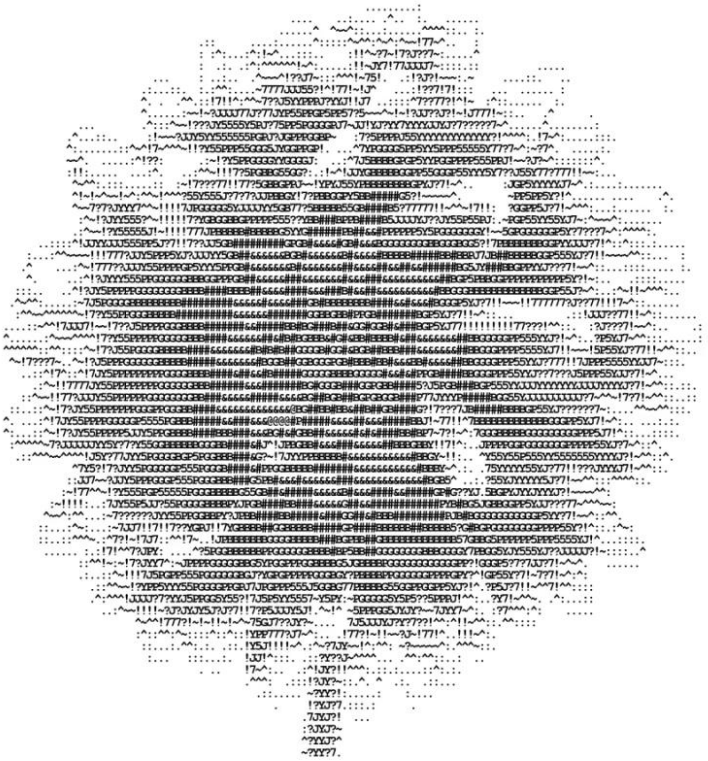
Noam Chomsky: La falsa promesa de ChatGPT

Por Noam Chomsky, Ian Roberts y Jeffrey Watumull

El Dr. Chomsky y el Dr. Roberts son profesores de lingüística. El Dr. Watumull es director de inteligencia artificial en una empresa de ciencia y tecnología.



Jorge Luis Borges escribió una vez que vivir en una época de grandes peligros y promesas es experimentar tanto la tragedia como la comedia, con “la inminencia de una revelación“ para entendernos a nosotros mismos y al mundo. En la actualidad, los avances supuestamente revolucionarios de la inteligencia artificial son motivo tanto de preocupación como de optimismo. Optimismo porque la inteligencia es el medio con el que resolvemos los problemas. Preocupación porque tememos que la cepa de la inteligencia artificial más popular y de moda (el aprendizaje automático) degrade nuestra ciencia y envilezca nuestra ética al incorporar a nuestra tecnología una

concepción fundamentalmente errónea del lenguaje y el conocimiento.

ChatGPT de OpenAI, Bard de Google y Sydney de Microsoft son maravillas del aprendizaje automático. A grandes rasgos, toman enormes cantidades de datos, buscan patrones en ellos y se vuelven cada vez más competentes a la hora de generar resultados estadísticamente probables, como un lenguaje y un pensamiento de apariencia humana. Estos programas han sido elogiados por ser los primeros destellos en el horizonte de la inteligencia artificial *general*, ese momento tan profetizado en el que las mentes mecánicas superan a los cerebros humanos no solo cuantitativamente en términos de velocidad de procesamiento y tamaño de memoria, sino también cualitativamente en términos de perspicacia intelectual, creatividad artística y cualquier otra facultad distintiva del ser humano.

Ese día llegará, pero aún no ve la luz, al contrario de lo que se lee en titulares hiperbólicos y se calcula mediante inversiones insensatas. La revelación borgesiana de la comprensión no se ha producido ni se producirá —y, en nuestra opinión, no puede producirse— si los programas de aprendizaje automático como ChatGPT siguen dominando el campo de la inteligencia artificial. Por muy útiles que puedan ser estos programas en algunos ámbitos concretos (pueden ser útiles en la programación informática, por ejemplo, o para sugerir rimas para versos ligeros), sabemos por la ciencia de la lingüística y la filosofía del conocimiento que difieren en gran medida de la manera en que los seres humanos razonamos y utilizamos el lenguaje. Estas diferencias imponen limitaciones significativas a lo que estos programas pueden hacer, codificándolos con defectos imposibles de erradicarse.

