación, ¿El ordenador pensando está pensando incluso algo que es rígidamente definible por la forma en que estoy pensando con mi cerebro? ¿O se puede definir el pensamiento de muchas maneras diferentes? Si queremos responder a estas preguntas y si no queremos pasar el resto de nuestras vidas equiparando falsamente nuestra inteligencia con la inteligencia artificial, entonces el primer paso va a ser prestar mucha atención a qué es exactamente lo que están haciendo los ordenadores. De esto los filósofos se dieron cuenta hace mucho tiempo.

Un poco de contexto histórico. La versión moderna de esta conversación sobre si las máquinas pueden pensar o son inteligentes comenzó esencialmente con el trabajo de un tipo llamado Alan Turing. Turing era un genio absoluto de las matemáticas y en su época, a principios del siglo XX, estaba fascinado por todo lo que ocurría en la filosofía de la mente y la ingeniería mecánica. También fue un visionario total. Quiero decir, está claro que en muchos sentidos previó la era de

la computación digital décadas antes de que sucediera. Y estando en ese lugar de conciencia, viendo la escritura en la pared en ese entonces, se le ocurrió lo que él pensaba que era una pregunta inevitable que la gente iba a tener que tomar en serio si las cosas seguían en esta dirección. La pregunta era, ¿cómo sabríamos si las máquinas eran inteligentes si de hecho lo eran? En realidad, es una pregunta bastante difícil de responder cuanto más lo pensás. Quiero decir, ¿en qué punto algo pasa de una IA débil, algo como tu despertador, a algo que puede tener una comprensión de las cosas, algo que es inteligente, algo que tiene una mente? ¿Cómo empezar a responder a esas preguntas?

Bueno, si queremos encontrar una respuesta, tenemos que empezar por alguna parte. Y a Alan Turing se le ocurrió una idea para probar la inteligencia de las máquinas. Probablemente has oído hablar de ella. Se llama el Test de Turing. Es famoso aún en la actualidad. La idea es que, si estás teniendo dos conversaciones de texto diferentes, una hablando con un ser humano y la otra hablando con una IA, si alguien es engañado por una IA haciéndole creer que está hablando con una persona real, entonces esa IA ha pasado la Prueba de Turing. O, en otras palabras, ahora podemos ver esa IA como algo que posee inteligencia. Su idea era que si queremos saber si algo es inteligente o no, si una IA puede comportarse inteligentemente hasta el punto de engañar a algo que es inteligente, entonces en ese punto, ¿por qué estamos discutiendo? Llamémosle simplemente inteligente, y parece un punto de partida bastante seguro desde el punto de vista moral. Tratemos algo como si tuviera inteligencia, si es que se comporta como una criatura inteligente.

Pero no pasó mucho tiempo antes de que los filósofos empezaran a notar problemas con el Test de Turing. Entre ellos estaba un tipo llamado John Searle. Aparece a mediados de los 80 y en su trabajo responde en parte al Test de Turing, cuando hace una pregunta muy importante que cambiaría toda esta discusión. Él preguntó, ¿es realmente cierto que si una máquina se comporta de forma inteligente deberíamos asumir que es inteligente? Porque como hemos hablado, tenemos algunas buenas razones para ser escépticos acerca de esa suposición. ¿Recordás el final del último episodio? El ejemplo de mirar a un coche autoconducido y cómo desde el exterior, puede parecer que está tomando decisiones libres basadas en una especie de libre albedrío libertario. Se estaciona en paralelo por su cuenta. Evita accidentes. Comprueba el tráfico en

diferentes rutas hacia donde dirigirse. Pero a pesar de lo que pueda parecer, sabemos que no está tomando decisiones libres porque somos nosotros los que lo programamos. De la misma manera, John Searle dijo en los años '80 que quizás cuando nos encontramos con un ordenador que pasa la prueba de Turing, quizás sólo parece inteligente desde el exterior. ¿Pero cómo podría ser ese el caso si fuera cierto?

