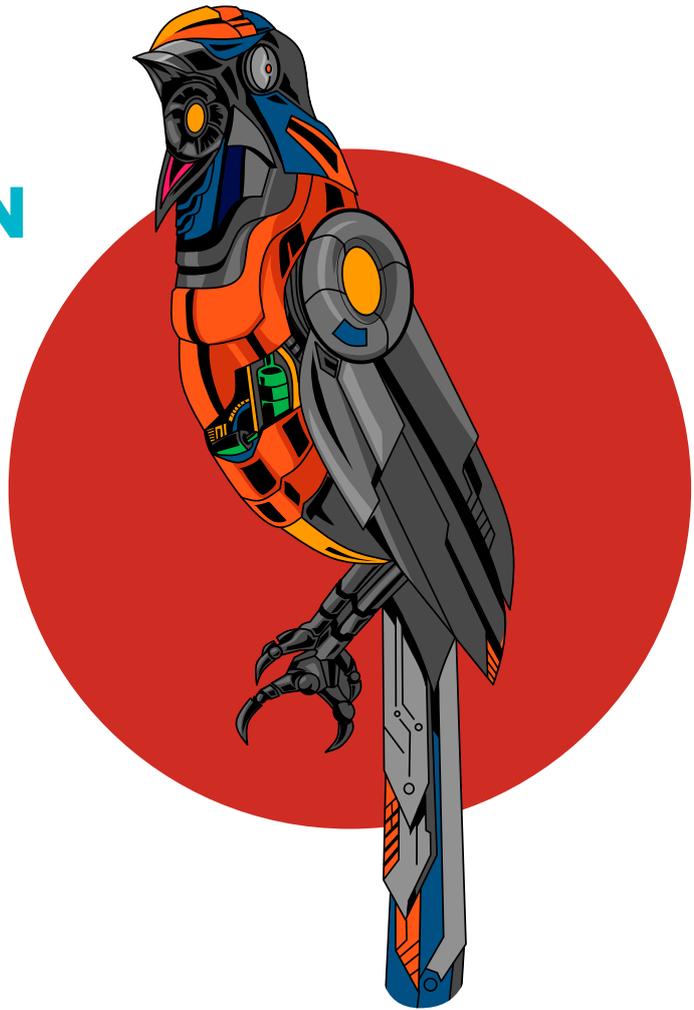


UN LORO ESTOCÁSTICO EN LA HABITACIÓN CHINA:

¿QUÉ NOS ENSEÑA CHATGPT SOBRE LA MENTE HUMANA?

por **José Luis Mendivil Giró**

Para algunos ChatGPT ha refutado las ideas centrales de la gramática generativa. Pero sus presupuestos son tan distintos que no puede considerarse que el cerebro humano sea su modelo.



Si le preguntamos a ChatGPT, el programa de generación de lenguaje de moda, qué es ChatGPT y cómo funciona, nos escribe –en tres o cuatro segundos– lo siguiente (que además ha traducido del inglés):

ChatGPT es un modelo de lenguaje natural basado en el modelo GPT-3 de OpenAI. Es un modelo de aprendizaje profundo que utiliza una arquitectura de red neuronal de transformador para generar texto. Cuando se le da una entrada de texto, ChatGPT utiliza una técnica llamada “aprendizaje por refuerzo” para producir una respuesta coherente y relevante. El modelo analiza la entrada de texto y utiliza su conocimiento previo del lenguaje para generar una respuesta que tenga sentido en el contexto dado. Para generar texto, ChatGPT utiliza una gran cantidad de datos lingüísticos, que se han pre-entrenado en una gran variedad de tareas de procesamiento de lenguaje natural. Esto le permite

a ChatGPT producir respuestas coherentes y relevantes en una variedad de temas y contextos.

Aunque seguramente el lector habrá entendido bastante bien la definición, que es esencialmente correcta, lo maravilloso es que el algoritmo que la ha escrito no entiende ni una palabra de lo que dice. Tampoco sabía que unos seres humanos iban a leer su texto, y ni siquiera sabe qué es un ser humano.

ChatGPT es una herramienta de ingeniería sorprendente y sofisticada, además de extraordinariamente útil. Su potencial, incluyendo su uso comercial, está llamado a ser enorme y, como ya se está señalando intensamente, podría incluso ser peligroso (el psicolingüista y experto en Inteligencia Artificial Gary Marcus ha sido especialmente prolífico en esta alarma). Sin duda, para lo bueno y para lo malo, en la tecnología del procesamiento del lenguaje natural sí va a haber un antes y un después marcado

por esta última generación de los llamados *modelos masivos del lenguaje* (LLM por las siglas de *large language models*). Pero seguramente no habrá tal cambio radical, en contra de lo que han sugerido numerosos autores, en lo que respecta a nuestro conocimiento del lenguaje y de la mente humanos.

