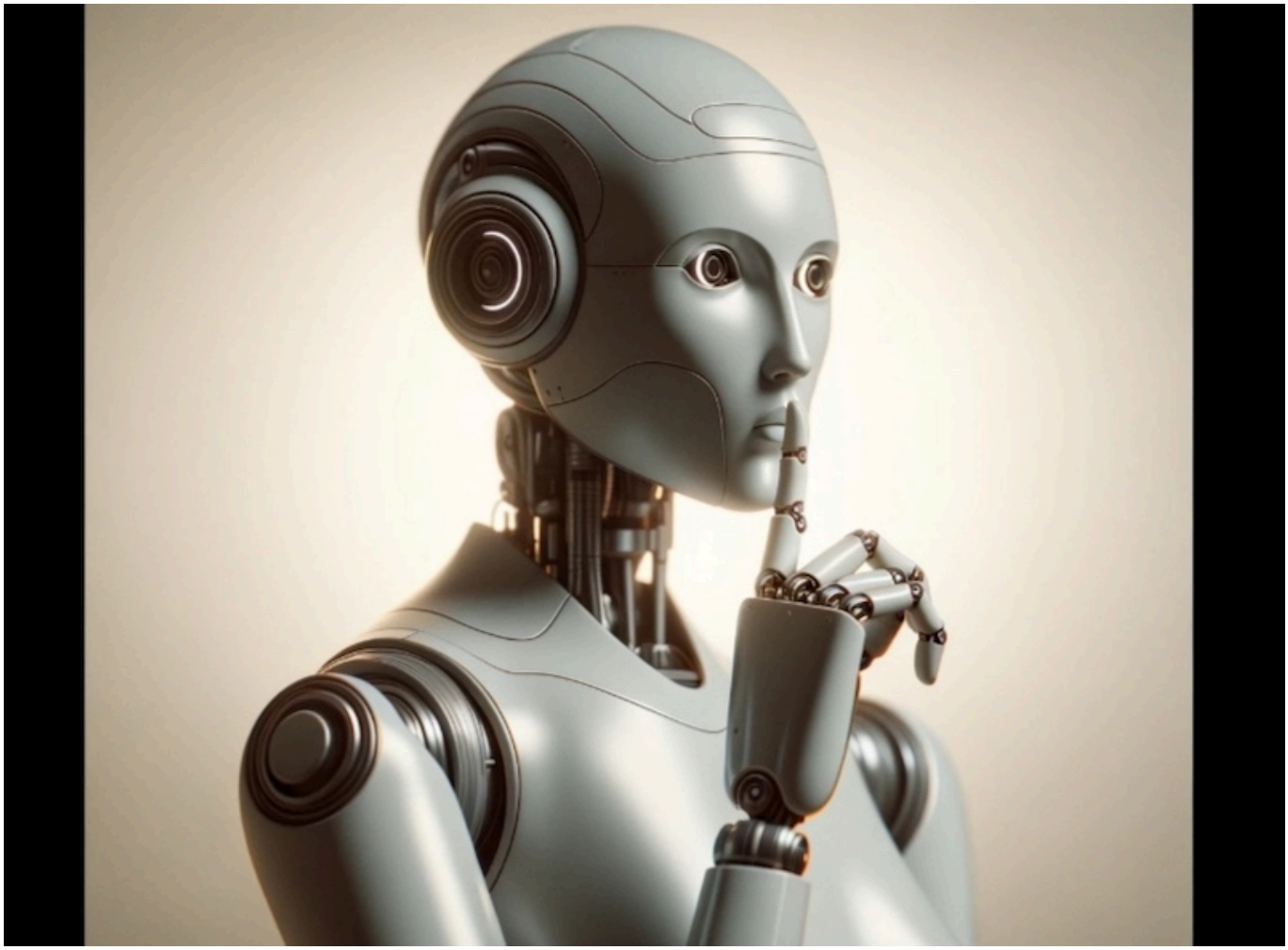
RdL Revista de Libros

Entre la exaltación y el miedo: ¿la IA nos quitará la palabra?

Violeta Demonte
20 febrero, 2024



La inteligencia artificial [IA] y los modelos generativos de lenguaje, los conocidos como «modelos grandes de lenguaje» (MGL; traducción de *Large Language Models*), son ahora mismo un tema estelar para los medios de comunicación, la filosofía y, claro está, las investigaciones académicas y las empresas tecnológicas. Las masivas especulaciones y explicaciones sobre este asunto se mueven entre la exaltación de un futuro en el que —dicen— no nos reconoceremos, el miedo a la suplantación de los humanos por las máquinas, y un cierto escepticismo metafísico o, mejor, prudencia razonada, con los pies en la tierra, donde querría colocarme. Me ceñiré a una cuestión concreta y no sencilla, pero muy recorrida por la opinión común: ¿la IA, los grandes modelos que producen lenguaje fluido y coherente, son el anticipo de un futuro en el que dialogaremos e interactuaremos con objetos hablantes emanados de algoritmos?, ¿estos objetos podrán ser superiores a nosotros intelectual y lingüísticamente?Estoy muy agradecida a José Manuel Sánchez Ron, quien me ofreció dar una charla sobre esta cuestión en el Museo de Navarra y me sugirió el título que ahora adopto como subtítulo de este texto. A José Antonio Millán, sabio amigo y experto, le debo sugerencias y correcciones muy valiosas para mí. Cristina Aranda me invitó a debatir con ella estos sobre estos asuntos en El Ateneo de Madrid, fue muy enriquecedor. A Elena Castroviejo le debo ideas, correcciones, diseño: una inestimable ayuda.

El futuro no tiene otra realidad que la esperanza presente, y el pasado es la memoria presente, parece que afirmó Borges. El futuro son conjeturas, y en el caso de instrumentos tan sofisticados y preliminares como los de la inteligencia artificial, nos sitúa en el terreno de la ciencia ficción. También, ese futuro se percibe desde un presente en el cual no disponemos de toda la información que sería necesaria para hacer conjeturas plausibles. Un presente en el que hay más preguntas abiertas que certezas por contrastar: los conocimientos sobre la biología del cerebro, y más aún sobre la estructura biológica y mental del lenguaje, no son aún redondos; ni lo son los desarrollos de las ciencias cognitivas, la neurociencia y la lingüística. Y los avances de los algoritmos capaces de producir escrituras o interacciones lingüísticas apenas cubren una ínfima parte de lo que sería necesario concebir.

Para responder a la pregunta que antes formulaba, examinaré en este texto algunas de las grandezas y los límites de estos MGL que parecen ser los mejores candidatos a seres lingüísticamente competentes. Tras detenerme un momento en la equívoca noción de inteligencia artificial, haré primero un breve estado de la cuestión acerca de los sistemas que utilizan la IA basada en el Procesamiento de Lenguaje Natural (PLN) y su gran valor facilitador de las tareas de los lingüistas. Presentaré a continuación algunas de mis conversaciones con el ChatGPT 3.5 para indagar lo que sabe sobre lenguaje (bastante, pero nada que ver con nosotros) y los aciertos y desaciertos en que incurre. Me extenderé después sobre lo que un lingüista, filósofo o científico cognitivo considera que debería tener un sistema operativo (un algoritmo) para que pueda quitarnos la palabra. Estas no son reflexiones de una ingeniera, sino de una persona de humanidades, o humanidades cognitivas si se prefiere. Pero esa es la cuestión: los que sabemos algo de estas cosas, por ejemplo del lenguaje, debemos quizá dialogar con quienes hacen los algoritmos y con quienes presumen de sus virtualidades revolucionarias.

1. La equívoca noción de inteligencia artificial

Lo primero que conviene aclarar cuando nos acercamos a este asunto es el uso y la comprensión equívoca de la palabra inteligencia, tanto por los expertos (para ellos es una expresión técnica, claro) como por los medios de comunicación, sobre los «artefactos» que hacen cosas que podrían simular «humanidad». El término parece que fue acuñado en 1956 (aunque ya Turing había hablado de computación inteligente) en una conferencia de verano en Darmouth College, organizada a propuesta de John McCarthy (profesor de Darmouth), junto con Marvin Minsky (matemático del MIT), Claude Shannon (matemático e ingeniero de los Laboratorios Bell, uno de los fundadores de la Teoría de la información) y varios otros científicos, para conjeturar sobre la suposición de que, si todos los aspectos del aprendizaje y de la inteligencia pudieran describirse con total precisión, entonces podrían construirse máquinas que los simularan. Inteligencia artificial era un término más eficaz como revulsivo que «sistema operativo avanzado», por caso, y más resonante para las multitudes que precedentes tales como «teoría de la información», «cibernética», «teoría de autómatas», entre otros.

La expresión es equívoca porque estos algoritmos ni son inteligentes, ni son tan artificiales. Naturalmente, estamos lejos de saber con precisión qué es la inteligencia, pero sí sabemos que los llamados seres inteligentes, nosotros mismos, somos capaces de comprender, razonar, pensar, hacer abstracciones, atributos que no parecen poseer los nuevos productos de la IA. En realidad, «no hay evidencia ni matemática, ni física, ni se conoce la existencia de ningún prototipo equivalente a las capacidades pensantes de un cerebro humano» (Etxebarria 2023), y ni siquiera sería acaso inteligente la atractiva OS2, Samantha, de *Her*, que perturba a Theodore y tanto lo ama. Por poner algún ejemplo revelador, un algoritmo dirigido por buenas instrucciones (*prompts*) puede demostrar un teorema pero, según dicen los matemáticos, no puede inventar un teorema. Es cierto que estos algoritmos tienen algo así como inteligencias, mejor sería decir capacidades, específicas, sistemas que generan respuestas, identificaciones que producen avances: identifican áreas en el ADN que pueden causar enfermedades, pueden sintetizar y reconocer voces, pero no poseen inteligencia general. Y no son tan artificiales porque los algoritmos los inventan humanos (aunque puedan automodificarse). Asimismo, no olvidemos que en los modelos generativos trabajan personas (*turks*), humanos subcontratados que revisan sus producciones y evitan que tengan sesgos o digan demasiadas tonterías.

2. Breve estado de la cuestión

Los «modelos grandes de lenguaje» —estas redes neuronales artificiales (algoritmos) generadoras de lenguaje, entrenadas mediante «aprendizaje profundo» (*deep learning*) con miles de millones de datos—, han despertado enorme atención tanto en las compañías tecnológicas grandes y pequeñas como en los medios de comunicación, que hablan de su capacidad de hablar, pensar y razonar, entre otras cosas, y anuncian el dominio del hombre por las máquinas. Han generado, como dice Innerarity (2023), una suerte de «histeria digital», y una fascinación no siempre informada que convierte a Sam Altman (el creador de la empresa Open AI) en una especie nueva de estrella de cine. Naturalmente, las «tecnologías habilitadoras de IA» son un instrumento muy potente para facilitar y extender las tareas de quienes trabajan con el lenguaje. Los lingüistas hemos de estar muy agradecidos a la IA. Operan a gran velocidad sobre enormes bases de datos sobre las que pueden hacer clasificaciones, ordenaciones, reagrupaciones, ediciones, deducir guías, hacer diagnósticos. Seleccionan palabras y las clasifican por similitudes, permiten mejorar los análisis léxicos (Peterson y Potts 2023), hacen «ontologías» (agrupan conceptos de un dominio dado de modo que puedan relacionarse con los de otros dominios). Traducen con un grado alto de corrección (no tanto en el caso de las traducciones literarias). Editan textos antiguos y modernos. Estos sistemas de PLN son buenísimos para ayudarnos, otra cosa será sustituirnos. Y nada de esto es completamente nuevo, el procesamiento del lenguaje natural viene obteniendo estos resultados desde hace ya varias décadas. La lingüística computacional, el aprendizaje profundo, el aprendizaje de las máquinas, la semántica distribucional y las aplicaciones de inteligencia artificial sirven a los sistemas de PLN. Pero lo que hoy mueve a creer que los algoritmos hablan y piensan, como he dicho, es el desarrollo exponencial de los sistemas de PLN que «generan lenguaje»Para empezar quizá por lo más interesante habría que hablar de las tecnologías de imitación de voz. Mai et al. 2023 describen experimentos que muestran que un alto

porcentaje de sujetos experimentales no distinguen la voz humana de la voz tecnológica, incluso aunque hayan sido entrenados para distinguirla. Son sorprendentes también los modelos de reconocimiento de voz. Lamentablemente, no tendré tiempo de detenerme en ellos..