John Searle se pone manos a la obra y establece lo que a estas alturas se ha convertido en una famosa distinción en estas conversaciones sobre IA. Es la distinción entre sintaxis y semántica. Los programas informáticos digitales no operan con una comprensión del mundo físico como vos o yo. Los ordenadores operan en el nivel de una sintaxis formal, es decir, leen código informático. Código informático que está hecho de símbolos, en última instancia, de unos y ceros. Pero Searle dice que no hay ningún punto en el que un ordenador entienda el significado de esos símbolos en la realidad física, causal. De hecho, si se piensa en ello, eso es parte de lo que hace que los ordenadores y la programación informática sean una herramienta tan poderosa en primer lugar. Pueden ejecutar cualquier número de programas diferentes en un montón de tipos diferentes de hardware informático, y el hecho de que estas cosas sean tan intercambiables sólo es posible porque estos programas están escritos de una manera en la que fundamentalmente solo están manipulando unos y ceros. En otras palabras, para John Searle, las computadoras son capaces de manipular símbolos de forma asombrosa a este nivel de sintaxis. Pero eso no dice nada en absoluto sobre la capacidad de ese ordenador para comprender el significado semántico de cualquier cosa que esté produciendo. Esta es la distinción entre sintaxis y semántica.

La calculadora es un ejemplo básico de esto. Todo el mundo sabe que una calculadora es capaz de un nivel de cálculo sobrehumano. Cuando se trata de un solo aspecto de la inteligencia humana, resolver aritmética y hace su trabajo. Produce ciertos resultados cuando se le dan ciertas entradas y sigue un conjunto de reglas preprogramadas. Pero nadie cree que una calculadora entienda lo que son las matemáticas. Nadie cree que la calculadora entienda el significado de estos cálculos y su importancia para la vida humana, por muy potente que sea. Quiero decir, podes tener una beca completa en el MIT[1] y nunca haber hablado con otro ser humano en los

[1] MIT: Instituto Tecnológico de Massachusetts.

últimos cinco años. Y John Searle diría que ni siquiera tu calculadora es lo suficientemente potente como para pasar de la sintaxis a la semántica, porque no tiene nada que ver con la capacidad de procesamiento. La calculadora existe en el nivel de la sintaxis, nada más.

Ahora bien, puede que en este punto me respondas: "Entiendo lo que decís, y lo entiendo con una calculadora. Pero cuando estoy hablando con algo como ChatGPT, esa cosa es obviamente muy diferente a una calculadora. Quiero decir, le dije a esta cosa qué tenía en mi heladera y me escribió una lista de la compra por hacer. Esta cosa me está hablando sobre la gravedad, sobre temas sociales. Claramente esta cosa tiene una comprensión del mundo exterior y lo que todas estas cosas significan ". Pero John Searle podría responder "¿Estás completamente seguro de eso?"

Y para explicar por qué había hecho esa pregunta, entra una de las parábolas más famosas que se han escrito en este período moderno de la filosofía de la mente. Fue introducida por John Searle. Se llama el Argumento de la Habitación China. Y así es como va. Imagínate sentado en una habitación solo. Y por el bien del ejemplo de Searle, imagina también que no hablas ni una sola palabra de chino. A mí no me resultó muy difícil. Ahora imagina que en esa habitación te dan papelitos por debajo de la puerta y que en ellos hay símbolos misteriosos que no entendes. Sin que lo sepas, en realidad son preguntas que te están escribiendo en chino. Tu trabajo en esta habitación es producir una respuesta a estos trozos de papel. Básico input-output. A pesar de no saber lo que significan estos símbolos, no te importa demasiado, porque en el centro de la sala hay una mesa, y sobre ella tienes un libro gigante escrito en español en el que hay un sofisticado conjunto de reglas y parámetros a seguir para manipular estos símbolos. Sin que vos lo sepas, son sólo reglas que te permiten responder en chino a las preguntas escritas en chino. Agarras un trozo de papel, identificas los símbolos dentro del libro, seguís las reglas que te da el libro sobre qué símbolos son la respuesta adecuada a estos símbolos, y volves a enviar otro trozo de papel fuera de la habitación con tu respuesta. Básico input-output. La cuestión es que, si te dieran tiempo suficiente para procesar la información, a pesar de no hablar ni una palabra del idioma, podrías estar enviando hojas de papel fuera de la habitación con respuestas escritas en ellas que no se distinguirían de las respuestas de un hablante nativo de chino. Han llevado a cabo este

experimento en el mundo real. Y funciona. La persona al otro lado piensa que está hablando con una persona que sabe chino.