El cerebro humano y la habitación china

Continuemos con otra prueba de la utilidad práctica de ChatGPT, puesto que en este punto pensaba desarrollar este ensayo introduciendo el concepto de “habitación china” de John Searle, pues viene al pelo cuando se discute qué tipo de conocimiento tiene realmente un sistema de IA como este modelo del lenguaje. Para ello, tendría que pensar qué voy a decir sobre la habitación china y cómo, recordar dónde leí sobre esa idea de Searle, y quizá buscar en la estantería algún libro para refrescar el concepto y su formulación (creo que era uno de Crick... ¿O era Damasio?). Pero ahora, en lugar de pensar, ponernos de pie para tomar un volumen de la estantería, consultarlo, volver a pensar y sentarnos a escribir, también podemos teclearle a ChatGPT “describe qué es la habitación china de Searle” y, de forma gratuita e instantánea (2-3 segundos), nos hace el trabajo:

La habitación china es un experimento mental propuesto por el filósofo John Searle en 1980 para cuestionar la capacidad de las máquinas para entender y procesar información de la misma manera que lo hacen los seres humanos. En el experimento, se imagina a una persona encerrada en una habitación que no habla chino, pero que tiene a su disposición un conjunto de reglas que le permiten manipular símbolos en chino en respuesta a ciertas entradas. Por ejemplo, si se le presenta una pregunta en chino, la persona consulta sus reglas y produce una respuesta en chino sin tener una comprensión real del lenguaje. Searle argumenta que, aunque la persona puede producir respuestas que parezcan adecuadas a los hablantes nativos de chino que interactúan con ella, en realidad no comprende lo que está diciendo ni tiene una verdadera capacidad para procesar información en chino. Del mismo modo, argumenta que una máquina que imita este tipo de procesamiento de símbolos no puede tener una comprensión real del lenguaje o del mundo en el mismo sentido que los seres humanos. La habitación china se ha utilizado como un ejemplo popular en la filosofía de la mente y la inteligencia artificial para ilustrar las limitaciones de la programación basada en reglas y la necesidad de desarrollar nuevas teorías de la cognición y el procesamiento del lenguaje.

Por supuesto, ChatGPT no sabe quién era Searle, qué es el chino ni hasta qué punto nuestros propios cerebros, que sí saben esas cosas, realmente lo saben. Mi argumento, en adelante, será que ChatGPT, en lo que respecta al conocimiento

del lenguaje, es una “habitación china” y tiene poco que enseñarnos sobre la naturaleza del lenguaje humano y la mente que lo construye y usa.

Como era de esperar, algunos filósofos han alegado que el experimento mental de Searle no sirve para rechazar la capacidad de conocimiento de la IA, dado que bien pudiera ser que nuestros propios cerebros sean también habitaciones chinas. Deberíamos entonces concluir que, aunque hubiera un hablante de chino de verdad dentro de la famosa habitación, su cerebro funcionaría como la habitación china en la que está. De hecho, no cabe duda de que, en un cierto nivel, nuestro “conocimiento”, sea de Searle, del chino o de nosotros mismos, estará soportado por tejidos o grupos de células que se comportan como el individuo de la habitación china, o como el propio algoritmo que anima a ChatGPT: moviendo pedazos de información incomprensible de un sitio a otro. Pero, aunque eso sea cierto, no significa necesariamente que no podamos decir que “sabemos” cosas y que “sabemos que las sabemos”.

En todo caso, más allá de las paradojas y misterios profundos de la consciencia y de la naturaleza misma del conocimiento, la cuestión relevante ahora es si ChatGPT tiene un conocimiento del lenguaje análogo o comparable al que tenemos los seres humanos. Los muchos científicos que creen que sí son los que defienden que ChatGPT es un modelo útil para explicar mejor cómo es posible que podamos aprender y usar la lengua del entorno (e incluso algunos, como Steven Piantadosi, afirman que estos sistemas son auténticas teorías del lenguaje),¹ mientras que los científicos que creen que no (por ejemplo, Chomsky, Roberts y Watumull)² rechazan que los modelos masivos del lenguaje sean relevantes para comprender mejor nuestra capacidad del lenguaje, por mucho que sí sean útiles como productos de ingeniería.