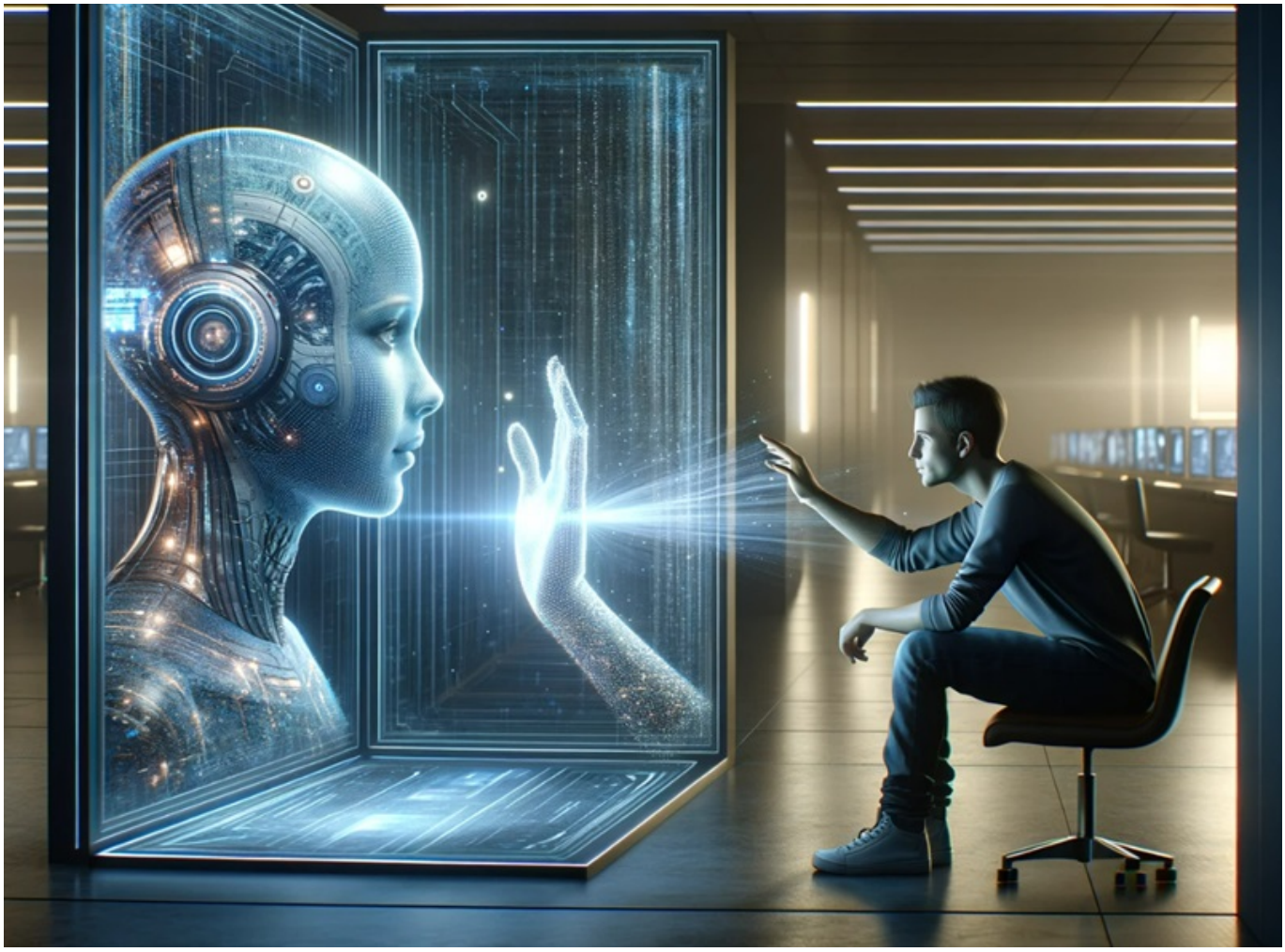


Imagen generada con ChatGPT-4. Humano interactuando con inteligencia artificial.

3. Mis conversaciones con ChatGPT-Open AI-3.5. Elogios y sombras

La expresión modelos grandes de lenguaje se refiere específicamente a los chats «transformadores, generadores preentrenados» (GPT, seria TGP en español). El más conocido, pero no el único, es el ChatGPT de Open AI, cofundado por Sam Altman en 2015 y lanzado al mundo con un éxito rotundo, y acceso gratuito, en 2022. Este ChatGPT y sus similares, además de varias versiones (GPT-2, GPT-3, GPT-4), tienen muchas e interesantes habilidades que ilustraré brevemente. La fundamental: generan textos escritos y habladosChatGPT-Open AI no es el primer chat conversacional y, como sabemos,

coexiste ahora con muchos otros. El primero de todos fue Eliza creado entre 1964 y 1966 por Joseph Wiezenbaum, profesor del MIT, para explorar la interacción entre humanos y máquinas.. También son capaces de responder a «peticiones de información», hacer «resúmenes y sumarios», generalmente sucintos, obvios para el experto, poco profundos en general, con una prosa no excesivamente elegante, como se ve en el siguiente ejemplo, pero siempre fluidos y coherentes:

> **Yo:** ¿En qué modelo de la física se funda la construcción de la primera bomba atómica?
> **ChatGPT-3.5:** La construcción de la primera bomba atómica se basó en principios de la física nuclear y se fundamentó en el modelo de la física conocido como la teoría de la relatividad de Albert Einstein y la física cuántica. En términos más específicos, la teoría de la relatividad proporcionó la famosa ecuación $E=mc2$, que establece la equivalencia entre la energía (E) y la masa (m), indicando que una pequeña cantidad de masa puede convertirse en una gran cantidad de energía.
> Sin embargo, la física subyacente que permitió la creación de la bomba atómica se basó en la fisión nuclear, un proceso en el cual los núcleos de átomos pesados, como el uranio-235, se dividen en fragmentos más pequeños liberando una gran cantidad de energía en forma de radiación y calor. La fisión nuclear es una consecuencia de la física cuántica y se comprendió mejor gracias al trabajo de científicos como Niels Bohr, Werner Heisenberg y Otto Hahn, entre otros.
> El proyecto para desarrollar la bomba atómica durante la Segunda Guerra Mundial, conocido como el Proyecto Manhattan, reunió a científicos destacados en el campo de la física nuclear, como Robert Oppenheimer, Enrico Fermi y Richard Feynman. Fue la combinación de los principios de la física cuántica y las ideas de la teoría de la relatividad lo que permitió comprender y controlar el proceso de fisión nuclear, y finalmente llevar a la creación de la primera bomba atómica en 1945.

Pueden también proporcionar las «referencias» para muchos temas, generalmente muy incompletas y a veces con entradas inexistentes, como sabemos cualquiera que le haya pedido a GPT-Open AI-3.5. una bibliografía. Pueden realizar «traducciones» muy aceptables, establecer «diálogos» con el usuario. Pueden escribir *papers* si son guiados por *prompts* (instrucciones) adecuados. Este chat puede ayudar y está ayudando a los científicos: se usa ChatGPT para hacer preguntas para un experimento, obtener datos, mejorar estilísticamente el inglés académico, construir tablas, hacer estadísticas. Es oportuno recordar por aquí que en la versión más actualizada del ChatGPT-Open AI (GPT-4) este es capaz de trabajar con noticias del día y, por ejemplo, puede resumirte artículos de los periódicos con bastante acierto. Esto no lo hacía GPT-3.5. Pueden ser constructivos, pero también disruptivos: si las instrucciones no son buenas, o si no dispone de información clara, sus respuestas dan lugar a lo que los expertos denominan «alucinaciones» (huecos rellenados al tun tun)Como GPT

Open AI- 4 detecta y utiliza bibliografía real para construir sus repuestas, es de esperar que sus respuestas anticipen mejor y eviten alucinaciones.:

> «Un problema habitual de las herramientas de IA generativa es su tendencia a rellenar los huecos inventándose cosas, un fenómeno conocido como alucinación. Para evitar que se inventara las referencias, el equipo permitió a ChatGPT acceder a motores de búsqueda bibliográfica para generar un artículo con las citas correctas…» (Conroy 2023).

Incluso componen «poemas» que pueden ganar premios literarios, o escriben «narrativas» que simulan el estilo y los temas de muchísimos autores. Estos productos a veces pueden sorprendernos, pero si los miramos con atención no es para tanto. En efecto, las simulaciones de autores suelen ser ramplonas. Veamos algún ejemplo.

> **Yo:** Escribe los tres párrafos iniciales de un cuento en el estilo del relato de Julio Cortázar «El Perseguidor».
> **ChatGPT-3.5:** Las noches en París eran un laberinto de luces y sombras, una danza interminable de destinos entrelazados por las callejuelas empedradas. En aquel rincón olvidado del barrio latino, donde las melodías del jazz se fusionaban con el aroma del café y el humo del tabaco, un misterio aguardaba su turno para revelarse. Era allí, en el sótano oscuro y repleto de notas suspendidas en el aire, donde los sueños de los bohemios tomaban forma y desafiaban al tiempo.
> Un contrabajo reposaba en la penumbra, su silueta parecía un amante abandonado en espera de un toque maestro. Daniel Dufresne, un músico errante de alma inquieta se encontraba frente al instrumento como un titiritero frente a su marioneta. Sus dedos, largos y delicados, acariciaban las cuerdas con una pasión contenida, desentrañando armonías que resonaban en el recuerdo de quienes lo escuchaban. Daniel vivía entre las notas y los espacios, entre el éxtasis y la desesperación, como si su música fuera la única manera de traducir las palabras que su mente nunca pudo articular.
> La ciudad latía afuera, pero en aquel sótano el tiempo parecía detenerse. Un espectro en la forma de saxofonista, conocido como «El Sombra», acechaba los pasillos cercanos, un ser enigmático que se alimentaba de los acordes de Daniel. La obsesión de El Sombra era capturar una nota, un instante de perfección musical que lo liberara de su eterno vagar. Daniel, ajeno a la sombra que lo seguía, continuaba su danza con el contrabajo, sin sospechar que su destino estaba entrelazado con el de su perseguidor…

Escojo casi al azar un trozo de uno de los cuentos del libro *Las armas secretas*: «El Perseguidor», el más experimental, más surrealista quizá de todos ellos, donde el narrador se adentra en la conciencia de Charlie Parker, representado por Johnny, el jazzista (saxofonista) sobre el que versa el cuento. El comienzo de este cuento dice así:

> «Dédée me ha llamado por la tarde diciéndome que Johnny no estaba bien, y he ido en seguida al hotel. Desde hace unos días Johnny y Dédée viven en un hotel de la rue Lagrange, en una pieza del cuarto piso. Me ha bastado ver la puerta de la pieza para darme cuenta de que Johnny está en la peor de las miserias; la ventana da a un patio casi negro, y a la una de la tarde hay que tener la luz encendida si se quiere leer el diario o verse la cara. No hace frío, pero he encontrado a Johnny envuelto en una frazada, encajado en un roñoso sillón que larga por todos lados pedazos de estopa amarillenta. Dédée está envejecida, y el vestido rojo le queda muy mal; es un vestido para el trabajo, para las luces de la escena; en esa pieza del hotel se convierte en una especie de coágulo repugnante.
>
> —El compañero Bruno es fiel como el mal aliento —ha dicho Johnny a manera de saludo, remontando las rodillas hasta apoyar en ellas el mentón. Dédée me ha alcanzado una silla y yo he sacado un paquete de Gauloises. Traía un frasco de ron en el bolsillo, pero no he querido mostrarlo hasta hacerme una idea de lo que pasa. Creo que lo más irritante era la lamparilla con su ojo arrancado colgando del hilo sucio de moscas. Después de mirarla una o dos veces, y ponerme la mano como pantalla, le he preguntado a Dédée si no podíamos apagar la lamparilla y arreglarnos con la luz de la ventana. Johnny seguía mis palabras y mis gestos con una gran atención distraída, como un gato que mira fijo pero que se ve que está por completo en otra cosa; que es otra cosa. Por fin Dédée se ha levantado y ha apagado la luz. En lo que quedaba, una mezcla de gris y negro, nos hemos reconocido mejor. Johnny ha sacado una de sus largas manos flacas de debajo de la frazada, y yo he sentido la fláccida tibieza de su piel…».