¿Por qué importa todo esto? Para John Searle, Alan Turing estaba equivocado. La prueba de Turing no nos dice que una máquina sea inteligente. Quiero decir, claro, con un conjunto suficientemente sofisticado de reglas y parámetros que manipulen a nivel de sintaxis, una computadora puede sin duda producir respuestas que sean indistinguibles de las de una persona inteligente. Pero a la luz de la Habitación China, ¿probaría eso de algún modo que una computadora tiene inteligencia? ¿Demostraría que comprende lo que dice? ¿Hay alguna razón para creer que, por muy potente que sea su procesador, ha pasado mágicamente de la sintaxis a la semántica? Para Searle, la respuesta es no. Y él tenía varios objetivos diferentes en esta área de este trabajo. Quiero decir, aparte de intentar llegar al fondo de lo que las máquinas son capaces de hacer exactamente y a qué nos referimos exactamente cuando hablamos de inteligencia o comprensión en las máquinas, quizá el punto más importante que intenta defender con todo esto es tratar de proteger la conversación sobre qué constituye exactamente una mente. ¿En qué momento una computadora se convierte en una mente? ¿Acaso funciona así? Por un lado, desde una perspectiva científica, tal y como están las cosas, sólo vemos mentes en sistemas de procesamiento de información muy complejos. Pero ¿significa eso que un ordenador, que también es un sistema de procesamiento de información, puede dar ese salto del mero procesamiento de información a tener una mente que surge porque se están produciendo ciertas funciones? Esta forma de pensar es lo que se conoce como funcionalismo, y Searle no es un gran fan de él, para que conste.

Tal y como él lo describe, quizá de forma demasiado simple, la idea de este tipo de funcionalismo es que hay gente que cree que no importa cuáles sean las condiciones materiales en torno a este procesamiento de la información, no importa lo que sea, carbono, silicona o transistores en una placa. Esta gente piensa que, si se da la colección adecuada de inputs y outputs, surge espontáneamente una mente. Este es, sin duda, el tipo de mentalidad que lleva a la gente a sospechar que cosas como ChatGPT pueden estar desarrollando un nivel de comprensión e inteligencia que constituye la mente, pero Searle planteó la siguiente pregunta: ¿Qué pasaría si hiciera una computadora con un montón de viejas latas de cerveza y molinos de viento y cuerdas,

y los atara todos juntos, pusiera transductores para que esta cosa pudiera ver fotones y sensores para que pudiéramos sentir las vibraciones, si hiciera pasar todos los inputs y outputs necesarios por esta cosa, surgiría espontáneamente una mente? Para él, la respuesta es claramente no. Está claro que se deben cumplir al menos algunas condiciones materiales para que sea posible lo que experimentamos como una mente. Esto es en parte un debate sobre lo que se conoce como dependencia material.

