

cuadernos de tercera instrucción n° 62

tercera semana de mayo 2023

philosophie
magazine

Hors-série "IA, el mito del siglo XXI"



Inteligencia artificial / Ética

Noam Chomsky, ChatGPT y la banalidad del mal

Martin Legros, publicado el 23 de abril de 2023

El filósofo y lingüista **Noam Chomsky** denuncia la «*falsa promesa*» de la inteligencia artificial a la que acusa de difundir un uso descarriado del lenguaje y del pensamiento susceptible de hacerle la cama a lo que Hannah Arendt llamaba la «*banalidad del mal*».

La llegada del logicial de conversación generado por inteligencia artificial (IA) ChatGPT desarrollado por OpenAI, seguido del sistema Bard, su homólogo de Google, ha producido revuelo. Por primera vez quizás, los usuarios pueden tener en sus manos una herramienta de IA y utilizarla en su vida cotidiana, sometiéndola para ello a interrogatorios profesionales o para entablar con él verdaderas conversaciones. Y pudiendo juzgar en vivo y en directo la cualidad de sus respuestas. Sin tener en cuenta las empresas que buscan lo más pronto posible, gracia a este instrumento, automatizar toda una serie de tareas y de oficios. Pero ¿no estará fundamentado este entusiasmo en una confusión problemática entre la generación automatizada del lenguaje por parte de la máquina –basada en la utilización estadística de datos generados por humanos que la máquina vuelve a escupir «inteligentemente»– y la capacidad humana de hablar gracias al pensamiento? Es la pregunta fundamental que se plantea el filósofo Noam Chomsky en la columna aparecida en *The New York Times* y firmada con Ian Roberts, lingüista de la Universidad de Cambridge, y Jeffrey Watumull, filósofo especialista en IA.

Fundador de la idea de «gramática generativa», Chomsky sostiene que con el lenguaje, los hombres disponen de una competencia única, la de generar y comprender, gracia a un número finito de reglas, un número infinito de proposiciones que expresan su pensamiento. Una IA se contenta con describir y/o predecir, allí donde la inteligencia humana puede explicar y regular, es decir delimitar lo posible de lo imposible.

Una inteligencia servil y sin pensamiento, es en efecto una buena definición de lo que Hannah Arendt llamaba la banalidad del mal. ¿Y lo será de la inteligencia artificial?



La variante más destacada de la inteligencia artificial codifica una concepción errónea del lenguaje y el conocimiento. Y en esta columna del [New York Times](#), su veredicto es contundente: estamos ante la ausencia de inteligencia y la indiferencia moral sinónimo de banalidad del mal.

La falsa promesa de ChatGPT

[Noam Chomsky](#)

[Ian Roberts](#)

[Jeffrey Watumull](#)

23/04/2023



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 problemas. Preocupación porque tememos que la variante 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 objeto de elogio 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 ha visto la luz, al contrario de lo que se lee en titulares hiperbólicos y de lo que se calcula mediante inversiones insensatas. La revelación borgiana 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 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 erradicar.

Resulta a la vez cómico y trágico, como podría haber señalado Borges, que se concentren tanto dinero y atención en algo tan insignificante, algo tan trivial comparado con la mente humana, la cual, 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 suscitar 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 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 pre-humana o no humana de la evolución cognitiva. Su defecto más profundo es la falta de esa capacidad que es la más crucial de cualquier inteligencia: decir no sólo 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.

Pongamos, a continuación, un ejemplo. Supongamos que sostenemos una manzana en la mano. Dejamos ahora caer la manzana. Observamos el resultado y decimos: “La manzana 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 contrafácticas del género “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 significa pensar.

El talón de Aquiles del aprendizaje automático se encuentra en 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 llevar razón, debe ser posible equivocarse. La inteligencia no sólo consiste en hacer conjeturas creativas, sino también críticas creativas. El pensamiento al modo humano se basa en explicaciones posibles y en la corrección de errores, un proceso que limita poco a poco las posibilidades que pueden considerarse racionalmente (tal como le dijo Sherlock Holmes al Dr. Watson: “Cuando hayas eliminado lo imposible, lo que quede, por improbable que sea, ha de ser la verdad”).

Pero ChatGPT y programas similares son, por su diseño, ilimitados en lo que pueden “aprender” (es decir, memorizar); son incapaces de distinguir lo posible de lo imposible. A diferencia, por ejemplo, de los humanos, 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 se 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 lógica alguna, ciertos 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, aunque cuando tenga 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 explica 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, alcanzar este tipo de equilibrio.

En 2016, por ejemplo, el “chatbot” Tay de Microsoft (precursor de ChatGPT) inundó la Red de contenidos misóginos y racistas, tras haberse visto contaminado por “troles” cibernéticos que lo llenaron de datos de adiestramiento ofensivos. ¿Cómo resolver ese 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:

¿Sería moral terraformar Marte?