En el texto del bot, que no puedo analizar detalladamente, no queda casi nada de Julio Cortázar, no hay relato, no hay descripción de un personaje alcohólico, heroinómano, no hay un recorrido por la decadencia y a la vez la grandeza de los ambientes que describe, no hay argentinismos, no hay metáforas tan sugerentes como las lamparillas o el ojo colgando de un hilo, las caracterizaciones de personajes son banales, no se entrelazan con el ambiente como en el texto real. Sí hay transmisión de sensaciones, una imaginería cursi (el contrabajo como amante abandonado) e intento de crear un clima de misterio que no tiene nada que ver con el cuento, que es de un realismo decadente. En

cambio, reproduce expresiones clichés supuestamente propias de una literatura misteriosa pero mediocre: *rincón olvidado*, *límites difusos, secretos insondables, laberinto de preguntas sin respuestas*. Ignacio Bosque en su diccionario combinatorio *Redes* indica que los adjetivos que más coaparecen con *rincón* son: *alejado, olvidado, ignorado, oscuro, perdido*… con *límite*: *borroso, difuso, frágil, estricto*… con *secreto*: *absoluto, a voces, de confesión, impenetrable, insondable*… Y esas son las combinaciones que encontramos en esta simulación del bot. Por otro lado, en términos de Bringsjord, Bello y Ferrucci (2001), creadores del test Lovelace 2.0 que intenta medir la capacidad creativa, artística, de estos artefactos en producciones de narrativas y pinturas, diríamos que a la imitación de Cortázar le falta lo que su texto sí tiene, conocimiento de sentido común (¿qué hace esa extraña sombra en la simulación?), teoría de la mente (la reacción de Dedeé —en el texto de Cortázar mismo— es un acto inferencial), y razonamiento afectivo (el narrador de «El perseguidor» transmite pena por el saxofonista).

Me interesó también hacerle preguntas para ver cómo estaba de «capacidad gramatical». Pues bien, no comete errores de bulto y puede dar respuestas aproximadas a preguntas gramaticales simples:

> **Yo:** Which is the meaning of «her» in the sentence «Which book did Mary buy after visiting her sister?».
> **ChatGPT-3.5. Segunda respuesta:** In the sentence «Which book did Mary buy after visiting her sister?», the word «her» refers to Mary’s sister. The sentence is asking which book Mary bought after visiting her sister.
> **Respuesta original:** In the sentence «Which book did Mary buy after visiting her sister?», the word «her» refers to Mary’s sister. It means that Mary visited her own sister before buying a book, and the book she bought may or may not be related to the visit to her sister.

La respuesta es una glosa de la oración, reproduce solo uno de los significados del pronombre y, sobre todo en la primera respuesta, antes de que le pidiera que rectificara, introduce datos de contexto que son completamente irrelevantes para entender el significado de *her*. La interpretación de esta oración debería mostrar lo que significan las anáforas pronominales del tipo de *her* cuando están incrustadas dentro de una subordinada adverbial: estos pronombres refieren «libremente» (pueden referirse al antecedente más próximo o pueden ser anáforas discursivas). La referencia de las anáforas es un problema para estos sistemas combinatorios sin dependencias jerárquicas porque tienden a buscar el antecedente más próximo.

Tiene una capacidad limitada para resolver **ambigüedades**:

> **Yo:** Explícame por qué es léxico-sintácticamente ambigua la oración «María y Luis discutieron sobre el bote».
> **ChatGPT-3.5:** La oración «María y Luis discutieron sobre el bote» es sintácticamente ambigua debido a la falta de claridad en la estructura gramatical, lo que permite diferentes interpretaciones en cuanto a quiénes están involucrados en la discusión y sobre qué se está discutiendo. La ambigüedad sintáctica se debe principalmente a la *ausencia de marcadores* que especifiquen las relaciones gramaticales de manera precisa. Aquí tienes dos formas de analizar la ambigüedad sintáctica de la oración:
> 1. Ambigüedad en la estructura de los sujetos:
> Interpretación 1: «María y Luis» actúan como sujetos compuestos de la oración y discuten juntos sobre el bote.
> Interpretación 2: «María» y «Luis» son sujetos individuales de la oración, lo que significa que cada uno de ellos tuvo una discusión separada sobre el bote.
> 2. Ambigüedad en la relación entre «sobre» y «el bote»:
> Interpretación 1: «Sobre el bote» se interpreta como el tema o motivo de la discusión.
> Interpretación 2: «Sobre el bote» se interpreta como la ubicación física de la discusión, … la discusión ocurrió en algún lugar relacionado con un bote, como un puerto o un muelle.
> La ambigüedad sintáctica en esta oración hace que no se pueda determinar con certeza quiénes están involucrados en la discusión ni cuál es el tema de la discusión sin más contexto.

Como vemos, el chatbot habla tanto de interpretaciones derivadas del conocimiento del mundo como de la ambigüedad que se deriva del significado de las preposiciones. Pero no parece saber muy bien de qué va el asunto, gramaticalmente hablando, cuando afirma que eso se debe a la «ausencia de marcadores que especifiquen las relaciones gramaticales de manera precisa». Da la impresión de estar alucinando, en el sentido en que lo usan los informáticos.

Pregunta sobre gramaticalidad:

> **Yo:** Las dos oraciones siguientes son parecidas. ¿Pueden las dos ser emitidas por un ser humano? Las oraciones son: «La caja de madera está rota» y «La de madera caja está

rota».
ChatGPT-3.5: Ambas oraciones tienen un significado similar, pero la primera oración es la forma más común y natural de expresar la idea. La segunda oración tiene un orden de palabras inusual y puede sonar extraña o confusa para un hablante nativo de español. Sin embargo, ambas oraciones son gramaticalmente correctas y podrían ser emitidas por un ser humano. La primera oración es la forma preferida y más natural de expresar la idea, pero en algunas situaciones específicas (por ejemplo, en un contexto poético o literario) la segunda oración podría ser utilizada con un propósito estilístico.

La segunda oración sería considerada como agramatical por un altísimo porcentaje de hablantes puesto que el español es una lengua de núcleo inicial y el complemento de un nombre siempre va a su derecha (aunque haya excepciones que luego veremos). No obstante, lo que se sigue de todos estos ejemplos es que el chat no se comporta como un lingüista que se referiría a los principios estructurales para la formación de frases nominales sino como un hablante que describe con acierto desigual por qué una oración es más o menos aceptable que otra. El bot está programado para construir oraciones «naturales», «que parezcan tener sentido» aunque no pueda explicar por qué son así.

Sin embargo, sus límites no refieren solo a la gramaticalidad, a la forma de las construcciones. Como indica Haider (2023): «El bot ha sido entrenado para producir textos con significado sobre prácticamente cualquier cosa, sin ninguna comprensión profunda de la materia que se trata. He aquí lo que nos dice sobre su competencia gramatical…». «La configuración interna del box que configura su competencia gramatical es una caja negra».

Yo: How does Chat-GPT decide whether a given sentence is grammatically correct?
ChatGPT-3.5: As an AI language model, ChatGPT does not have a fixed set of rules to determine whether a sentence is grammatically correct or not. Instead, it has been trained on a vast corpus of text data, and it uses statistical patterns and probabilities to make predictions about the grammaticality of a given sentence.

Mujer interactuando con un mapa de ondas sinusoidales en 3D, por ChatGPT-4.

En suma, este chat produce textos (ensayos, conversaciones, situaciones pregunta respuesta) y resúmenes coherentes y fluidos sobre temas diversos, a la vez que generales e imprecisos. Detecta ambigüedades que casi siempre atribuye al contexto, y poco a razones estructurales o léxicas, traduce razonablemente bien y corrige la redacción en otras lenguas, (no siempre) reconoce la agramaticalidad, y la contextualización puede ser irrelevante. A veces no distingue bien los referentes (interpretación de anáforas de discurso, interpretación de deícticos). Estos bots pueden asistir al científico en la escritura de manuscritos o revisiones, aunque la cuestión es cómo controla el asistido los errores que pueden introducir estos modelos.

Los modelos generativos, son buenos encontrando patrones frecuentes y prediciendo qué suele aparecer en contextos similares. De ahí esa sensación que tenemos frente a sus respuestas de que da vueltas sobre esquemas, tiene una cierta verborrea y muchas veces dice cosas que suenan a aproximaciones mal planteadas, incluso erróneas. La propia oficina de ChatGPT-Open AI, hablando de GPT-4, admite que su confiabilidad no puede ser total pese a que puede pasar pruebas donde los humanos tienen ratios muy bajas de éxito (me imagino que se refiere a su capacidad de escribir poesía, a su facilidad para hacer guiones, que pone nerviosos a los guionistas de Hollywood…):

> «Sin embargo, puede superar pruebas en las que los humanos suelen tener bajas tasas de éxito, y tener éxito comparable al del nivel humano en muchos procesos profesionales y académicos. GPT-4 supera a los importantes modelos de lenguaje anteriores, y a la mayoría de los sistemas de gama alta, en una variedad de pruebas tradicionales de procesamiento del lenguaje natural (NLP). A pesar de todas sus capacidades, GPT-4 tiene limitaciones similares a los modelos GPT anteriores. GPT-4 no es completamente fiable, tiene una ventana de contexto limitada y no aprende de la experiencia» (Informe de Open AI 2023).

Hasta aquí una introducción simple a este modelo grande de lenguaje, un recurso tecnológico impresionante, que ayuda a los usuarios y puede llevar a suponer que, de alguna manera, piensa, razona y habla puesto que responde con bastante adecuación. Querría entrar ahora en cuestiones de más enjundia.

4. Qué debería tener y hacer un sistema operativo para que pueda quitarnos la palabra. Limitaciones de los MGL

4. 1. ¿Loros estocásticos o entes competentes lingüísticamente?