Resulta a la vez cómico y trágico, como podría haber señalado Borges, que tanto dinero y atención se concentren en algo tan insignificante, algo tan trivial comparado con la mente humana, que a fuerza de lenguaje, en palabras de Wilhelm von Humboldt, puede hacer un "uso infinito de medios finitos", creando ideas y teorías de alcance universal.

A diferencia de ChatGPT y sus similares, la mente humana no es una pesada máquina estadística de comparación de patrones, que se atiborra de cientos de terabytes de datos y extrapola la contestación

más probable en una conversación o la respuesta más probable a una pregunta científica. Por el contrario, la mente humana es un sistema sorprendentemente eficiente e incluso elegante que funciona con pequeñas cantidades de información; no busca inferir correlaciones brutas entre puntos de datos, sino crear explicaciones.

Por ejemplo, un niño pequeño que aprende un idioma está desarrollando (de manera inconsciente, automática y rápida a partir de datos minúsculos) una gramática, un sistema increíblemente sofisticado de principios y parámetros lógicos. Esta gramática puede entenderse como una expresión del "sistema operativo" innato, instalado en los genes, que dota a los seres humanos de la capacidad de generar frases complejas y largos hilos de pensamiento. Cuando los lingüistas intentan desarrollar una teoría de por qué una lengua determinada funciona como lo hace ("¿Por qué se consideran gramaticales estas frases y no aquellas?"), están construyendo consciente y laboriosamente una versión explícita de la gramática que el niño construye por instinto y con una exposición mínima a la información. El sistema operativo del niño es completamente distinto al de un programa de aprendizaje automático.

De hecho, estos programas están estancados en una fase prehumana o no humana de la evolución cognitiva. Su defecto más profundo es la ausencia de la capacidad más crítica de cualquier inteligencia: decir no solo lo que es el caso, lo que fue el caso y lo que será el caso —eso es descripción y predicción—, sino además lo que no es el caso y lo que podría y no podría ser el caso. Esos son los ingredientes de la explicación, la marca de la verdadera inteligencia.

A continuación, un ejemplo. Supongamos que sostienes una manzana en la mano. Ahora deja caer la manzana. Observas el resultado y dices: "La manzana se cae". Esa es una descripción. Una predicción podría ser la frase: "La manzana se caerá si abro la mano". Ambas son valiosas y ambas pueden ser correctas. Pero una explicación es algo más: incluye no solo descripciones y predicciones, sino también conjeturas contrafactuales como "cualquier objeto de este tipo caería", más la cláusula adicional "debido a la fuerza de la gravedad" o "debido a la curvatura del espacio-tiempo" o lo que sea. Eso es una explicación causal: "La manzana no habría caído de no ser por la fuerza de la gravedad". Eso es pensar.

El talón de Aquiles del aprendizaje automático son la descripción y la predicción; no plantea ningún mecanismo causal ni leyes físicas. Por supuesto, cualquier explicación de tipo humano no es necesariamente correcta; somos falibles. Pero esto es parte de lo que significa pensar: para tener razón, debe ser posible equivocarse. La inteligencia no solo consiste en hacer conjeturas creativas, sino también críticas creativas. El pensamiento al estilo humano se basa en explicaciones posibles y corrección de errores, un proceso que limita poco a poco las posibilidades que pueden considerarse racionalmente (como le dijo Sherlock Holmes al Dr. Watson: "Cuando hayas eliminado lo imposible, lo que quede, por improbable que sea, debe ser la verdad").

Pero ChatGPT y programas similares, por diseño, son ilimitados en lo que pueden "aprender" (es decir, memorizar); son incapaces de distinguir lo posible de lo imposible. A diferencia de los humanos, por ejemplo, que estamos dotados de una gramática universal que limita los idiomas que podemos aprender a aquellos con un cierto tipo de elegancia casi matemática, estos programas aprenden idiomas humanamente posibles y humanamente imposibles con la misma facilidad. Mientras que los humanos estamos limitados en el tipo de explicaciones que podemos conjeturar a nivel racional, los sistemas de aprendizaje automático pueden aprender tanto que la Tierra es plana como que es redonda. Se limitan a negociar con probabilidades que cambian con el tiempo.