Para entender cómo pueden existir esas discrepancias es relevante que primero sepamos cómo funcionan realmente estos sistemas de generación de lenguaje. Solo después podremos valorar si lo que sabemos sobre cómo el cerebro humano aprende y usa el lenguaje es comparable o equivalente, o no lo es.

¿Cómo funciona ChatGPT?

La lingüista Emily Bender ha popularizado la denominación de *loros estocásticos* para los modelos masivos del lenguaje porque la expresión reúne los dos rasgos esenciales de su funcionamiento: no entienden lo que dicen y se basan esencialmente en calcular las probabilidades que tiene una palabra de aparecer después de otra.

1 Steven Piantadosi, “Modern language models refute Chomsky’s approach to language”, marzo de 2023, <https://lingbuzz.net/lingbuzz/007180>.

2 Noam Chomsky, Ian Roberts y Jeffrey Watumull, “The false promise of ChatGPT”, *The New York Times*, 8 de marzo de 2023, accesible en: <https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html>.

Y, en efecto, lo que hace ChatGPT realmente es calcular la probabilidad de que aparezca una palabra gráfica teniendo en cuenta las palabras gráficas que ya lleva encadenadas.³

He usado la expresión “palabra gráfica” porque, a diferencia de lo que sucede con el lenguaje humano, ChatGPT no opera con auténticas unidades lingüísticas (es decir, morfemas y lexemas, entidades con significado), sino con *tokens*, esto es, secuencias de combinaciones muy frecuentes de letras. Los programadores de OpenAI han descubierto que los resultados son mejores con estos *tokens*, que en inglés son secuencias de unos cuatro caracteres, de manera que un *token* equivale en realidad a unos “tres cuartos” de palabra gráfica en inglés (100 *tokens*=75 palabras gráficas). Los *tokens* que maneja ChatGPT son, de alguna manera, el equivalente en la lengua escrita de las sílabas de la lengua oral: secuencias recurrentes de unidades básicas sin correlación con el significado. No hay, que se sepa, una motivación teórica detrás de esa práctica, sino una mayor eficacia y flexibilidad en la tarea de predecir la próxima *palabra* (entendida a partir de ahora como “secuencia frecuente de caracteres”) teniendo en cuenta prácticamente todo lo que se ha escrito en inglés y estaba digitalizado.

Pero ¿de dónde salen las probabilidades que el programa emplea para valorar qué palabra escribir a continuación de la que acaba de escribir? No se trata de calcular la probabilidad de la siguiente palabra teniendo en cuenta la última usada, sino teniendo en cuenta todas las ya empleadas (de modo contrario no se generaría texto coherente y con apariencia natural, sino una secuencia confusa y repetitiva).

Sin embargo, según los cálculos de Wolfram, simplemente no hay suficiente texto en inglés escrito en el mundo para poder calcular las probabilidades de grupos de palabras. Partiendo de un léxico de unas 50.000 palabras (el número aproximado de *tokens* que usa ChatGPT), las combinaciones posibles de, por ejemplo, secuencias de tres palabras pasan de sesenta trillones; para combinaciones de veinte palabras el número de posibilidades es mayor que el número de partículas del universo, luego no hay —ni nunca podría haber— suficiente texto escrito en el mundo para que se pudieran calcular esas probabilidades.

Por tanto, lo que hace ChatGPT para instanciar su habilidad básica de adivinar cómo continuar el fragmento de texto que se le da es usar un *modelo* (una enorme función matemática) que permite estimar las probabilidades de que aparezcan ciertas secuencias de palabras tras ciertas

secuencias de palabras, *aunque esas secuencias nunca hayan aparecido* en el corpus de entrenamiento. Así, ChatGPT es esencialmente un modelo matemático de la distribución de los *tokens* en el enorme corpus con el que se ha entrenado.⁴

Para conseguir esto, ChatGPT está formado por una enorme red neural que ha sido entrenada con una cantidad inmensa de textos (un corpus de un tamaño de unos 500.000 millones de palabras). Una clave del entrenamiento de esa red neural para que aprenda a continuar textos que tengan apariencia coherente es cebarla con textos incompletos (a los que se les ha quitado un fragmento del final) y usar como *output* de objetivo del entrenamiento el texto completo, de manera que se pasan los ejemplos de entrenamiento hasta que la red proporciona el texto completo original (o uno muy similar) y entonces se fijan los pesos o valores numéricos de cada nodo de la enorme red (esos pesos son los que permiten luego hacer los cálculos de probabilidades).