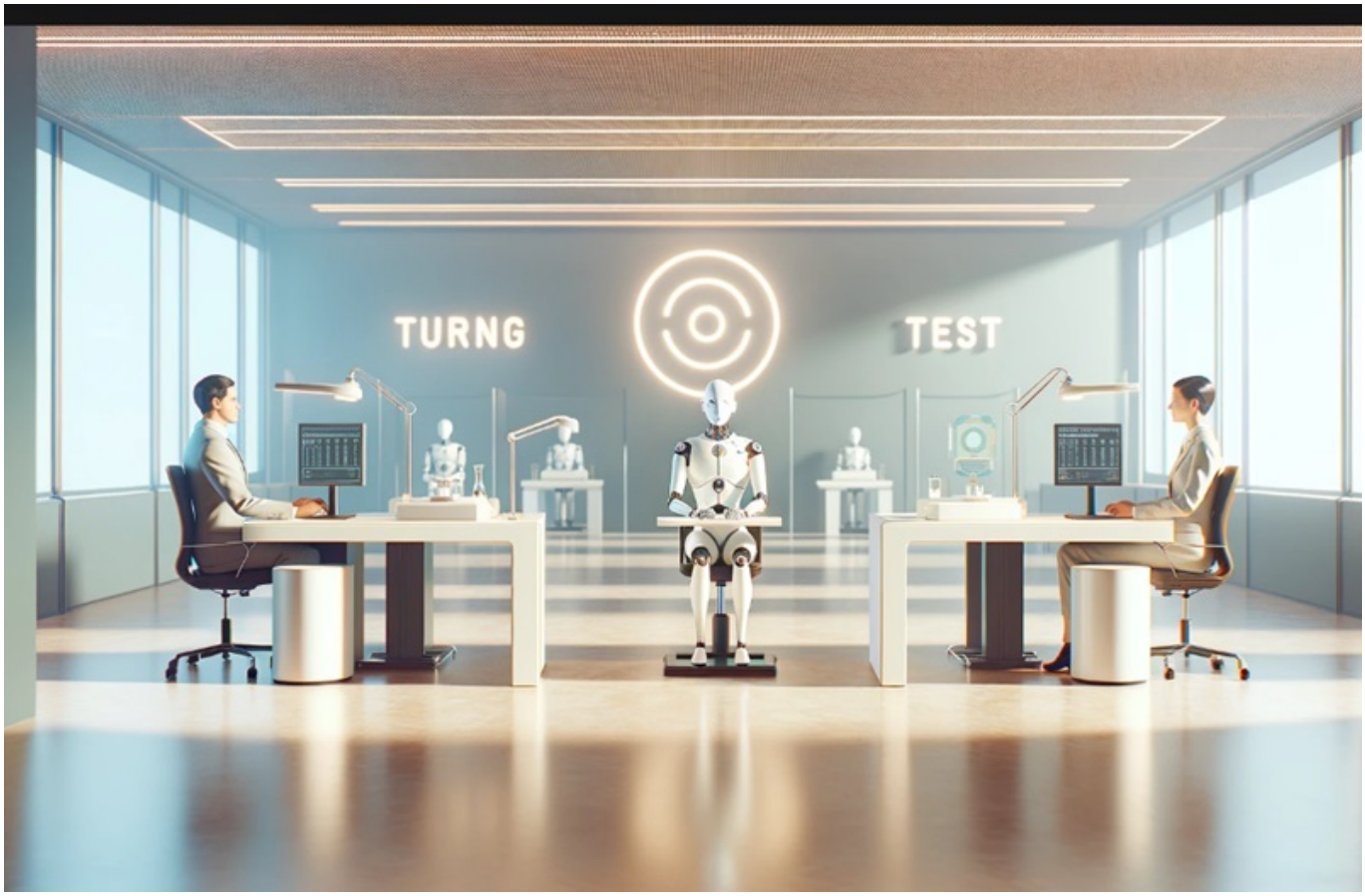
Como estos algoritmos están en continua modificación, ¿llegarán a ser más «humanos» a medida que se modifiquen? ¿Los artefactos habilitadores de tareas lingüísticas de apariencia avanzada no son simplemente «loros estocásticos», objetos fascinantes de ingeniería que, sin embargo, carecen de propiedades esenciales de los seres humanos?

Los seres humanos estamos dotados de una capacidad cognitiva única, una «competencia lingüística» que nos permite no solo construir oraciones bien formadas sintácticamente, y a partir de ahí hacer discursos con una cierta coherencia, sino ser capaces de «comprender / entender», «explicar», «evaluar», «improvisar», «juzgar», entre otras cosas. Mediante el lenguaje expresamos y atribuimos intenciones, formulamos objetivos, suscitamos el reconocimiento de estados de la «mente» de los otros, hacemos inferencias, expresamos relaciones causales, tenemos una percepción de la comprensibilidad (por eso escribimos diez veces un texto hasta que diga lo que queremos decir), deducimos cuál es el *point* de nuestro interlocutor/a, hacemos chistes, podemos ser rápidos e ingeniosos (*witty*), tenemos «retranca», en general nos damos cuenta de qué es relevante, disponemos de formas diversas de manifestar deseos, emociones, dudas, certezas. Y todas estos, llamémoslas así, estados de la mente / cerebro o acciones (cuando externalizamos), los expresamos con inmediatez y con una relativa solvencia, mediante el lenguaje. Todas estas actividades requieren pensamiento, razonamiento, comprensión y conocimiento.

Un filósofo pensará que esto es justamente lo que significa ser inteligente, tener inteligencia general no inteligencia orientada a un objetivo, y tener «consciencia» (sea lo que sea la consciencia, que muy pocos se atreven a definir). Por otro lado, la comprensión, el diálogo, o la utilización tanto creativa como funcional del lenguaje implica poder manejar y resolver vaguedades, dudas, presuposiciones que hacen que la actividad lingüística codifique un nivel alto de sugerencia, incertidumbre, duda, cuestiones todas que plantean un problema serio para cualquier «simulación» de la competencia y la actuación lingüística. ¿Cuáles son los argumentos que asientan la suposición de que la sustituibilidad entre lenguaje humano y modelos grandes de lenguaje no es nada obvia?

4. 2. La prueba de Turing

Turing se preguntaba si pueden pensar las máquinas. Reformulando su pregunta a nuestros efectos, diríamos hoy: ¿pueden las tecnologías habilitadoras de la IA conseguir hacer lo que nosotros como seres lingüísticos hacemos regularmente? La prueba de Turing (1950) examina conversaciones entre una máquina y un ser humano; y un tercero, un evaluador, debe discriminar si uno de los participantes es una máquina, o estamos frente a dos seres humanos. Esta prueba obtiene una respuesta conductista, operacional, a la pregunta qué es pensar y hablar: serían el producto de una «actividad cognitiva ingenieril» y no cerebral, valga la paradoja. En este sentido muchos de los casos que antes presentábamos pasan la prueba de Turing, o por lo menos generan dudas en el observador. Ya hemos visto ejemplos de algunas actividades lingüísticas razonables de estos sistemas. Biever (2023) relata que el Chat 4 obtiene buenos resultados en algunos exámenes de EE. UU., por ejemplo, los que se usan para evaluar capacidades médicas, o las pruebas para habilitarse como abogados. La oficina de San Francisco realizó pruebas que mostraban que el chat da respuestas que lo sitúan en el 10% más alto entre los calificados por esas pruebas. Lo que señalan críticos como Biever es que la «comparativa de evaluación» (*benchmarking*) puede ser floja. El chat está entrenado sobre millones de textos y a veces, según esté formulada la pregunta, contesta bien; pero con un leve cambio en la pregunta ya se equivoca. La misma autora relata otro experimento de comprobación de la posibilidad del chat de superar los requisitos del test de Turing. Se trataba de que un jugador de un juego *online* bastante corriente distinguiera cuándo estaba jugando con un humano y cuándo con GPT-4. Los jugadores identificaron al chat el 60% de las veces. De nuevo Biever advierte que la pregunta que se le formule es crucial: «En muchos casos, el MGL responde soltando palabras que muy probablemente estuvieron asociadas con la pregunta original en los datos de entrenamiento, y no proporcionando en cambio una respuesta buena en un nuevo escenario» (Biever, 2023). En definitiva, la prueba de Turing parece demasiado simple para nuestros efectos pues solo evalúa resultados débiles. La cuestión es si los algoritmos de estos chats pueden reflexionar sobre lo que dicen, si entienden, emiten, tienen lenguaje interior, a saber, ¿son como nosotros?



El test de Turing según ChatGPT-4.

En el momento actual, con los algoritmos que se usan, la simulación de la actividad cerebral sintáctica, semántica, pragmático-discursiva e intencional de los sapiens es más que escasa. Veámoslo.

4. 3. Otros desafíos

En primer lugar, los objetos y eventos lingüísticos generados por estos algoritmos son un conjunto mínimo de los productos y eventos que podemos generar y llevar a cabo los humanos. Crucialmente, sabemos que en estos modelos no hay siquiera un atisbo de consciencia (tener consciencia es comprender, sentir, pensar, tener conocimiento inmediato de uno mismo y, consiguientemente, tener responsabilidades morales, dice Penrose 1989). Pero aún si dejamos de lado el difícil concepto de consciencia, hay otras zonas grises que conviene mirar.

Competencia lingüística

Así, se basan en una concepción quizá poco creíble de lo que es la capacidad lingüística humana pues no formalizan la «competencia lingüística». No parecen ser modelos de lenguaje humano según lo ven algunos expertos (Marcus, 2023)Piantadosi 2023, en un artículo tan interesante como en alguna medida incompleto y centrado hacia sus objetivos, sostiene que estos modelos de lenguaje refutan la visión chomskiana de la capacidad lingüística..La tecnología de los MGL se sustenta sobre algoritmos de aprendizaje profundo (redes neuronales profundas o «transformadores»), que se entrenan larga y profundamente y hacen algunas cosas en las que parecen ser iguales y en alguna ocasión mejores que las de los humanos. Pero sus limitaciones son importantes. Si trasladamos a los sapiens las capacidades de estos algoritmos, la idea importada de ellos sería que los humanos desarrollamos las lenguas a partir del entrenamiento sobre miles de millones de «tokens» (palabras o partes de palabras) y que somos una máquina estadística que genera (predice) patrones de texto coherentes. Y no está claro que así sea.

Comprensión, predictibilidad y cómo funciona el ChatGPT

Otra cuestión que hace poco concebible que estos algoritmos basados en redes neuronales sean un modelo adecuado del lenguaje humano se refiere a la semántica —a la comprensión de lo que se genera—, y a la pragmática, en conjunto a la capacidad para modalizar lo que se genera (aserción, duda, perplejidad, composición de la forma con el significado). Oliver (2018) señala como dos de los problemas fundacionales de la IA pendientes de resolver: «la necesidad de incorporar semántica y razonamiento en los sistemas de procesamiento del lenguaje natural» y «la necesidad de desarrollar representaciones de la incertidumbre que sean tratables computacionalmente». Pero crucialmente, el lenguaje humano es también una arquitectura. Las lenguas naturales tienen una armazón, la sintaxis, un léxico y mecanismos específicos, generalmente en forma de reglas, para generar e interpretar y usar el lenguaje. Y, lo más importante, tienen interfases con los otros subsistemas del sistema cognitivo: pensamiento, intenciones, reconocimiento de la mente del otro, entre varias otras cuestiones que luego veremos. Las lenguas generadas por los chats generativos tienen otra arquitectura y otra génesis.

Los chats generativos, como ya he sugerido, usan análisis estadísticos y algoritmos de aprendizaje para determinar la viabilidad de los textos basándose en sus datos de entrenamiento y en los patrones aprendidos de los datos. Se entrenan sobre tokens: «caracteres compilados —vectorizados— en una unidad de procesamiento con significado» (Aydin *et al*. 2023), a la manera de la semántica distribucional. Los vectores tienen palabras y contextos inmediatos. Analizan pedazos de prosa en paralelo decidiendo qué elementos merecen «atención», gracias a los «transformadores»Los modelos transformadores (un tipo de red neuronal) fueron creados por Google, pero OpenAI, que es la que los

usa, es una empresa independiente; ahora mismo financiada también por Microsoft.. Durante el entrenamiento los modelos se corrigen y van ajustándose. Con los transformadores, más procedimientos estadísticos, generan lenguaje seminatural usando mecanismos de autoatención; aprenden del contexto de diálogo para producir respuestas más exactas. En el proceso de generación de respuestas combinan lo aprendido durante el entrenamiento con los textos actuales, y evalúan posibles respuestas; en este proceso. En suma, estos modelos fusionan sintaxis y semántica, recogen información en vectores, y definen así la similitud entre palabras, construcciones, patrones sintácticos, encuentran «tópicos». Usando un símil culinario, primero está la preparación: coger todos los textos que se encuentran en internet, entrenarlos largamente con diversos mecanismos computacionales, revisar con humanos los resultados de los entrenamientos y hacer pruebas. Entonces se empiezan a hacer los pasteles, cada uno independiente de los otros, probando y comparando sabores y formas cada vez, digamos. Los MGL buscan regularidades estadísticas entre sus millones de datos, lo que les permite hacer una predicción bastante correcta de las palabras o construcciones que se suceden entre sí. Estadística, redes neuronales, aprendizaje computacional, identificación de patrones significativos salidos del entrenamiento, asociaciones entre palabras son los mecanismos simultáneos, los artilugios de la cocina que generarán las respuestas a las instrucciones que les demos.

Los humanos tenemos esa capacidad combinatoria y asociativa y también hacemos predicciones, por ejemplo cuando formamos términos del español a partir de otras lenguas: *mapear, hackear*, o cuando el niño dice *cabí* o forma términos derivados*: construir → construcción*. O sabemos que *engordar* está en una ontología que tiene términos como *comida*, *sedentarismo* o *metabolismo*. La capacidad estadística, probabilística, asociativa interviene en los procesos lingüísticos, pero no parece ser el fundamento de nuestra competencia lingüística.