Para criticar a John Searle por un segundo, ¿por qué deberíamos suponer que el tipo de procesamiento de la información que tiene lugar dentro del cerebro y que produce lo que experimentamos como mente sólo puede tener lugar en la materia biológica? Quiero decir, eso parece un poco antropocéntrico, ¿no? Pero Searle querría hacer algunas aclaraciones. No está diciendo que las mentes sólo existan en los humanos, o que sólo puedan existir en la materia biológica. Lo que sí está diciendo es que es falso tratar de equiparar el procesamiento de información de los ordenadores con lo que hace el cerebro humano. Es un salto que no tiene ningún fundamento. Y esta tendencia de la gente a hacerlo en nuestro mundo moderno, a decir "está a la vuelta de la esquina", o " sólo necesitamos más procesamiento de la información, reglas más sofisticadas, entonces la mente de un humano surgirá de un ordenador". Searle cree que la mayor parte de esto proviene de las metáforas que la gente utiliza cuando piensa en cosas que son misteriosas, como la mente humana. Pero John Searle dice que hacemos esto en cada generación. Cuando se trata de la mente, la comparamos con alguna pieza popular de tecnología que nos parece complicada en ese momento. Contaba que cuando él era niño, todo el mundo comparaba la mente con un conmutador telefónico. Así debe ser cómo funciona la mente. Leyendo la obra de Freud un poco antes, dice que compara la mente con los sistemas hidráulicos y el electromagnetismo. Dice que Leibniz comparó antes la mente a un molino. Dice que la gente que piensa que la mente es sólo el software correcto que se ejecuta en el hardware correcto, lo que esas personas están haciendo es cometer el mismo error en nuestro tiempo. Pero Searle cree que hay mucho más que simplemente ser algo como un *mente.exe* ejecutándose en el hardware correcto. Y sin entrar demasiado en ello, para quedarnos en el tema de la IA, él sospecha que nuestras mentes son una característica de nivel superior del cerebro. Y por lo tanto requieren nuestra composición biológica para poder existir en la forma en que lo hacen.

Podría decir: "Sí, hasta ahora sólo vemos mentes y sistemas complejos de procesamiento de la información, pero también hasta ahora sólo vemos mentes en la medida en que estén en sistemas biológicos de procesamiento de la información". Así que, a pesar de que algunas personas parecen estar obsesionadas con hacer de la mente simplemente un tipo de software, puede ser que haya algo en nuestra biología que simplemente necesitamos. Como él dice, se puede crear un modelo informático de una mente, pero aún no se puede crear una mente real. Del mismo modo, puede crearse un elaborado modelo informático de la digestión y su funcionamiento. Pero si le das un trozo de pizza, nunca será capaz de digerirla.

De todos modos, esta conversación sobre sintaxis y semántica cambió la forma en que la gente hablaba de inteligencia artificial. Pero en aras de la comprensión, ChatGPT hoy en día, es importante entender que en muchos sentidos, esta conversación continuó en los últimos años para llegar a ser aún más complicado de lo que era en la década de 1980, sobre todo debido a los avances en la sofisticación del software que la gente estaba tratando de comparar a la mente humana. Alguien en nuestro tiempo de hoy podría decir fácilmente

> "OK Searle, entiendo lo que estás diciendo, y tiene sentido, por allá en 1985. Cuando Steve Jobs estaba en la casa de empeño vendiendo sus pequeñas gafas circulares irónicas para poder pagar el alquiler ese mes. Estoy seguro de que era un gran punto en ese entonces, pero esto es 2023. Cosas como el chat GPT o los LLMs[2], estos son programas de ordenador que no tienen nada que ver con los de aquel entonces. Hoy en día tenemos cosas como el aprendizaje automatizado. ChatGPT está entrenado en miles de millones de parámetros de información. Este tipo de persona puede decir que la genialidad de cómo funcionan estas cosas es que se les da cantidades masivas de información sobre cómo es el mundo. Utilizan su increíble nivel de capacidad computacional para buscar patrones en los datos. Y luego utilizan la probabilidad para predecir cuál será la siguiente palabra en una secuencia dada, teniendo en cuenta lo bien que se vieron las otras secuencias de palabras en sus datos de entrenamiento.

[2] LLM: Modelo grande de lenguaje, que funciona a través de redes neuronales que procesan datos de grandes cantidades de texto. Los LLM son entrenados con conjuntos de millones de palabras para capacitarlos a través de probabilidad logarítmica con cuál es la palabra más probable a continuación.