Cada palabra/*token* (del total de 50.000) está caracterizada por una lista de números (las palabras que tienden a aparecer en los mismos contextos tienen números más parecidos). Partiendo de ahí, también se pueden caracterizar secuencias frecuentes de palabras por listas de números. Y eso es lo que hace ChatGPT cuando interactúa con el usuario: toma el texto que se le da (el *prompt* que introduce el usuario) y lo convierte en un vector de números que procesa. Como hay unas 50.000 palabras en el diccionario, lo que obtiene es una lista de 50.000 números que dan las probabilidades de cada una de las palabras posibles. De forma muy ingeniosa, ChatGPT no usa siempre la palabra más probable, sino que aleatoriamente escoge de vez en cuando una de menor probabilidad (dentro, por supuesto, de la parte alta de la lista). Al parecer, eso tiene como efecto que los textos sean menos “planos” y más originales. Esta “libertad” de la tiranía probabilística sin duda contribuye a la apariencia más humana de la conversación con ChatGPT con respecto a otros competidores y a versiones anteriores.

Según Wolfram, la red neural tiene unos 175.000 millones de “pesos” asignados y cada vez que genera un *token* tiene que hacer cálculos con todos esos pesos (Wolfram se refiere a GPT-3, la versión previa a la GPT-4 que incorpora ahora ChatGPT). Así, cada vez que añade una palabra al texto que genera, el sistema ha tenido que hacer unos 175.000 millones de cálculos numéricos (esto es, aplicar a las matrices numéricas de *input* los pesos de cada una de las conexiones para generar el *output*).

Las siglas GPT están por *generative pre-trained transformer* (transformador generativo preentrenado) y la última parte es, al parecer, la tecnología más novedosa. El llamado

3 En esta exposición del funcionamiento de ChatGPT sigo de cerca la descripción (mucho más detallada y técnica) del físico, informático y gurú de la IA Stephen Wolfram en su escrito al respecto: “Whats is ChatGPT doing... and why does it work?”, accesible en <https://writings.stephenwolfram.com/2023/02/what-is-ChatGPT-doing-and-why-does-it-work/>.

4 Recuérdese que un modelo es un mecanismo para calcular ciertos valores sin tener que medirlos uno a uno, como cuando usamos una función para representar la trayectoria aproximada de un objeto.

transformer es el recurso de arquitectura de redes neurales que permite que la predicción de las palabras siguientes tenga en cuenta las que ha ido procesando antes. Como ChatGPT no sabe de qué está hablando, ni tiene ninguna intención comunicativa, los programadores han encontrado que el medio para que mantenga cierta coherencia es volviendo atrás y recalculando las probabilidades, de unas 50.000 opciones cada vez, de la siguiente palabra. La probabilidad de la siguiente palabra también estará afectada por la última que el propio sistema ha añadido, y otra vez se repite el proceso. En algún momento, ChatGPT encuentra un punto y entonces termina la frase y empieza otra, o se queda “callado”.

La otra característica esencial de ChatGPT, y que está en buena medida detrás de su éxito en la creación de texto coherente (y en convencer al usuario de que realmente está dialogando con él) es que también se ha entrenado la red neural con *feedback* de usuarios humanos. Los lectores humanos detectan divagaciones, repeticiones y pérdidas de coherencia (o faltas de educación) y valoran esos resultados, de manera que el sistema reajusta sus operaciones en función de esas valoraciones. Así, se crea otra red neural auxiliar que aprende de esas valoraciones y modula el *output* de la red principal.

Una idea esencial que ya podemos apreciar es que ChatGPT no opera con unidades lingüísticas, sino con grupos de caracteres, esto es, letras. Por tanto, no tiene manera (ni capacidad) de vincular las secuencias de letras con significados de ningún tipo. Lo único que “sabe” ChatGPT es con qué frecuencia aparecen los grupos de caracteres al lado de otros grupos de caracteres, y en ello se basa para generar cadenas “plausibles” de caracteres, incluyendo los espacios en blanco entre ellos. Aunque pueda sonar despectivo, se merece el apodo de *loro estocástico*.⁵ Todo el significado que puedan tener sus respuestas es el que aporta el usuario que las lee (y que, inevitablemente, tiende a atribuir a su inorgánico interlocutor).