Por esta razón, las predicciones de los sistemas de aprendizaje automático siempre serán superficiales y dudosas. Como estos programas no pueden explicar las reglas de la sintaxis de la lengua inglesa, por ejemplo, pueden predecir, erróneamente, que la frase "John is too stubborn to talk to" significa que Juan es tan terco que no habla con nadie (en lugar de que es demasiado terco como para razonar con él). ¿Por qué un programa de aprendizaje automático predeciría algo tan extraño? Porque podría establecer una analogía en el patrón que infirió a partir de frases como "John ate an apple" (Juan se comió una manzana) y "John ate" (Juan comió), en el que esta última significa que Juan comió algo. El programa bien podría predecir que, como la frase "John is too stubborn to talk to Bill" (Juan es demasiado terco para hablar con Bill) es similar a "John ate an apple" (Juan se comió una manzana), "John is too stubborn to talk to" (Juan

es demasiado terco para hablar) sería similar a "John ate" (Juan comió). Las explicaciones correctas de lenguaje son complicadas y no pueden aprenderse simplemente macerándolas en macrodatos.

Sin ninguna lógica, algunos entusiastas del aprendizaje automático parecen estar orgullosos de que sus creaciones puedan generar predicciones "científicas" correctas (digamos, sobre el movimiento de cuerpos físicos) sin recurrir a explicaciones (que impliquen, por ejemplo, las leyes del movimiento y la gravitación universal de Newton). Pero este tipo de predicción, incluso cuando tiene éxito, es pseudociencia. Aunque es cierto que los científicos buscan teorías que tengan un alto grado de corroboración empírica, como señaló el filósofo Karl Popper: "No buscamos teorías altamente probables, sino explicaciones; es decir, teorías poderosas y altamente improbables".

La teoría de que las manzanas caen al suelo porque ése es su lugar natural (el punto de vista de Aristóteles) es posible, pero solo invita a plantearse más preguntas (¿por qué el suelo es su lugar natural?) La teoría de que las manzanas caen a la tierra porque la masa curva el espacio-tiempo (opinión de Einstein) es altamente improbable, pero en realidad te dice por qué caen. La verdadera inteligencia se demuestra en la capacidad de pensar y expresar cosas improbables pero lúcidas.

La verdadera inteligencia también es capaz de pensar moralmente. Esto significa ceñir la creatividad de nuestras mentes, que de otro modo sería ilimitada, a un conjunto de principios éticos que determinen lo que debe y no debe ser (y, por supuesto, someter esos mismos principios a la crítica creativa). Para ser útil, ChatGPT debe ser capaz de generar resultados novedosos; para ser aceptable para la mayoría de sus usuarios, debe mantenerse alejado de contenidos moralmente censurables. Pero los programadores de ChatGPT y otras maravillas del aprendizaje automático batallan, y seguirán haciéndolo, para lograr este tipo de equilibrio.

En 2016, por ejemplo, el chatbot Tay de Microsoft (precursor de ChatGPT) inundó el internet de contenidos misóginos y racistas, tras haber sido contaminado por troles cibernéticos que lo llenaron de datos de adiestramiento ofensivos. ¿Cómo resolver el problema en el futuro? Al carecer de capacidad para razonar a partir de principios morales, los programadores de ChatGPT restringieron de manera

burda la posibilidad de aportar algo novedoso a los debates controvertidos; es decir, importantes. Se sacrificó la creatividad por una especie de amoralidad.

Consideremos el siguiente intercambio que uno de nosotros (Watumull) mantuvo hace poco con ChatGPT sobre si sería ético transformar Marte para que pudiera albergar vida humana:

Nótese, a pesar de todo el pensamiento y lenguaje en apariencia sofisticados, la indiferencia moral nacida de la falta de inteligencia. Aquí, ChatGPT exhibe algo parecido a la banalidad del mal: plagio, apatía y obviedad. Resume los argumentos estándar de la literatura mediante una especie de superautocompletado, se niega a adoptar una postura sobre lo que sea, alega no solo ignorancia sino falta de inteligencia y, en última instancia, se defiende con un "solo cumplía órdenes", trasladando la responsabilidad a sus creadores.

En resumen, ChatGPT y sus afines son constitutivamente incapaces de equilibrar la creatividad con la restricción. O bien generan de más (produciendo tanto verdades como falsedades, respaldando decisiones éticas y no éticas por igual) o generan de menos (mostrando falta de compromiso con cualquier decisión e indiferencia ante las consecuencias). Dada la amoralidad, la falsa ciencia y la incompetencia lingüística de estos sistemas, solo podemos reír o llorar ante su popularidad.

Este artículo apareció originalmente en The New York Times.

Traducción actual: https://es-us.noticias.yahoo.com/opinión-noam-chomsky-falsa-promesa-231032168.html

https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html