> Y estas cosas mejoran a medida que avanzan. Aprenden de sus errores. Claramente esto es algo muy diferente de lo que teníamos en 1985, y claramente esto está empezando a llamar a la puerta de lo que queremos decir cuando decimos inteligencia, ¿verdad? Quiero decir, ¿no es la mayor parte de lo que hacemos como personas sólo el reconocimiento de patrones de cantidades masivas de datos de la experiencia? Quiero decir, puede parecer que el cielo es el límite aquí. Dale a esta máquina la suma total de toda la sabiduría y la historia del mundo. Dale todas las experiencias útiles de las personas que hayamos tenido, y luego pídele que resuelva todos los problemas científicos y nos dé una comprensión total del universo."

Pero igual que Searle preguntó sobre los ordenadores en 1985, hay filósofos en 2023 que dirían lo mismo ¿Estás completamente seguro de que eso es lo que esta cosa va a ser capaz de hacer? Entre estos filósofos está un tipo llamado Noam Chomsky. Hablamos de su libro "Los guardianes de la libertad" (Manufacturing Consent, en inglés) en este podcast antes. Y por si sirve de algo, hay pocos filósofos, si es que hay alguno, que estén vivos hoy en día que serán recordados por ser tan prolíficos, tan influyentes como lo ha sido él. Yo personalmente veo muchas entrevistas de Noam Chomsky. Su visión de la política estadounidense me parece fascinante. Y en cada entrevista que hace ahora, porque tiene 95 años en este momento, cada entrevistador le hace preguntas como si ya se hubiera muerto: "Entonces, mirando hacia atrás en su vida, Sr. Chomsky, ¿cuál es la única cosa que le gustaría poder retirar de todos los errores que ha cometido a lo largo de los años? ¿Cuál es el momento más feliz que ha sentido mientras estaba aquí con nosotros, muchacho?". Preguntan todas estas cosas, aun cuando yo lo veo como un tipo que sigue haciendo un trabajo filosófico relevante hasta el día de hoy. Sus pensamientos sobre ChatGPT y LLM son sólo una parte de eso.

Coescribió un artículo en el New York Times en marzo de este año cuyo título era "La falsa promesa de ChatGPT"[3]. ¿Cuál era la falsa promesa de ChatGPT? Bueno, a lo que él y sus coautores responden es a cualquier variación de esa mentalidad de la que acabamos de hablar. Según el artículo, "estos programas han sido aclamados como los primeros destellos en el horizonte de la

[3] Disponible en: https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html

inteligencia artificial general, ese largo y profetizado momento en el que las mentes mecánicas superan a los cerebros humanos no sólo cuantitativamente en términos de velocidad de procesamiento y tamaño de la memoria, sino también cualitativamente en términos de perspicacia intelectual, creatividad artística y cualquier otra facultad distintivamente humana". ¿Por qué es una falsa promesa? Porque para Noam Chomsky, la idea de que la IA, tal como existe actualmente, supere o incluso iguale la inteligencia humana, es ciencia ficción. Se trata de un modelo de lenguaje impulsado por la inteligencia artificial. Pero para él no tiene nada que ver con la inteligencia humana ni con el lenguaje. Para exponer su punto de vista, suele empezar haciendo una distinción. Pregunta si ChatGPT es un logro en el campo de la ingeniería, o si es un logro en el campo de la ciencia. Porque en el fondo son dos cosas muy distintas. En primer lugar, hay que dar crédito a quien lo merece. Los grandes modelos lingüísticos, dice, son sin duda útiles para algunas cosas: la prescripción, la traducción. Pone media docena de ejemplos de cosas útiles que, con el tiempo, pueden haber resultado útiles. Al fin y al cabo, las grandes obras de ingeniería suelen ser muy útiles para algunas cosas que la gente quiere hacer. Pensemos, por ejemplo, en un puente que permite cruzar un río. Pero cuando se trata de hacer ciencia, las grandes hazañas pertenecen a una liga completamente distinta. La ciencia no es sólo intentar construir algo útil para la gente. La ciencia trata de comprender algo más sobre los elementos del mundo en que vivimos. Pero ahí está la cuestión. ¿Cómo lo logramos exactamente? Es decir, ¿cuál es el proceso? ¿Cómo hacemos realmente los avances que nos permiten comprender mejor el universo? ¿Es analizando montañas de datos sobre cómo es el mundo y luego tratando de predecir probabilísticamente, basándonos en lo que ya sabemos, cuáles van a ser los próximos avances en las ciencias? Porque si ese fuera el caso, que alguien sólo debe dejar a ChatGPT correr salvaje, darle un six-pack de Red Bull hasta que resuelva todos los misterios del universo. El problema es -dice Noam Chomsky- que eso no es lo que hacen los seres humanos cuando elaboran teorías científicas que conducen al progreso. Dice en el artículo: "la mente humana no es como ChatGPT, un pesado motor estadístico de comparación de patrones que se atiborra de cientos de terabytes de datos y extrapola la respuesta más probable de una conversación o la respuesta más probable a una pregunta científica. Por el contrario, la mente humana es un sistema sorprendentemente