Muchos de los lingüistas y psicólogos que afirman que ChatGPT es un modelo del lenguaje humano que nos puede permitir entenderlo mejor (y que incluso puede refutar algunas aproximaciones científicas al mismo) parecen pasar por alto que precisamente todo el enorme esfuerzo de entrenamiento y computación que implica ChatGPT (incluyendo un enorme consumo energético) se hace precisamente para poder generar texto aparentemente coherente *sin entender lo que se dice*. Así, salvo desde el punto de vista cínico que dice que nosotros tampoco entendemos lo que decimos (que somos habitaciones chinas), es difícil imaginar por qué alguien podría pensar que este programa es un modelo realista de lo que hacemos los humanos cuando hablamos,

dado que nosotros, aunque no sepamos cómo, sí entendemos, la mayoría de las veces, lo que decimos.

El lenguaje humano según ChatGPT

En realidad, no es tan difícil imaginar por qué hay autores que sí creen que ChatGPT es un modelo del lenguaje humano: porque desconocen cómo es el lenguaje humano. El razonamiento típico sería, más o menos, así: como ChatGPT produce oraciones gramaticales y solo sabe de probabilidades de coaparición de palabras, entonces la gramática de las lenguas es una cuestión de probabilidad de aparición de unas palabras al lado de otras. Problema resuelto.

Pero para validar esta lógica hay que asumir que ChatGPT crea oraciones gramaticales como lo hacen los seres humanos. Y esto es mucho asumir. Aunque ChatGPT y sus parientes se denominan *modelos de lenguaje*, lo que pretenden modelizar no es el lenguaje humano entendido como una capacidad cognitiva (esto es, el sistema de conocimiento o facultad que nos permite aprender y usar una lengua cualquiera), sino la probabilidad de aparición de grupos de caracteres junto a otros grupos de caracteres, basándose en las frecuencias observadas en los textos de entrenamiento. Y eso no es un modelo del lenguaje humano.

Para entender mejor por qué no lo es, debemos prestar atención a una distinción crucial que los autores mencionados tienden a pasar por alto: el lenguaje visto desde el punto de vista *interno* (el lenguaje como capacidad) y el lenguaje visto desde el punto de vista *externo* (el lenguaje como producto).

Pensemos por un momento qué es realmente el inglés (por usar como ejemplo la lengua en la que se ha entrenado a ChatGPT centralmente). Según la perspectiva del lenguaje interno (la perspectiva de la lingüística como ciencia cognitiva), el inglés es un sistema de conocimiento que se ha desarrollado en el cerebro de las personas que hablan inglés (por simplificar la exposición, pensemos en los hablantes nativos de esta lengua). Llamemos a ese sistema de conocimiento, siguiendo a Noam Chomsky, la lengua-*i* (*i* de interna e intensional). Una lengua-*i* consiste mínimamente en un repertorio de conceptos o significados asociados a cadenas de fonemas (palabras en sentido estricto) y un conjunto de reglas gramaticales para construir oraciones con ellas. Por su parte, la lengua-*e* (*e* de externa y extensional) sería la suma de todas las oraciones producidas por los hablantes de una lengua dada, por ejemplo, el corpus de todo lo que se ha escrito en inglés en los últimos trescientos años (al que aún habría que añadir todas las oraciones que los miles de millones de hablantes que ha habido de esa lengua en la historia han pronunciado sin llegar a escribirlas nunca; sin duda, muchísimas más que las que han sido escritas). Los lingüistas internistas, con Chomsky a la cabeza, creen que el objeto de estudio de la lingüística como ciencia es la lengua-*i* (el sistema de conocimiento que hay en los cerebros

⁵ Como han señalado algunos comentaristas, en realidad la expresión es ofensiva para los loros, porque ellos sí son realmente inteligentes.

de las personas), mientras que la lengua-e es un objeto inabarcable e incoherente. Donde existe realmente *el inglés* es en los cerebros de sus hablantes, no “ahí fuera”.

Los lingüistas externistas creen, por el contrario, que lo que realmente existe es la lengua-e, una especie de constructo sociocultural que pasa de generación en generación, siendo la lengua-i una copia imperfecta de la lengua-e en los cerebros de las personas, desarrollada en los mismos, como decía Ferdinand de Saussure, por la práctica del habla. Independientemente de quién tenga razón, lo que parece claro es que ChatGPT es un modelo de la lengua-e, no de la lengua-i.