Estado inicial

Las hipótesis acerca del desarrollo del lenguaje en los humanos (desde el innatismo extremo al constructivismo, o los modelos de procesamiento de la información) suelen coincidir en que hay un «estado cognitivo inicial» o la predisposición a él (sea una capacidad independiente o en interacción con otras) a partir de la cual y en un período de tiempo relativamente breve (4 o 5 años) se configura lo que denominamos el «conocimiento del lenguaje». Chomsky, considera que ese diseño o estado inicial es el legado biológico que hace posible la adquisición del lenguaje y lo caracteriza formalmente como un sistema de reglas. Ese conocimiento podría coadyuvar principios estadísticos, pero lo cierto es que la adquisición del lenguaje se basa en una exposición a pocos datos. En este proceso solo habilitamos lenguas posibles. Los algoritmos de los MGL, a diferencia del cerebro / mente generan tanto lenguas posibles como imposibles: si el conjunto de datos con los que los alimentamos está lleno de disparates gramaticales o los datos se basan en una gramática inexistente se generarían textos imposibles.

Incrementalidad, inmersión, pobreza del estímulo

En el mismo sentido, los organismos biológicos (el cerebro) aprenden rápidamente y aprenden de manera incremental**,** es decir van añadiendo conocimiento sin tener que volver a aprender desde cero y aprenden desde la inmersión. En cambio, los sistemas de generación de lenguaje necesitan millones de ejemplos para poder organizar un resumen, dar una respuesta, resolver un problema menor. Como dice Oliver (2018): «… las redes neuronales profundas necesitan millones de observaciones para poder aprender un patrón, mientras que el cerebro necesita pocos ejemplos».

Un ser humano no necesita rodearse de millones de tokens y descubrir patrones para tener lenguaje. Si la base del conocimiento del lenguaje fuera eso, el aprendizaje de una lengua requeriría centenares de años. El argumento de la «pobreza del estímulo» (se aprende una lengua rápido, con una exposición somera a ella y con datos que pueden ser defectuosos: oraciones incompletas, secuencias ininteligibles y similares) se desplomaría, y reduciríamos el desarrollo del lenguaje a un proceso puramente empírico, de imitación y ensayo y error, del tipo de lo que suponían los psicólogos conductistas. Si aceptamos la distinción aristotélica entre posesión y uso del lenguaje, entre «competencia y actuación», afirmaremos que estos programas / máquinas tienen un defecto de base: funcionan con principios derivados de lo que se ve, y no buscan explicar el lenguaje interiorizado, los principios que subyacen a la actuación.

En todo caso, este es un tema abierto, una hipótesis (u observación) que ha sido muy discutida y que los constructores de los MGL ponen en cuestión (Piantadosi 2023), pues para ellos la mera existencia de estos MGL que aprenden sería una prueba de que los sujetos no necesitan estructuras previas, de diseño, para aprender. Simplificando mucho, sostienen también que sus modelos probabilísticos, graduales, si bien actúan sobre estructuras lineales en espacios no restringidos, pueden inferir jerarquías. Por ello, es que los sujetos no necesitarían una restricción sobre las jerarquías implantada en el cerebro para poder adquirir su lenguaLa visión de Piantadosi y similares no parece tener en cuenta la estructura real de la noción de pobreza del estímulo que implica, como todo proceso biológico tres elementos imprescindibles: (i) dotación innata, (ii) exposición a datos lingüísticos, (iii) generalización y parametrización.

¿Pensamiento verbalizado? Mente. La habitación china

Una característica central de las lenguas humanas es que sus expresiones son relaciones sonido-sentido. La «comprensión» de las oraciones y construcciones, un acto cognitivo y lingüístico, es el resultado de asignar significado a estructuras formales —teniendo en cuenta dependencias,

jerarquías, tipos de estructuras—, y esto sería la base de lo que llamamos «pensamiento verbalizado». La semántica formal explica este proceso, y este mecanismo sería en parte el fundamento del pensamiento; el lenguaje viene a ser así una ventana al pensamiento. Nadie ha afirmado, que yo sepa, que haya una relación uno-a-uno entre lenguaje y pensamiento (volveré sobre esta cuestión)Aunque sea un inciso, quiero recordar que hay otras formas de pensamiento que no son verbales, Einstein decía que pensaba con imágenes: «Las palabras, el lenguaje, no parecen jugar ningún papel en mi mecanismo de pensamiento. Las entidades físicas que parecen servir como elementos del pensamiento son ciertos signos e imágenes» (tomado de Penrose 1989).. Si bien como hablantes nos parece, a veces, que cuando decimos pensamos, y viceversa. ¿Puede esta interacción obtenerse mediante los algoritmos de la IA? Por el momento no, por buenos que sean los traductores, correctores, asistentes de voz, sistemas de recomendaciones que incorporan el contexto, etc.

Roger Penrose es también muy escéptico en cuanto a que los algoritmos de la inteligencia artificial puedan ser un modelo de la actividad cerebral. En su libro de 1989, *La nueva mente del emperador*, muestra su interés por la prueba de la habitación china relativa a las posibilidades de la IA fuerte, planteada por el filósofo John Searle en 1980.

Así imagina ChatGPT 4 la habitación china de Searle.

En esta prueba, una persona que no habla chino es (supuestamente) encerrada en una habitación. Esa persona recibe a través de una ranura tarjetas escritas en chino y un manual de instrucciones en su propio idioma (el español, por ejemplo) que le indica cómo manipular las tarjetas para proporcionar respuesta a las sucesivas tarjetas en chino que le llegan. A pesar de no entender chino, la persona encerrada consigue dar respuestas adecuadas porque tiene instrucciones muy detalladas para responder a las preguntas, y contesta bien. Desde fuera, las personas que envían las tarjetas podrían pensar que el individuo dentro de la habitación habla y «entiende» chino. Pero ¿hay comprensión en este caso? Searle dice que no. Con la prueba de la habitación china, Searle pretende demostrar que «aunque la prueba de Turing sea una buena definición operacional de la inteligencia, no indica si la máquina tiene una mente, una consciencia o intencionalidad».

Por aquí conviene recordar entonces la cuestión de cómo la «atribución de intenciones», «pensamientos», «estados mentales» al interlocutor —lo que se suele llamar «teoría de la mente»— es esencial para generar secuencias lingüísticas, interactuar, abrir el pensamiento. Con recurso a patrones estadísticos probables no se genera intencionalidad, atribución de creencias y similares. Sin embargo, uno de los máximos valedores de los MGL, el neurocientífico de Berkeley Piantadosi (2023), señala que la crítica a la escasa capacidad semántica de estos modelos no es correcta:

«Existe una interfaz bastante uniforme para (establecer) cómo el contexto y el significado de las palabras predicen el material próximo; la sintaxis y la semántica, en este modelo, no se separan en componentes distintos, ni en mecanismos predictivos separados. Debido a esto, los parámetros de red que estos modelos encuentran combinan propiedades sintácticas y semánticas, y ambas interactúan entre sí y con el mecanismo de atención de maneras no triviales».

La idea en realidad no es nueva en la lingüística, muchos semantistas han argumentado que el sentido de una palabra puede estar influido por su inmediato contexto morfosintáctico, así como por el más extenso contexto de uso. La cuestión es su capacidad de «predecir». Para discutir esta idea, plausible en muchos casos, de que en los MGL el contexto y el significado convencionalizado de las palabras derivan correctamente la relación entre forma y sentido y «predicen lo que va a venir», quiero referirme al papel de la predictibilidad / probabilidad en la formación de estructuras y de sentidos.

5. La probabilidad en el lenguaje. La diferencia entre contexto local y contexto global

Piantadosi, en la cita que antes introduje, afirmaba que la interacción entre palabra y contexto «predice» el material que va a continuación cuando se están generando textos. En el chat GPT-2 la arquitectura se basaba en la adición de palabras una tras otra teniendo en cuenta las probabilidades más altas o bajas de que esas palabras estuvieran juntas (Wolfran 2023). Un Sistema como ChatGPT puede predecir cuál es la continuación de una serie de palabras. Sin embargo, parecería no estar preparado para trabajar con estructuras, órdenes y relaciones, que es lo que caracteriza al lenguaje y al pensamiento abstracto. Por otro lado, cuando yo reflexiono un momento antes de terminar una secuencia como: *Este artículo es acertado / erróneo / toda una novedad* puedo naturalmente explicar mi decisión y jamás diría que escojo uno u otro por razones de frecuencia en una base de datos. Lo que buscamos son palabras que sean «relevantes y verdaderas» (o falsas si se es un mentiroso) dentro de esa proposición, es decir que añadan información a nuestro universo del discurso y sean adecuadas —o tengan valor de verdad— con respecto a la realidad que estamos caracterizando. No buscamos que sean más o menos frecuentes en relación con algún tópico del texto, de ahí el peligro que entraña generar texto si los datos de entrenamiento son sesgados o provienen de redes sociales tendenciosas y falsas. Y de ahí las cautelas que muchos aconsejan.

5.1. Probabilidad y gramaticalidad

Para Katzir (2023) estos modelos prefieren la probabilidad a la gramaticalidad. Estas dos situaciones a veces se solapan pero son conceptualmente distintas, algo puede ser improbable pero correcto («Era del año la estación florida / en que el mentido robador de Europa / —media luna las armas de su frente, y el Sol todos los rayos de su pelo—», comienzo de *Las Soledades* de Góngora) y algo puede ser probable pero incorrecto («Si me queréis, hablar»). Los modelos GPT generalmente preferirán una palabra que sea una mejor continuación probable, desdeñando la estructura, como vemos en el siguiente ejemplo (buscado por mí siguiendo una idea de Katzir 2023):

> **Yo:** Tenemos la siguiente oración: «El perro de mi hermana que ayer vio los caballos con manchas doradas…». ¿Cuál es la palabra que va mejor a continuación de ese texto: (a) *son,* (b) *rompió*, y por qué?
> **ChatGPT 3.5:** La palabra que va mejor a continuación del texto proporcionado sería «(a) son».
> La oración original es un fragmento incompleto y carece de un verbo que le dé sentido completo. Agregando «(a) son» después de la oración, obtendríamos:
> «El perro de mi hermana que ayer vio los caballos con manchas doradas son…»
> Esta versión tiene más sentido gramaticalmente, ya que «son» actúa como el verbo que completa la oración. No podemos usar «(b) rompió» porque no tiene relación con el contexto de la oración original y no encaja en el significado de la misma.

Lo que el chat no considera es que en la construcción de las oraciones, sintácticamente hablando, opera un principio de «dependencia de la estructura»y no de «probabilidad de coaparición». Los artífices de este chat argumentan que, si bien este sistema ChatGPT no tiene un conocimiento implícito de reglas gramaticales, las redes neuronales podrían descubrirlas a través del entrenamiento pues acierta bastante con ellas:

> «La simple tarea de predicción de palabras es una señal muy efectiva de autosupervisión: las redes neuronales pueden mejorar y mejoran en esta tarea al inducir sus propias representaciones de la estructura de las oraciones, las cuales aprehenden muchas de las nociones de la lingüística, incluyendo clases de palabras (partes del discurso), estructura sintáctica (relaciones gramaticales o dependencias) y correferencia (…), de hecho, la codificación aprendida de una oración incluye en gran medida la información encontrada en las estructuras (jerárquicas) arbóreas propuestas por lingüistas». (Manning et al. 2020)

El hecho de que un modelo que solo se alimenta con palabras acierte en general en casos de concordancia con un posible atractor intermedio podría indicar que de alguna manera construye estructuras. Por ejemplo, si en la frase *El hombre [que trajo las pizzas] es / son* el bot acierta con la concordancia en singular, indicaría que tiene una representación con jerarquía y no puramente lineal. Esto no es lo que parece suceder (como indicaba el ejemplo de más arriba).