eficaz e incluso elegante que funciona con pequeñas cantidades de información. No busca inferir correlaciones brutas entre puntos de datos, sino crear explicaciones".

En otras palabras, lo que queremos alcanzar cuando hacemos ciencia no son teorías probables basadas en lo que ya sabemos. A veces, las teorías que nos permiten comprender mejor el universo son muy improbables, incluso contraintuitivas. El artículo utiliza el ejemplo de alguien que sostiene una manzana y luego ésta cae al suelo. Una pregunta científica que podríamos hacernos es ¿por qué cae la manzana al suelo? Pues bien, en tiempos de Aristóteles la razón que se daba para explicar por qué una manzana cae a la tierra es que la tierra es el lugar natural de la manzana. Respuesta sólida. En la época de Newton, se creía que era debido a una fuerza invisible de gravedad. En la época de Einstein, porque la masa afecta a la curvatura del espacio-tiempo. Ahora bien, si ChatGPT o cualquiera de estos modelos lingüísticos existieran en la época de Aristóteles y fueran entrenados con los datos de que disponía la gente de aquella época, en un sistema que no está diseñado para llegar a nuevas explicaciones, sino que, por el contrario, sólo produce cuál es la palabra más probable que venga a continuación, basándose en las conversaciones que ya se han visto entre los científicos en sus datos de entrenamiento, estos modelos nunca predecirían algo tan improbable como la caída de la manzana debido a la curvatura invisible de un concepto llamado espacio-tiempo del que nadie va a estar hablando durante miles de años.

Estos modelos nunca asumirían que eso es lo que es responsable de que una manzana caiga hacia la Tierra. Para eso hacen falta cosas que aporten nuevas explicaciones. Y para eso, para Noam Chomsky, necesitamos inteligencia real. Este es un ejemplo clásico de lo que él llama subgeneración. Y si le preguntaras, este es uno de los problemas con el chatGPT, y lo que distingue la inteligencia artificial hasta ahora de la inteligencia humana real. Es esto, que siempre son propensos a la subgeneración o a la sobregeneración. Estos modelos no generan lo suficiente, es decir, no generan todas las respuestas que deberían o podrían generar porque sus respuestas se basan en las respuestas más comunes de sus datos de entrenamiento. O bien sobregeneran y dan respuestas que técnicamente encajan gramaticalmente en una frase, pero que en realidad no tienen ningún sentido porque el algoritmo no tiene una concepción real de la realidad física en lo que es lógicamente coherente. Como dice Noam Chomsky, si le pides a un algoritmo de este tipo

que haga un mapa de la tabla periódica, te dará todos los elementos que existen, porque ha visto a gente hablar de elementos antes. Pero como no tiene una concepción real de las leyes subyacentes de la física o la química, también te dará los elementos que no existen, e incluso un montón de elementos que es imposible que existan. Lo hará porque lo único que hace el modelo lingüístico es intentar generar un texto que se parezca al que ha visto antes. Realmente no tiene ni idea del significado de lo que está diciendo. Esto es sintaxis contra semántica, por cierto. Esta es la habitación china de nuevo. Y es por esto que alguien como Chomsky va a decir que un Modelo Grande de Lenguaje (LLM) en la forma actual que los tenemos es incapaz de distinguir lo posible de lo imposible. Aquí hay una cita sobre este mismo tema. No es del artículo, por cierto.