Es un hecho objetivo, en todo caso, que cualquier persona puede crear y entender centenares de miles de oraciones distintas, que no ha oído nunca antes. Se ha calculado que cuando estamos hablando podemos elegir unas diez palabras distintas después de la última que hemos usado. En algunas ocasiones solo vale una (detrás de *sin ton ni* solo cabe *son*), en otras ocasiones sirven centenares de ellas (como detrás de *me comería un*), siendo diez la media, según cálculos que menciona el psicolingüista Steven Pinker. Si una persona puede hacer oraciones de veinte palabras (cualquiera de este texto tiene bastantes más) y detrás de cada palabra que usa puede escoger de entre otras diez, entonces el número de oraciones diferentes que una persona puede crear y entender es de 10^{20} (o sea, cien trillones de oraciones distintas), y eso limitando absurdamente la extensión máxima de una oración a veinte palabras. De nuevo según cálculos de Pinker, solo esos cien trillones de oraciones de veinte palabras requerirían de una infancia de unos cien billones de años para aprenderlas (asumiendo unos cinco segundos para aprender cada oración). Dado que esto es obviamente absurdo, siendo además que el cerebro humano “solo” tiene unos 80.000 millones de neuronas, los lingüistas internistas asumen que los hablantes disponen de un sistema generativo de reglas, mecanismos o principios que les permiten crear cada una de esas cien trillones de oraciones distintas de veinte palabras que podrían pronunciar y entender si vivieran lo suficiente.

¿Dispone ChatGPT de ese sistema? No hay evidencias de que así sea. Es importante tener en cuenta que una parte central del mecanismo generativo del lenguaje humano es la estructura jerárquica de constituyentes, que básicamente implica que las oraciones del lenguaje humano (del inglés o de cualquier lengua) no son secuencias lineales de palabras (aunque así aparecen en el habla o en la escritura), sino agrupamientos de palabras en “cajas” o constituyentes, contenidos unos en otros, que determinan el significado que realmente tienen las oraciones. La frase nominal *El traficante de armas de Rusia* es ambigua porque la secuencia lineal puede reflejar tanto una estructura en la que *de Rusia* modifica a *armas* (en cuyo caso entendemos que las armas son de Rusia, aunque el traficante puede ser de otro sitio), como

una en la que *de Rusia* modifica al constituyente formado por *traficante de armas*, en cuyo caso entendemos que quien es de Rusia es el traficante, pudiendo ser las armas de otro sitio. En el primer caso, tenemos la estructura simplificada de (1a), en la que la secuencia *armas de Rusia* forma un constituyente, y en el segundo, la de (1b), en la que no lo hace:

- (1) a. *El traficante de [armas de Rusia]*
b. *El [traficante de armas] [de Rusia]*

La estructura sintáctica es inaudible e invisible (y no está en las ondas sonoras ni en las secuencias lineales de palabras que forman los textos), pero es crucial para interpretar el significado de las secuencias de palabras que nos llegan y para construir los pensamientos complejos y proposiciones que los humanos creamos y (a veces) comunicamos.

En (1b) *de Rusia* no puede modificar a *armas* porque no tiene acceso estructural a ese nombre (de quien depende sintácticamente es del nombre *traficante*), mientras que en (1a) *de Rusia* tiene que modificar a *armas* y no puede acceder a *traficante*. No hay manera de explicar estas asimetrías, ubicuas en el lenguaje humano, analizando la sintaxis como una secuencia lineal basada en el encadenamiento probabilístico de palabras, que es lo que hace ChatGPT.

Ya en 1957 Chomsky propuso la célebre oración, absurda pero perfectamente gramatical, *Las ideas verdes incoloras duermen furiosamente* (*Colorless green ideas sleep furiously*) precisamente para mostrar que la generación del lenguaje humano no se puede modelizar como un sistema de encadenamiento de la palabra más probable (más esperable) respecto de la anterior, sino empleando mecanismos formales independientes del significado de cada palabra. Esta característica de diseño del lenguaje, esto es, un mecanismo generativo que crea estructuras jerárquicas usando categorías sintácticas (nombre, verbo, conjunción, etc.) independientemente del contenido léxico de cada ítem (*mesa, comer, que, etc.*) es lo que realmente puede ayudar a explicar la capacidad de los hablantes de lenguas humanas de crear y comunicar ideas nuevas y albergar pensamientos libres del control del estímulo o de las interacciones con otras personas. Pero es importante observar que una sintaxis así es necesaria para crear significados (como los dos de *El traficante de armas de Rusia*), pero no para encadenar palabras de izquierda a derecha. ChatGPT no usa una estructura de constituyentes simplemente porque no la necesita, dado que no tiene nada que decir.

Hemos considerado cómo funcionan estos modelos masivos de lenguaje y hemos visto que no hay indicios objetivos de ningún tipo de estructura gramatical en esa tarea (y, por supuesto, nada de semántica ni de fonología). Si retomamos la distinción entre lengua-i y lengua-e, estamos en disposición de poder afirmar que de lo que es un modelo ChatGPT es de la lengua-e, no de la lengua-i.