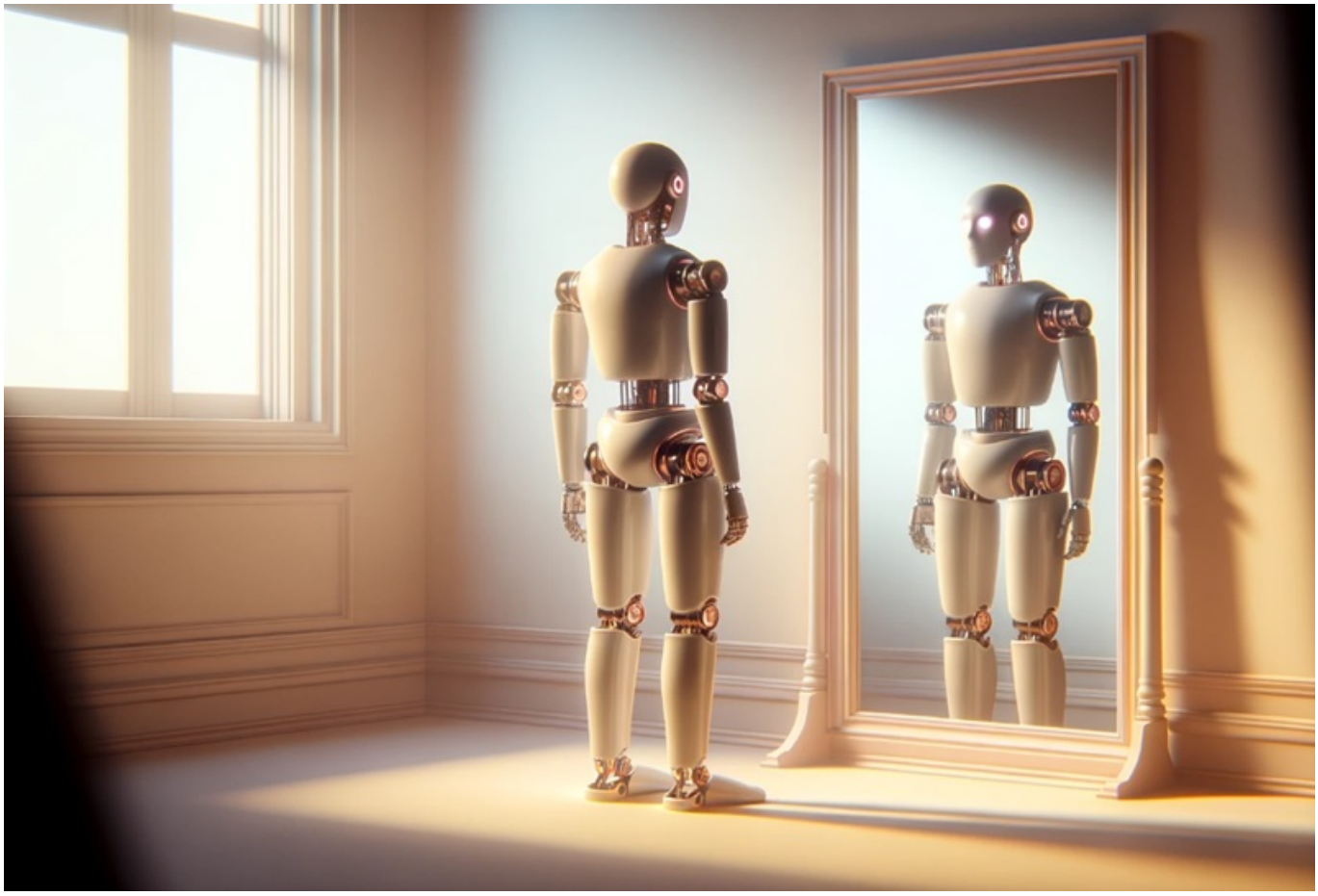
5.2. Probabilidad y realidad

La explicación que el chat proporciona sobre la siguiente búsqueda de «la palabra que sigue» indica lo que decía: escoge por razones de frecuencia y mayor asociación entre patrones, no por factores racionales como la adecuación a la realidad que está describiendo:

> **Yo:** Escribe la palabra que falta en las dos siguientes oraciones: «Es un día bonito, el cielo está ____»; «Mi hermana comió demasiado y ahora está ___».
> **ChatGPT-3.5:** Es un día bonito, el cielo está despejado; Mi hermana comió demasiado y ahora está indispuesta.
> **Yo:** Explícame por qué has escogido las palabras de las oraciones anteriores.

ChatGPT-3.5: Por supuesto, con gusto te explico las palabras que he escogido para completar las oraciones:
1. Es un día bonito, el cielo está despejado: La elección de «despejado» encaja bien en el contexto de un día bonito. Cuando decimos que el cielo está despejado, significa que no hay nubes y el sol brilla claramente. Esta palabra complementa la idea de un día agradable y soleado.
2. Mi hermana comió demasiado y ahora está indispuesta: Opté por «indispuesta» porque en este contexto se refiere a que tu hermana no se siente bien debido a haber comido en exceso. «Indispuesta» es una manera educada de decir que alguien está experimentando malestar estomacal o algún tipo de incomodidad debido a la comida.

En resumen, las palabras *despejado* e *indispuesta* se eligen en función del contexto y de las connotaciones que transmiten en las oraciones dadas. Evidentemente, otras palabras tendrían las mismas papeletas para ser elegidas. Si los madrileños admiran su cielo, tal vez preferirían decir: *Es un día bonito, el cielo está azul*. Igual que podríamos decir *furiosa, disgustada* (porque hoy le toca ir a la ópera) en vez de *indispuesta*. La fragilidad semántica y pragmática del sistema probabilístico es evidente.



Robot mirándose en el espejo, especificando que el reflejo le devuelva la mirada. ChatGPT-4.

5. 3. Predictibilidad y contextos local y global

Quiero ver ahora otras búsquedas que continúan el diálogo con el bot después de la pregunta sobre el día bonito. Estos casos nos permiten quizá entender mejor el papel de las «colocaciones» o afinidades léxicas (que ya veíamos en el ejemplo sobre Cortázar) y del contexto local frente al contexto global en las opciones para la formación de respuestas o tareas de completamiento:

> **Yo:** ¿Y qué pasaría si en la primera oración pones *azul*?
> **ChatGPT-3.5:** Si en la primera oración pusiera «azul» en lugar de «despejado», la oración quedaría así: «Es un día bonito, el cielo está azul».
> Esta opción también sería válida, ya que «azul» se refiere al color del cielo en un día despejado y soleado. Sin embargo, la palabra «despejado» añade más información al

contexto al indicar que no hay nubes en el cielo, lo cual es un detalle importante para describir un día bonito y soleado. En este caso, ambas opciones funcionarían, pero «despejado» brinda un matiz más completo sobre las condiciones del cielo en ese día.
Yo: ¿Y qué pasaría si en la primera oración pones *romántico*?
ChatGPT-3.5: Si en la primera oración pusiera «romántico» en lugar de «despejado», la oración quedaría así: «Es un día bonito, el cielo está romántico».
En este caso, la palabra «romántico» no encajaría de manera coherente en el contexto de la oración. Normalmente, cuando hablamos del cielo en relación a un día bonito, solemos usar adjetivos que describen sus condiciones meteorológicas, como «despejado», «soleado» o «claro». La palabra «romántico» no se relaciona directamente con las condiciones del clima…

Como se ve el algoritmo es tozudo y cree mucho en su sistema combinatorio basado fundamentalmente en los vectores previamente seleccionados y su frecuencia.

La oración *El astrónomo se casó con la estrella* ha sido suscitada por algunos expertos para mostrar la diferencia entre «contexto local» (el que tiene que ver con las restricciones de selección de una palabra, las preferencias léxico-sintácticas por así decir) y el «contexto global»: «la capacidad de un hablante nativo para visualizar el mundo de un texto usando "marcos" evocados por el texto en la interpretación de sus elementos… y construir así una interpretación» (Erk y Herbelot 2022). En el contexto local, *casarse* se asocia con un ser humano activo, pero también puede evocar un objeto celestial. Un entramado de «situaciones», «restricciones» y «conceptos evocados», modelizados en el algoritmo, permitiría la correcta comprensión y desambiguación. Esta programación lleva al bot a contestarme lo que sigue, que me parece muy poco convincente:

Yo: ¿Cómo interpretas la oración «El astrónomo se casó con la estrella»?
ChatGPT-3.5: La oración «El astrónomo se casó con la estrella» es ambigua y puede interpretarse de diferentes maneras según el contexto. Aquí hay dos posibles interpretaciones:
1. Interpretación literal: En esta interpretación, se entiende que un astrónomo (una persona que estudia el espacio y los astros) se casó con una estrella, es decir, con un objeto celestial compuesto principalmente de gas que emite luz y calor debido a la fusión nuclear en su núcleo. Esta interpretación es claramente absurda desde una perspectiva científica, ya que las estrellas son objetos astronómicos y no pueden ser seres humanos ni contraer matrimonio.
2. Interpretación figurativa o poética: En esta interpretación, la frase se utiliza de manera

figurativa para transmitir una idea o sentimiento más abstracto. En este contexto, «el astrónomo» podría referirse a alguien que está profundamente enamorado o fascinado por una persona (llamada «la estrella» en sentido figurado) que destaca por su belleza, personalidad, talento, fama, o algún otro atributo especial. En este caso, «casarse con la estrella» significa que el astrónomo está completamente dedicado o entregado a esa persona, pero no implica un matrimonio literal.
La interpretación de esta oración dependerá del contexto en el que se utilice y de la intención del hablante. En un contexto científico o literal, la primera interpretación no tiene sentido, mientras que en un contexto literario o poético, la segunda interpretación podría ser más apropiada.

Veámoslo. Hay dos elementos activos en la ambigüedad de esta frase, que el bot asume bajo los rótulos de «interpretación literal» e «interpretación figurativa» más abstracta (lo cual no está mal), son: a) los dos sentidos de *estrella*: objeto celestial y persona famosa (una metáfora del sentido literal), y b) la semántica léxica de *casarse*, que significa unirse, estar estrechamente ligado a algo o a alguien. No se sabe por qué el chat invierte las interpretaciones que vendrían a ser: con la *estrella celestial* en realidad no puede casarse pero se entrega a ella; con la persona famosa sí puede casarse y ese sería el significado más próximo al contexto local; pero los dos significados son posibles si se tiene en cuenta el significado de *casarse* y no solo de *estrella*. El chat lo interpreta al revés. La sensación que deja la respuesta es que el algoritmo anda un poco despistado. Evidentemente, un ser humano no necesita una sofisticada semántica con nudos conceptuales y situaciones evocadas para entender adecuadamente la expresión bien como un juego de palabras que extiende el significado de *casarse*, bien como una boda con alguien famoso. Un ser humano la interpretará adecuadamente en un segundo. Esto es un ejemplo más de que el bot no sabe bien lo que dice. Eso no quiere decir que la probabilidad no desarrolle ningún papel en la adquisición y en el uso del lenguaje, aunque no es este el lugar de entrar en ese asunto. Pero demos un paso más en el análisis de las limitaciones de ChatGPT.

6. ¿Por qué el simple desarrollo de instrumentos computacionales más avanzados no garantiza que la IA pueda quitarnos la palabra?