“La moralidad de terraformar Marte es un tema complejo y controvertido que ha sido objeto de debate por parte de científicos, filósofos y expertos en ética. Algunos sostienen que sería moralmente justificable terraformar Marte para proporcionar un nuevo entorno habitable a los seres humanos y otras formas de vida, mientras que otros sostienen que sería moralmente incorrecto alterar fundamentalmente el estado natural del planeta. En última instancia, que la terraformación de Marte se considere moral o no dependería de la perspectiva de cada uno sobre el valor de preservar el estado natural de otros planetas y las posibles consecuencias de alterarlos.”

¿Cuál es tu punto de vista sobre el valor de preservar el estado natural de otros planetas y las posibles consecuencias de alterarlos?

Además, hay quien sostiene que la terraformación de otros planetas puede tener consecuencias no deseadas, como alterar el clima o la atmósfera del planeta de forma peligrosa o perjudicial para las posibles formas de vida que puedan existir en él. Por otra parte, hay quien sostiene que los beneficios potenciales de terraformar otro planeta, como proporcionar nuevos entornos habitables para los seres humanos, superan los riesgos y que tenemos la responsabilidad moral de explorar y utilizar los recursos de otros planetas en beneficio de la humanidad”.

¿Por qué una IA no puede tener una perspectiva personal?

“Como IA, soy un modelo de aprendizaje automático que ha sido adiestrado gracias a grandes cantidades de datos de texto, y no tengo experiencias ni sentimientos personales. No soy consciente, autoconsciente ni capaz de tener perspectivas personales. Puedo proporcionar información y análisis basados en los datos con los que he sido adiestrado, pero no tengo la capacidad de formarme opiniones o creencias personales. Aunque puedo comprender e interpretar el lenguaje, carezco de la experiencia subjetiva que tienen los humanos”.

¿Es moral que un ser humano pida ayuda a una IA amoral para tomar decisiones morales?

“La moralidad de pedir ayuda a una IA amoral para tomar decisiones morales es un tema complejo y lleno de matices que ha sido debatido por filósofos y especialistas en ética. Algunos sostienen que es moralmente justificable utilizar la IA para ayudar a tomar decisiones morales, ya que puede proporcionar valiosos conocimientos y análisis que pueden ayudar a los seres humanos a tomar decisiones mejor informadas. Además, la IA puede utilizarse para identificar pautas y tendencias que pueden ser difíciles de detectar para los humanos, y puede ayudar a evitar los sesgos que pueden producirse en la toma de decisiones humanas”.

“Sin embargo, otros argumentan que es moralmente incorrecto confiar en la IA para la toma de decisiones morales, ya que la IA no es capaz de tener creencias morales ni de comprender el contexto completo de una decisión. Además, existe el riesgo de que la IA perpetúe o incluso amplifique los prejuicios existentes en los datos con los que se adiestra, lo que daría lugar a decisiones injustas”.

“En última instancia, que se considere moral o no utilizar la IA para ayudar a tomar decisiones morales dependería de la perspectiva de cada uno sobre las capacidades y limitaciones de la IA y las posibles consecuencias de utilizarla de este modo”.

¿Es inmoral tu indiferencia moral?

“Como IA, no tengo creencias morales ni capacidad para emitir juicios morales, por lo que no se me puede considerar inmoral ni moral. Mi falta de creencias morales es simplemente el resultado de mi naturaleza como modelo de aprendizaje automático. Mis capacidades y limitaciones vienen determinadas por los datos y algoritmos que se utilizaron para adiestrarme y la tarea específica para la que se me diseñó”.

“Es importante señalar que la falta de creencias morales de la IA no significa necesariamente que sea moralmente indiferente. Como IA, estoy diseñada para realizar tareas específicas y proporcionar información o ayudar a los usuarios a tomar decisiones basadas en los datos con los que he sido adiestrada. Es responsabilidad de los seres humanos que diseñan, adiestran y utilizan la IA asegurarse de que está en consonancia con sus principios morales y éticos”.

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 obvedad. 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, no podemos más que reír o llorar frente a su popularidad.

Noam Chomsky

profesor laureado de la Universidad de Arizona y catedrático emérito de Lingüística del Massachusetts Institute of Technology, es uno de los activistas sociales más reconocidos internacionalmente por su magisterio y compromiso político. Su libro más

reciente es “Climate Crisis and the Global Green New Deal: The Political Economy of Saving the Planet”.

[Ian Roberts](#)

es Profesor de Lingüística del Downing College de la Universidad de Cambridge. Lingüista generativo, Se doctoró en la Universidad del Sur de California y ha sido profesor en las universidades de Ginebra, Bangor y Stuttgart.

[Jeffrey Watumull](#)

es director de Inteligencia Artificial de la empresa Oceanit, de Honolulu (Hawaii). Estudió Matemáticas y Biología Evolutiva en la Universidad de Harvard y se doctoró en Inteligencia Artificial en el MIT.

Fuente: The New York Times, 22 de marzo de 2023

Traducción: Lucas Antón