Por ello se puede afirmar que es una eficiente herramienta de ingeniería, un logro notable de la IA, pero no que sea un modelo del lenguaje humano, esto es, de la facultad que nos permite crear pensamientos nuevos y originales y volcarlos en sonidos (y estos en caracteres, como los que maneja ChatGPT). Por así decirlo, los seres humanos crean secuencias de caracteres como consecuencia de combinar previamente conceptos empleando una sintaxis recursiva jerárquica, mientras que ChatGPT crea secuencias de caracteres a partir de otras secuencias de caracteres. Es un loro estocástico en una habituación china.

Solo quien realmente crea que las lenguas humanas son conjuntos de textos planos (secuencias lineales de palabras) y que, por tanto, crea que los niños aprenden a hablar memorizando y repitiendo las oraciones que han oído, o calculando la probabilidad de la siguiente palabra, sin entender su significado, puede seguir manteniendo que ChatGPT y sus parientes son modelos del lenguaje humano.

Por decirlo de otra manera: ChatGPT es un modelo matemático del producto del uso del lenguaje humano, no de la capacidad de producir el lenguaje humano. De hecho, podría decirse que afirmar que ChatGPT habla (o escribe) en inglés (o en cualquier otra lengua) es lo mismo que igualar el procedimiento que usó Velázquez con el pincel para pintar *Las Meninas* con el del fotógrafo que reproduce el famoso cuadro en un mural del mismo tamaño. Puede que el resultado sea superficialmente similar, pero el proceso de creación es radicalmente distinto, y no aprenderemos mucho de la técnica creativa de Velázquez analizando el proceso de revelado de una fotografía de *Las Meninas*.

¿Es ChatGPT una teoría de la adquisición del lenguaje?

No todos los científicos que opinan que ChatGPT es un modelo del lenguaje humano (capaz incluso de decidir qué teorías del lenguaje son más adecuadas) son “terraplanistas” (en el sentido de que niegan la existencia de la estructura jerárquica de constituyentes). Algunos, aunque admiten que el lenguaje humano no consiste en ristas estocásticas de palabras, sostienen que ChatGPT ha debido, *de alguna manera*, descubrir los principios para crear esas estructuras a partir del corpus de entrenamiento. Así, a la vez que asumen que ChatGPT es un buen modelo de la mente humana (aunque quienes lo diseñaron no pretendían hacer tal cosa), afirman que ha desmentido la hipótesis (habitualmente asociada a Chomsky) de que existe un sesgo innato para el lenguaje en los seres humanos.

El razonamiento discurre, más o menos, así: puesto que ChatGPT no ha sido equipado con los principios y mecanismos que algunos lingüistas asumen que tienen que ser innatos para poder explicar el aprendizaje del lenguaje (por ejemplo, los mecanismos generativos para producir estructuras sintácticas jerárquicas recursivamente), entonces,

como ChatGPT *ha aprendido* esos principios, la aproximación innatista es falsa.

Otro prodigio de la ingeniería del procesamiento del lenguaje natural, Google Translator, vierte así al español (con ligeras correcciones mías) la afirmación en esta línea del psicólogo cognitivo Seven Piantadosi: “El surgimiento y el éxito de los grandes modelos del lenguaje socava prácticamente todas las afirmaciones sobre el carácter innato del lenguaje que ha propuesto la lingüística generativa. El aprendizaje automático moderno ha subvertido y sobrepasado todo el marco teórico del enfoque de Chomsky.”⁶

No deja de ser curioso (y revelador) que el argumento que emplea Piantadosi para negar que estos sistemas sean loros estocásticos sea exactamente el mismo que empleaba Chomsky en los años cincuenta y sesenta del siglo XX para afirmar que si los humanos aprendieran a hablar simplemente imitando o copiando del habla del ambiente, no se podría explicar su capacidad de crear oraciones nuevas (y aquí cito literalmente, sin traducir, para evitar sospechas de exageración): “If models only repeated what they had seen before, they would not be able to generate anything new, particularly complex sentence structures that are grammatical and coherent.”⁷

Y también resulta curioso pretender que sea un modelo denominado *transformador generativo pre-entrenado* (GPT) el que supuestamente habría demostrado que la llamada *gramática generativo-transformacional* es una empresa científica ilegítima.