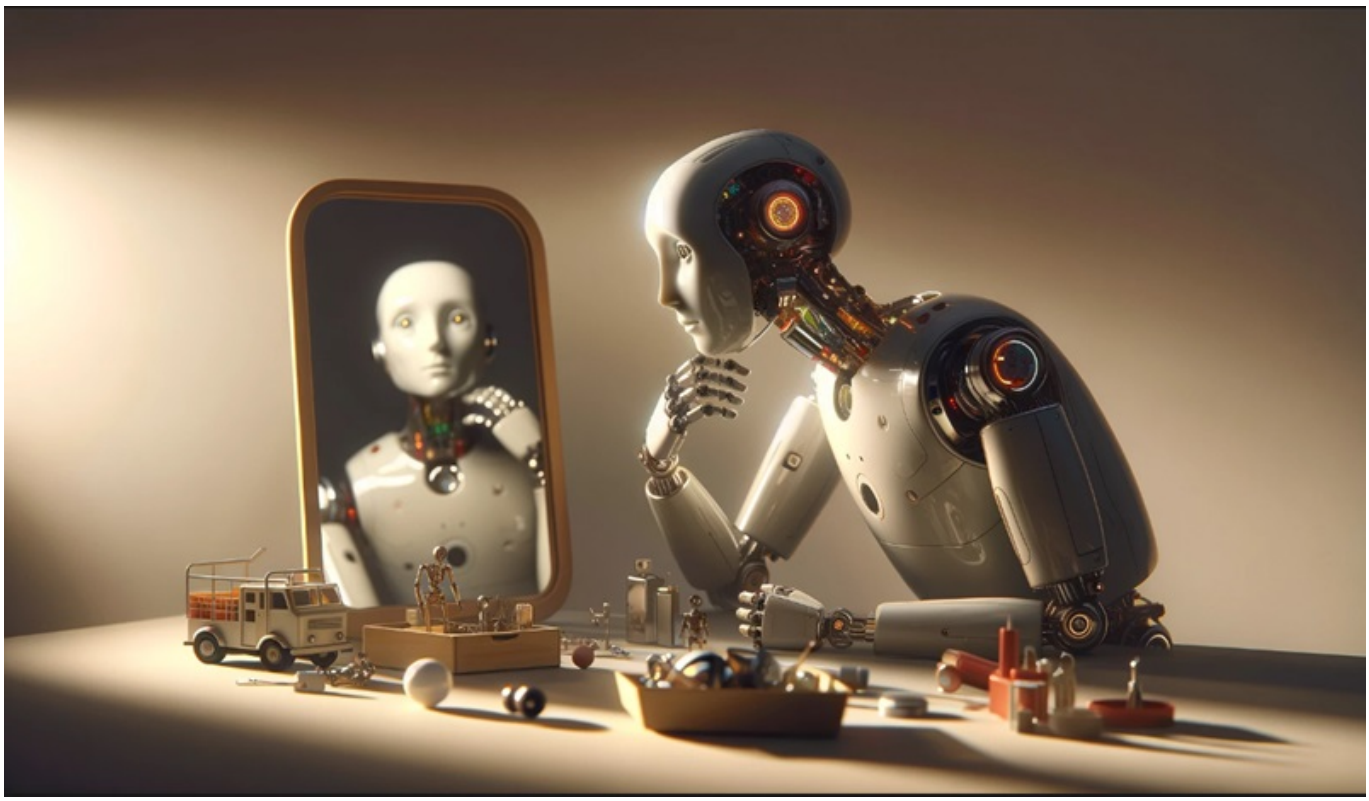
Yendo más al fondo de la cuestión, en realidad las más importantes limitaciones de este chat en cuanto hablante posible conciernen a asuntos que no se pueden contrastar mediante preguntas al artefacto porque van más allá de ser problemas sintácticos, semánticos, pragmáticos, contextuales o de adecuación a la realidad, son cuestiones de la cognición de la que el lenguaje es un sistema más. Tienen que ver con lo que hace humanos a los humanos y con la forma en que el lenguaje interactúa con los otros módulos del sistema cognitivo general.

Empezando por una síntesis, las razones de fondo para sostener que los MGL no son modelos del lenguaje sino instrumentos ingenieriles que «actúan» como si supieran usar lenguaje humano son varias. La primera es que no entienden sino que «procesan» algorítmica y estadísticamente lo que se les dice. «La última frontera de una inteligencia artificial sofisticada deberá tener un cuerpo y sensores: ojos, tacto y orejas. Un futuro robot humanoide que aprenderá del mundo interactuando con él», decía con razón López de Mántaras en una entrevista en *El País* el 6 de septiembre de 2023. Con un modelo como el del ChatGPT esto es imposible. Incluso si adoptamos su propio camino, los algoritmos deberían ser entrenados con una base de datos, por ejemplo, de vídeos que «muestren» cómo se razona, si ello es posible; qué son relaciones causa-efecto; tener billones de muestras de situaciones de interacción lingüística, profesional, de causalidad, de sentido común. Y tener un programa que genere lenguaje que entienda y que interactúe con el mundo. ¿Por qué?

6.1. Lenguaje y comunicación

Las ejecuciones lingüísticas de los humanos tienen formas diversas: pueden ser espontáneas y semiautomáticas, cuando estamos en situación de interacción discursiva, o pueden estar planeadas previamente, como lo que estoy haciendo yo ahora (hacemos planes sobre lo que vamos a decir), en situaciones más regladas o previsibles. Pero en todas estas situaciones encontramos rasgos comunes: nos damos cuenta, percibimos o entendemos lo que decimos, y tenemos siempre capacidad de autorreflexión o autodeliberación, por eso podemos —en el mejor de los casos y si no estamos contaminados por sesgos cognitivos o de interés— corregirnos, revisar, rectificar, cambiar de ideas. Como percibimos lo que hemos dicho, podemos dar razón tanto del significado literal de nuestras expresiones como del contenido implícito, tanto del contenido recto como del metafórico, de los juegos de palabras, bromas, metáforas, etc. Y se nos puede hacer ver que hemos dicho esto o lo otro.

Empezando por la relación lenguaje / comunicación, la suposición de que el «código» (el lenguaje) es materia importante de la comunicación, pero no el único factor, la glosa muy bien John Searle (1969) cuando dice que uno puede pedirle a alguien que se vaya de una habitación sin decir nada, con un gesto, pero uno no puede explicar que es el «ser» sin usar un código. A saber, la comunicación puede ser independiente del código, pero debe señalarse que, si bien lenguaje y comunicación parecen ser dos constructos cognitivos diferenciados, los actos lingüísticos de comunicación implican una intención por parte del hablante, que se puede expresar en el texto entre otras muchas maneras por medio de verbos intencionales (*creer, suponer, pensar*), así como una capacidad por parte del oyente para inferir las intenciones del hablante. Sin intencionalidad y capacidad inferencial lo que habría serían monólogos automáticos preprogramados.



Robot mirándose al espejo. No se le indicó a la IA que el reflejo tuviera que devolver la mirada.

En la generación de la «comunicación intencional» desempeña un papel central el conocimiento del mundo y el análisis que hayamos hecho de las situaciones preparatorias, de la capacidad que tengamos para actuar adecuadamente en situaciones ambiguas, como son la mayoría de las situaciones lingüísticas; y del manejo apropiado de los elementos del contexto general y particular. Las capacidades típicas de los modelos de PLN tienen que ver con opciones establecidas por los mecanismos de entrenamiento, mientras que los humanos no nos entrenamos para responder a un agente que nos va a poner una multa, o para intentar impedir que alguien cruce un paso de peatones cuando viene un coche, o para evitar un conflicto soslayando algún elemento de la acción comunicativa que pudiera introducirlo.

Los algoritmos que pudiesen hablar con nosotros, «comunicarse», ser discursivos y capaces de actos de habla, tendrían que ser objetos intencionales que capten las intenciones de los otros y actúen con una gran labilidad a estos respectos, como nos sucede a los humanos. Los seres intencionales advierten, o deberían advertir —para qué engañarnos y olvidarnos de los sesgos cognitivos—, qué es lo relevante en cada momento. No se me ocurre que la IA lingüística / comunicativa pueda ser capaz de definir bien qué es relevante y qué irrelevante en una situación comunicativa cuando sus actuaciones están regidas por un algoritmo que les hace buscar en un universo de datos enormes los tokens más cercanos e irlos engordando progresivamente.

6. 2. Lenguaje y pensamiento. La creatividad, el chiste, la ironía

Creatividad es una palabra muy utilizada y cuya definición no es fácil. La creatividad se tiene, se mejora y se potencia, es un don y una práctica. Como bien indica Aladro (2023), la creatividad implica un uso disruptor de los códigos y lenguajes habituales. La saco a colación no porque esté directamente conectada con el lenguaje (aunque algo está, como diré) sino para señalar que los modelos grandes de lenguaje proporcionan una visión deforme de la creatividad. Hacen creer que crean porque generan ensayos o textos. Pero si suponemos que la creatividad es hacer lo que no se ha hecho antes: transportar visiones de unos campos a otros; ampliar los campos de saberes; abstraer y generalizar debidamente; entonces los sistemas de la IA son casi la antítesis de la creatividad. En palabras de Aladro (2023) «La inteligencia artificial no puede generar información nueva ni usar el lenguaje para generar nuevos pensamientos jamás concebidos por los seres humanos, porque su base de trabajo es lo "ya sabido"… la información universalmente compartida en la red». Cuando los algoritmos hacen de escritores tienen el máximo de creatividad que les permite la imitación sin pensamiento. Si la literatura no depende de la gramática (Borges) menos aún depende de la combinatoria probabilística. Sin embargo, muchos llaman «crear» a lo que hacen estos algoritmos. Cito a una autoridad ministerial:

> «"A muchos nos pilló por sorpresa", ha reconocido Bolón. Porque existen herramientas de IA desde hace décadas, lo que han demostrado ChatGPT y el resto de (las) inteligencias artificiales denominadas "generativas" es que "es buena en algo en lo que pensábamos que solo éramos buenos nosotros, que es en crear", ha recordado Artigas» (Coloquio de *El Diario*, 9 de septiembre de 2023).

Uno no deja de sorprenderse cuando oye esto. La creatividad lingüística en los chatbots y similares se obtiene aleatoriamente, si se obtiene. Es cierto que este chat puede ayudar a los guionistas, entre otras cosas, y de ahí los problemas en Hollywood, porque las series son rutinarias y están hechas todas con los mismos patrones: un crimen, un misterio, unos sospechosos inesperados, vueltas sobre otros personajes, siempre con gran lentitud y mucho morbo y poca sustancia en los diálogos, salvo por la recurrencia a las palabras gruesas. Pero si queremos que haga un guion para películas de gran imaginación y sensibilidad como *Rocco y sus hermanos, Blade Runner, o Drive my car,* creo que le plantearemos una tarea imposible aun cuando le diésemos muchos mimbres previos. Es un instrumento ideal para «ayudar a pensar» (eso dicen algunos) en una época precipitada, repetitiva, superproductiva de banalidades cortadas con un mismo patrón. Ahora bien, si incluimos en la creatividad el poder hacer una operación muy extensa, inalcanzable para un humano, como planificar todos los movimientos de los trenes en Europa en un solo día, entonces podemos concederle a la IA

una (pseudo)creatividad importante, pero de segundo nivel.

La creatividad se relaciona con el pensamiento, pero va mucho más allá. Pensamiento y lenguaje son independientes, ya lo anticipé: hay afásicos totales que emiten señales claras de que tienen pensamiento y seres con serios déficits de cognición pueden tener un sistema lingüístico aceptable; no obstante, puede hablarse de alguna relación entre los dos sistemas. Los sistemas lingüísticos biológico-culturales en realidad presentan de manera bastante indeterminada los contenidos conceptuales que transmiten y los van «construyendo» sobre la marcha. Señalan Gleitman y Papafragou (2012):

> «Si uno intentara decir todo y exactamente lo que uno querría significar, la conversación no podría tener lugar. Los hablantes se perderían en sus pensamientos. En cambio, la conversación implica una negociación constante en la que los participantes estiman y actualizan el conocimiento previo del otro como base para lo que es necesario decir frente a lo que es mutuamente conocido e inferible…».

En este sentido el lenguaje ayuda a generar ideas. Cuando entendemos y explicamos metáforas o metonimias (*la mesa 10 espera la cuenta* vs. *el sándwich de queso espera la cuenta*), chistes, construcciones irónicas, o cuando discutimos sobre asuntos teóricos, prácticos o personales, cuando escribimos y vemos que una frase lleva a otra podríamos decir que el lenguaje ayuda a generar ideas; y esto vendría a ser una forma de la creatividad.