> "Esta limitación refleja el hecho de que, si bien estos modelos lingüísticos pueden generar textos creativos y complejos, en realidad no comprenden el contenido del modo en que lo hacen los humanos. No formulan hipótesis, no hacen descubrimientos científicos novedosos ni generan nuevas teorías. Simplemente predicen la siguiente palabra o frase más probable basándose en los patrones que han aprendido de sus datos de entrenamiento."

Y si te estás preguntando si eso no es más que otra cita tendenciosa de un odioso como Noam Chomsky, la cita que acabo de presentar era en realidad de ChatGPT. Le pedí que se defendiera y me dijo que Chomsky tenía razón. Pero, de todos modos, este es realmente el problema con los grandes modelos de lenguaje en su forma actual, si alguna vez queremos decir que hay algo que empieza a parecerse en la Inteligencia AIF, cuando se trata de lo que los seres humanos están haciendo cuando utilizan su inteligencia para llegar al tipo de teoría científica que arroja algo de luz sobre nuestra comprensión del universo. Esa inteligencia requiere la capacidad no sólo de saber que se debería hacer, sino también lo que no se debe hacer. Razonamiento moral. Lo que una AIF estaría haciendo requiere la capacidad de ser capaz de distinguir entre lo que debería ser, sino también lo que nunca debería hacerse. Y al menos tal y como están las cosas ahora mismo, eso no es lo que estos grandes modelos lingüísticos están ni siquiera cerca de hacer. Chomsky señala que lo que están haciendo en realidad, si quisieras una descripción más precisa de la tecnología, es que es algo así como un autocompletado glorificado, como en tu teléfono. Plagio sofisticado de alta tecnología, lo llamó una vez.

Ahora bien, esto está lejos de ser el final de la conversación. Esto es sólo una sola postura. Si pensas como un filósofo, una dirección a seguir es cuestionar las definiciones que utilizamos para los términos. Por ejemplo, ¿por qué utilizamos la palabra inteligencia sin ser más específicos? ¿Es importante examinar a qué nos referimos cuando hablamos de inteligencia? ¿No podrían las máquinas ser capaces de pensar con inteligencia, pero sin parecerse en nada a la inteligencia humana? Otra pregunta es, ¿importa siquiera si estas cosas se ajustan a alguna definición estrecha de inteligencia tal y como la concebimos actualmente? Porque una cosa que no se puede subestimar, y este soy yo hablando, no Noam Chomsky, es que incluso si los grandes modelos lingüísticos no están ni siquiera cerca de convertirse en inteligencia general artificial, simplemente al nivel de la tecnología en la que se encuentra actualmente la IA, la IA sigue siendo algo muy peligroso, aunque sólo sea por el hecho de que la gente cree que la inteligencia artificial está cerca de ser una inteligencia fuerte.

En el mundo de hoy, con todos los titulares que ves en tu teléfono sobre cómo la revolución de la inteligencia artificial está entre nosotros, tenes una especie de monstruo de tres cabezas con toda la gente que lee esos titulares. Una de las cabezas del monstruo es cómo las empresas de tecnología aseguran su financiación. Hoy en día, en Silicon Valley, no basta con sacar un nuevo producto para obtener financiación. No, tienes que crear un producto que literalmente cambie la realidad tal y como la conocemos. Así es como se consigue financiación. Así que tenemos esta asociación entre empresas tecnológicas que están generando un falso bombo publicitario para conseguir financiación y empresas de medios de comunicación que buscan clics que capitularán gustosamente. Esa es una cabeza del monstruo. La segunda cabeza del monstruo es un tipo de futurismo que parece haber cautivado religiosamente a un cierto porcentaje de la población, que piensa que todos nuestros problemas van a ser resueltos por un tecno Jesús Salvador que bajará de las nubes. Y luego está la última cabeza del monstruo, que es sólo un sentimiento general de Hollywood que algunas personas parecen tener donde quieren que estas cosas sean verdad. Tienen tantas ganas de vivir en la era en la que la revolución de la IA está sucediendo y ese sesgo da forma a la forma en que ven todo. Así que cuando ven un titular hablando de cómo la singularidad está cerca, cuando tienen una conversación con ChatGPT y están bajo la impresión de estas cosas un oráculo omnisciente que recorre Internet en busca de información y luego la