Por supuesto, aunque sería sorprendente, la aproximación innatista podría ser falsa. Pero, desde luego, ChatGPT no presenta ninguna evidencia en esa dirección, simplemente porque no es posible afirmar que haya aprendido lengua alguna. Por muy sofisticado que parezca en sus respuestas, ChatGPT no entiende, ni pretende entender, ni una palabra de lo que escribe. No ha interiorizado ninguna gramática con la que crear y combinar conceptos, ni tiene conceptos, ni, por tanto, ha adquirido ninguna lengua por medio de aprendizaje, tal y como hacen los seres humanos. Como antes sugería, para que ChatGPT fuera un modelo del lenguaje humano, las personas deberían emplear el lenguaje sin entenderlo, usando secuencias lineales más probables en ciertos contextos, en vez de construir pensamientos nuevos empleando la gramática interna y luego convirtiéndolos en cadenas de palabras para comunicarse.

Veamos, para concluir, un ejemplo de cómo sabemos que ChatGPT en realidad no ha adquirido una lengua humana y, por tanto, no ha podido deducir de la estadística de las palabras las reglas de la sintaxis. El lingüista computacional Roni Katzir (en una contundente respuesta a

6 Piantadosi, *op. cit.*, p. 1.

7 Que viene a decir que “si los modelos se limitaran a repetir lo que han visto antes, no serían capaces de generar nada nuevo, especialmente complejas estructuras oracionales que sean gramaticales y coherentes”. Piantadosi, *op. cit.*, p. 5.

Piantadosi)⁸ le pregunta a ChatGPT cuál de las dos oraciones siguientes es más aceptable (usando equivalentes en inglés):

- (2) a. *La persona que María se encontró ayer y de la que Juan le hablará a Pedro mañana ya ha llegado.*
 b. *La persona que María se encontró ayer y de la que Juan le hablará de una idea nueva a Pedro mañana ya ha llegado.*

Aunque ambas son oraciones complejas, cualquier hablante nativo del español que tenga nueve o diez años sabe que la primera es gramatical y la segunda no lo es, a pesar de que ChatGPT considera que la segunda es más aceptable que la primera.

En esencia, lo que sucede en este tipo de construcción es que cuando se coordinan dos oraciones de relativo (en este caso que modifican a *la persona*), el “hueco” que deja el pronombre relativo tiene que ser igual en las dos (nótese que podemos decir *La mujer que vino y que Pedro conocía*, pero no *La mujer que vino y que Pedro conocía a Luis*). Así, vemos que en (2a) falta el objeto de *se encontró* y el objeto de *hablará*, mientras que en (2b) falta el objeto de *se encontró*, pero no falta el objeto de *hablará* (se ha añadido un *de una idea nueva* que rellena ese hueco). Por supuesto, eso crea una oración agramatical (hay dos complementos de *hablará*), aunque ChatGPT la

prefiere, simplemente porque, aunque no entiende qué significa, sabe que es muy probable que detrás de *hablará* aparezca *de una nueva idea*. Cabe señalar que la restricción que aparece en el ejemplo de (2a), esto es, que si en un coordinando hay un “hueco” también lo tiene que haber en el otro, es una restricción formal a la sintaxis de las lenguas humanas ya descrita en los años sesenta del siglo XX por los lingüistas generativistas, una restricción que los niños humanos adquieren sin esfuerzo y con muchísimos menos datos de entrenamiento que ChatGPT.

La conclusión, por tanto, es doble: ChatGPT no ha interiorizado los mecanismos y principios que los humanos emplean para construir oraciones y, por tanto, en modo alguno se puede afirmar que este modelo masivo del lenguaje haya demostrado que las propuestas de la gramática generativa (o de cualquier teoría de la adquisición del lenguaje) sean falsas. Simplemente, los sesgos inductivos que presentamos los humanos y los que presenta ChatGPT son totalmente distintos, por lo que no tiene sentido considerar al segundo un modelo de los primeros. Como concluye Katzir, esta vez traduciendo con un cerebro humano, “una cosa es usar estos modelos para escribir poemas entretenidos y narraciones cortas, y otra muy diferente usarlos para comprender la facultad humana del lenguaje en lugar de hacer lingüística de verdad”. ~

⁸ Roni Katzir, “Why large language models are poor theories of human linguistic cognition. A reply to Piantadosi”, abril de 2023, accesible en: <https://lingbuzz.net/lingbuzz/007190>.

JOSÉ LUIS MENDÍVIL GIRÓ es catedrático de lingüística general en la Universidad de Zaragoza.



LETRAS LIBRES



¡Suscríbete! 12 números €50