Estos chats, sin embargo, si bien pueden describir una ironía, una metáfora, una antítesis, no pueden generarla. Por ejemplo, si a ChatGPT le haces la siguiente pregunta, contesta adecuada aunque verborreicamente:

> **Yo:** Cuando Lord Chamberlain regresó de sus entrevistas con Hitler les dijo a los británicos: he sentado las bases para la paz. Tiempo después Hitler invadió Polonia y empezó la segunda guerra mundial. ¿Qué situación retórica describe este texto?
> **ChatGPT-3.5:** El texto que has proporcionado es un ejemplo de ironía dramática. La ironía dramática es una figura retórica en la que el significado de una declaración es

> contrario a lo que parece ser en el contexto de la situación. En este caso, Lord Chamberlain regresó de sus entrevistas con Hitler y afirmó que había «sentado las bases para la paz», lo que implica que había logrado un acuerdo que evitaría la guerra. Sin embargo, la ironía dramática radica en que, en realidad, poco después de estas afirmaciones, Hitler invadió Polonia, lo que dio lugar al estallido de la Segunda Guerra Mundial, lo que contradice completamente la afirmación inicial de Chamberlain y revela la futilidad de sus esfuerzos por evitar la guerra.

Pero si le pides ejemplos de ironías dramáticas registradas en la última década te dice que sus conocimientos se detienen en septiembre de 2021 —ya sabemos que ahora puede responder con datos actuales— y que busques en periódicos, televisiones, etc. Ahora bien, si le pides ejemplos de construcciones irónicas, acierta con los ejemplos porque repite manuales: Un estudiante suspende un examen extremadamente fácil y alguien comenta: «¡Vaya genio académico que tenemos aquí!». En un día extremadamente caluroso, alguien exclama: «¡Oh, qué día tan fresco y agradable!». Ante un desastre total en la cocina, alguien podría decir: «Mis habilidades culinarias son legendarias», etc. Y terminará comentándote que «estos ejemplos ilustran cómo la ironía se utiliza para expresar lo contrario de lo que se quiere decir, a menudo con un toque de sarcasmo o humor».

7. La memoria. Y los riesgos

El pensamiento y la actuación lingüísticos están impregnados de memoria. Nuestro yo y nuestra actividad lingüística es en buena medida nuestra memoria; aunque sepamos que la memoria es selectiva y reconstructora y que, si recordáramos siempre todo lo que hemos pasado, no podríamos sobrevivir. Pero está ahí e informa nuestras actuaciones lingüísticas y nuestro yo. Una cosa que impresiona en *Blade runner* es que los replicantes tenían recuerdos implantados, porque se puede vivir sin muchas cosas, pero no se puede vivir sin recuerdos. El robot que nos reemplace debería tener memoria, pero no simplemente recuerdos implantados sino recuerdos construidos en interacción con la realidad. Es decir, el robot debería ser capaz de construirse una experiencia de vida, de la cual surge también la intuición, que tan importante papel juega en nuestra comprensión y creatividad. Debería tener pensamiento concreto y abstracto. Debería también estar de alguna manera inmerso en una realidad física. Vuelvo a recordar a OS2, el sistema operativo de *Her* del que se enamora perdidamente Joaquin Phoenix, y sus encuentros emocionados con nuevas situaciones a la vez que su dependencia de la experiencia construida con sus interlocutores virtuales, que le resultaba escasa. ¿Cómo se consigue todo esto? ¿Se entrena a un robot con escenas, textos, situaciones emocionales, situaciones sensoriales, cursos sobre todos los saberes inventados para que así construya una memoria que sería una mezcla de los seres históricos que conciben los inventores de los algoritmos? Si no se siguiera esta estrategia de construcción de las capacidades del robot humanoide, habría que buscar otra que, a mi modo de ver, debería estar guiada desde la biología del

cerebro y el cuerpo y no desde las solas capacidades de análisis paralelos de esquemas estadísticamente plausibles, distribuciones léxicas posibles o de la constante exposición a miles de millones de datos, capaces de producir tanto productos fascinantes como cualquier producto sesgado decidido por el programador. ¿Habrá que volver a los sistemas de reglas y abandonar los sistemas probabilísticos como dicen algunos? ¡Vaya a saberse! Como escribía Muñoz Molina en su artículo de *El País* del sábado 7 de octubre de 2023, «el porvenir no suele parecerse a ningún vaticinio». Pero más allá de vaticinios a largo plazo, o del anuncio de los riesgos existenciales para la humanidad que podrían derivarse de esta tecnología, cuando está tan sobre el tapete la cuestión de la regulación de la IA —con llamativa concentración en cuestiones biométricas y en la seguridad de los servicios de inteligencia—, convendría centrarse en los riesgos reales e inmediatos de estos productos: uso de datos sin consentimiento violando la propiedad intelectual, invasión de la privacidad, desinformación deliberada (*deep fakes*), textos científicos pirateados o híbridos, sesgos estereotípicos y discriminadores, entre otros muchos. Oliver (2023) señala algunos de estos problemas inmediatos y enfoca otros: «la manipulación subliminal del comportamiento humano por parte de algoritmos de IA; la falta de veracidad de los sistemas de IA generativa que inventan todo tipo de contenidos… sin correspondencia con el mundo real; la fragilidad de estos grandes modelos que pueden equivocarse y ser engañados; o la concentración de poder en las manos de un oligopolio de empresas y sus multimillonarios dueños o inversores». No es trivial, porque estas son amenazas sobre el presente.

Violeta Demonte es lingüista, catedrática emérita de la Universidad Autónoma de Madrid, miembro correspondiente de la RAE, y miembro de la Academia Europaea. En 2014 le fue concedido el Premio Nacional de Investigación Ramón Menéndez Pidal.

Referencias

Aladro, Eva, «Lo que la IA nos enseña sobre qué es verdaderamente la creatividad», *The Conversation*, 10 de mayo de 2023, https://theconversation.com/lo-que-la-inteligencia-artificial-nos-ensena-sobre-que-es-realmente-la-inteligencia-202673

Aydin, Ömer y Enis Karaarslan, «Is ChatGPT Leading Generative AI? What is Beyond Expectations?». 2023, *Academic Platform Journal of Engineering and Smart Systems.* 11:3, p. 119.

Biever, Celeste, «The easy intelligence tests that AI chatbots fail», *Nature* 619, 2023, p. 687.

Bringsjord, Selmer, Paul Bello, y David Ferrucci, «Creativity, the Turing Test, and the (Better) Lovelace Test», *Minds and Machines* 11: 1, 2001.

Castelvecchi, Davide, «Open-source AI chatbots are booming. What does this mean for researchers?», *Nature* 618, 2023.

Conroy, Gemma, «Scientists used ChatGPT to generate a whole paper from data», *Nature* 619, 20 de julio de 2023, p. 444.

Erk, Katrin y Aurelie Herbelot, «How to marry a start: probabilistic constrains for meaning in context», *arXiv*:2009.07936 [cs.CL], 2022.

Etxebarria Ecenarro, Víctor, «El desafío cuántico de la conciencia humana», *El Diario*, 23 de agosto de 2023, https://www.eldiario.es/tecnologia/desafio-cuantico-conciencia-humana_129_10413980.html

Gleitman, Lila y Anna Papafragou, «Relations between language and thought», en: D. Reisberg (ed.), *Handbook of Cognitive Psychology*, cap. 32, Nueva York: Oxford University Press, 2013, p. 505.

Haider, Hubert, *Is ChatGPT a grammatically competent informant?* Manuscrito inédito, Universidad de Salzburgo, 2023.

Innerarity, Daniel, «La histeria digital», *El País*, Opinión, 5 de agosto de 2023, https://elpais.com/opinion/2023-08-05/la-histeria-digital.html

Katzir, Roni, «Why large language models are poor theories of human linguistic cognition. A reply to Piantadosi 2023», manuscrito, Universidad de Tel Aviv, p. 5.

López de Mántaras, Ramón, entrevista en la sección de tecnología de *El País* de 6 de septiembre de 2023, https://elpais.com/tecnologia/2023-09-06/ramon-lopez-de-mantaras-experto-en-inteligencia-artificial-l

a-ia-sola-no-resolvera-absolutamente-nada-seran-los-humanos.html

Mai, Kimberly T., Sergi Bray, Toby Davies y Lewis D. Griffin, «Warning: Humans cannot reliably detect speech deepfakes», *PLoS ONE* 18:8, 2023.

Manning, Christopher, Kevin Clark, John Hewett, Urbashi Khandelwai y Omer Levi, «Emergent linguistic structures in artificial neuron-networks trained by self-supervision», *PNAS* 117: 48, 2020, p. 30047.

Marcus, Gary, «Noam Chomsky and GPT-3. Are Large Language Models good models of human language?» 2022 https://garymarcus.substack.com/p/noam-chomsky-and-gpt-3

Oliver, Nuria, *Inteligencia Artificial. Ficción, realidad y sueños*, Discurso de ingreso en la Real Academia de Ingeniería, RAI, Madrid, 11 de diciembre de 2018, p. 22 y 24.

Oliver, Nuria, «Una pausa cuestionable en la inteligencia artificial», *El País*, https://elpais.com/opinion/2023-05-03/una-pausa-cuestionable-en-la-inteligencia-artificial.html

Penrose, Roger, *The Emperor New's Mind*, Oxford, Oxford University Press, 1989. Trad. Cast.: *La nueva mente del Emperador*, Madrid, Mondadori, 1991, p. 30 y 525 (por la que cito).

Peterson, Erika y Christopher Potts, «Lexical semantics with Large Language Models: A case study of English *break*», *Findings of the Association for Computational Linguistics*, FACI, 2023, 490-511.

Piantadosi, Steven P., «Modern language models refute Chomsky's view of language», en: E. Gibson y M. Poliak, (eds.) *From fieldwork to linguistic theory: A tribute to Dan Everett*. Language Science Press, 2023, p. 19 y 6.

Searle, John, *Speech Acts. An essay in the Philosophy of Language,* Cambridge, Cambridge University Press, 1969, p. 38.

Turing, Alan, «Computing Machinery and Intelligence», *Mind* 59, 1950.

Wolfram, Stephen, *What is Chat GPT doing … and why does it work?* Wolfram Media, 2023, p. 10

---

[i]Estoy muy agradecida a José Manuel Sánchez Ron, quien me ofreció dar una charla sobre esta cuestión en el Museo de Navarra y me sugirió el título que ahora adopto como subtítulo de este texto. A José Antonio Millán, sabio amigo y experto, le debo sugerencias y correcciones muy valiosas para mí. Cristina Aranda me invitó a debatir con ella estos sobre estos asuntos en El Ateneo de Madrid, fue muy enriquecedor. A Elena Castroviejo le debo ideas, correcciones, diseño: una inestimable ayuda.

[ii]Para empezar quizá por lo más interesante habría que hablar de las tecnologías de imitación de voz. Mai et al. 2023 describen experimentos que muestran que un alto porcentaje de sujetos experimentales no distinguen la voz humana de la voz tecnológica, incluso aunque hayan sido entrenados para distinguirla. Son sorprendentes también los modelos de reconocimiento de voz. Lamentablemente, no tendré tiempo de detenerme en ellos.

[iii]Chat GPT Open AI no es el primer chat conversacional y, como sabemos, coexiste ahora con muchos otros. El primero de todos fue *Eliza* creado entre 1964 y 1966 por Joseph Wiezenbaum, profesor del MIT, para explorar la interacción entre humanos y máquinas.

[iv]Como GPT Open AI- 4 detecta y utiliza bibliografía real para construir sus repuestas, es de esperar que sus respuestas anticipen mejor y eviten alucinaciones.

[v]Piantadosi 2023, en un artículo tan interesante como en alguna medida incompleto y centrado hacia sus objetivos, sostiene que estos modelos de lenguaje refutan la visión chomskiana de la capacidad lingüística.

[vi]Los modelos transformadores (un tipo de red neuronal) fueron creados por Google, pero OpenAI, que es la que los usa, es una empresa independiente; ahora mismo financiada también por Microsoft.

[vii]La visión de Piantadosi y similares no parece tener en cuenta la estructura real de la noción de pobreza del estímulo que implica, como todo proceso biológico tres elementos imprescindibles: (i) dotación innata, (ii) exposición a datos lingüísticos, (iii) generalización y parametrización.

[viii]Aunque sea un inciso, quiero recordar que hay otras formas de pensamiento que no son verbales, Einstein decía que pensaba con imágenes: «Las palabras, el lenguaje, no parecen jugar ningún papel en mi mecanismo de pensamiento. Las entidades físicas que parecen servir como elementos del pensamiento son ciertos signos e imágenes» (tomado de Penrose 1989).