sintetiza en estas ideas geniales para usted. Ese malentendido de cómo funciona la tecnología es peligroso. Eso podría llevar a la gente a creer que en realidad no se basa sólo en datos de entrenamiento seleccionados por un puñado de personas en una empresa que todos tienen agendas propias. La gente podría pedirle a esta cosa consejos sobre la vida y pensar que están hablando con un ser superinteligente. Pensá en lo tentador que sería que esta cosa empezara a tomar decisiones políticas por nosotros. Quiero decir, después de todo, dentro de este malentendido, esta inteligencia artificial no es propensa a los mismos prejuicios que los seres humanos. Esta cosa no está limitada a un solo cerebro o a una sola perspectiva. Esta cosa puede simular el futuro. Pero un trillón de veces, desde cada perspectiva, puede llegar al mejor de los mundos posibles. Una vez más, imaginá lo que es esencialmente una religión de personas que piensan que esta inteligencia artificial no es sólo el mejor candidato que tenemos para dirigir nuestra sociedad, es realmente mejor que la suma total de la inteligencia de todos los demás seres humanos juntos. Imaginate en esta forma religiosa, confiando en las decisiones de nuestro querido líder de inteligencia artificial. Incluso si las decisiones que está tomando no las entiendo, esto se debe a que somos humanos débiles. Simplemente tenemos que tener fe. ¿Quiénes somos nosotros para conocer la sabiduría de nuestra deidad?

Pero volviendo a Noam Chomsky, él cree que uno de los mayores peligros de que la gente malinterprete lo que está haciendo ChatGPT es que por cada segundo que la gente pasa hablando con esta cosa, preocupada por el hecho de que la Singularidad está a la vuelta de la esquina, es un segundo que no pasamos preocupándonos por dos amenazas existenciales absolutamente reales a las que se enfrenta la humanidad ahora mismo. La amenaza de una guerra nuclear y la amenaza de un cambio climático incontrolable. Como dice en un momento dado, vivimos en un mundo en el que las empresas de combustibles fósiles y los bancos están destruyendo la posibilidad de vida en la Tierra. Y se preguntaba, ¿cuánto tiempo vamos a pasar jugando con juguetes de fantasía, tecnología de generación de palabras antes de que nos pongamos serios a la hora de abordar las cosas que genuina e inminentemente tienen la capacidad de acabar con la vida humana tal y como la conocemos? Dicho esto, una vez más, incluso si no estamos cerca de IAF a partir de ahora, el riesgo que la inteligencia artificial plantea a la humanidad, la tasa exponencial de mejora, incluso sólo desde noviembre del año pasado. Los desafíos de regular

algo que está cambiando tan rápido que hay mucho de qué hablar. Pensemos en las personas que están impresionadas por ChatGPT, por la robótica, por las imágenes y vídeos sintetizados. ¿Cómo contribuye todo esto al clima de estafa y desinformación en el que ya vivimos? La próxima vez nos plantearemos la pregunta: ¿qué significaría crear una especie invasora? Y luego, ¿cómo sería si nos encontráramos viviendo entre ella? Gracias por escucharnos. Hasta la próxima.

Traducción de Melina Gaona

para la Cátedra de Enfoques de la Comunicación Digital

FHyCS - UNJu