

MODOS DE COGNICIÓN

Implicancias para los
grandes modelos de
lenguaje (LLMs)

Katherine Hayles

Traducción: Gaspar Basualdo

Con la llegada de los grandes modelos de lenguaje —LLMs, por sus siglas en inglés—, como ChatGPT de OpenAI y otros afines, la inteligencia artificial ha ingresado en el espacio donde yo, como crítica literaria, y miles de personas como yo, pasamos la mayor parte de nuestro tiempo: el ámbito del lenguaje humano. Al mismo tiempo, la investigación sobre la cognición no consciente en formas de vida humanas y no humanas avanza rápidamente, con argumentos emergentes acerca de las capacidades cognitivas de plantas, células, robots e hibots —entidades híbridas que integran cuerpos vivos con mecanismos robóticos—. Nunca hubo tanto interés por la cognición, y nunca hubo tanta confusión terminológica en torno a cuestiones fundamentales, tales como si puede haber cognición sin cerebro, si los textos producidos por los LLMs tienen significados más allá de los que los lectores humanos proyectan sobre ellos, y si las IA como los LLMs son efectivamente inteligentes o si están diseñadas meramente para parecerlo.

En este texto propongo una definición de cognición, evalúo sus implicancias para distintos tipos de organismos y luego me concentro específicamente en los textos producidos por ChatGPT¹. Después de ofrecer información contextual sobre cómo la IA alcanza sus resultados, pregunto si sus respuestas son meras predicciones de la siguiente palabra, carentes de significado, y si tiene sentido —y, específicamente, qué tipo de sentido— pedirle que interprete un texto literario célebre por su ambigüedad: *The Figure in the Carpet*, de Henry James. Sostengo que los textos de ChatGPT son algo más que proyecciones probabilísticas, y que los LLMs sí poseen capacidades cognitivas. De hecho, desde mi punto de vista, son potencialmente la adaptación cultural/cognitiva más importante desde la invención del lenguaje. Como señalan Louise Amoore y sus coautores, los LLMs están penetrando con rapidez los sistemas sociales, económicos, políticos y financieros; en sus términos, los LLMs están creando un “modelo de mundo” que desplazará lo que denominan las “lógicas políticas” subyacentes de las culturas occidentales².

¹ Quiero agradecer a Ranjodh Singh Dhaliwal por sus útiles comentarios sobre este artículo.

² Amoore et al., “World Model”.

Hacia una definición de cognición

Podemos comenzar nuestra incursión en la densidad terminológica que rodea a la cognición examinando primero el caso especial de los seres humanos. Hasta donde sabemos, solo los humanos son capaces de pensamiento simbólico abstracto de manera sostenida, lo que nos permite demostrar teoremas matemáticos, componer sinfonías, esculpir objetos artísticos y escribir poemas³. Llamaré a tales actividades “pensar”, y se requiere de la conciencia para ponerlo en acto [*to enact*].

Sin embargo, el pensamiento es solo uno de los modos cognitivos de la humanidad, y probablemente no sea el más importante en nuestra vida cotidiana. También está activa la cognición implícita, que controla, entre otras capacidades, las funciones motoras y sensoriales. Definida habitualmente como la cognición que ocurre por fuera de la conciencia —es decir, por fuera de la experiencia subjetiva fenomenológica—, la cognición implícita garantiza que, una vez que hemos aprendido una actividad —como conducir un auto, andar en bicicleta o caminar hacia el trabajo—, podamos realizarla automáticamente. La cognición implícita trabaja de manera fluida junto con la conciencia; mientras caminamos, por ejemplo, podemos pensar en muchas cosas: qué queremos lograr en la reunión de hoy, o qué queremos decirle al jefe sobre haber llegado tarde. La cognición implícita tiene la ventaja de operar mucho más rápido que la conciencia. Mientras que la conciencia tarda medio segundo completo, desde el inicio de una sensación, en advertir qué está ocurriendo, la cognición implícita actúa en 200 milisegundos. Por eso los atletas competitivos practican sus movimientos una y otra vez, para que sus respuestas se automaticen mediante la cognición implícita y puedan reservar el ancho de banda limitado de la conciencia para tomar decisiones estratégicas durante el juego⁴.

³ Terrence Deacon sostiene esta afirmación en *Symbolic Species*, y en gran medida estoy de acuerdo con ella.

⁴ Libet y Kosslyn, *Mind Time*.

A pesar de su integración habitual, la conciencia y la cognición implícita pueden ser puestas en tensión mediante un experimento controlado, lo que permite a los investigadores determinar qué aporta cada una a nuestra cognición general. Scott Albert, John Krakauer y sus colegas, que trabajan con pacientes con discapacidad motora en la Universidad Johns Hopkins, diseñaron un experimento en el que se pedía a los sujetos que movieran un cursor en la pantalla hacia un punto utilizando un joystick. Intuitivamente, cabría esperar que mover el joystick hacia arriba desplazara el cursor hacia arriba⁵; pero en este experimento había rotaciones de 20, 40 y 60 grados respecto de la correspondencia uno a uno entre joystick y cursor. Para dominar el ejercicio, el paciente debía activar el control consciente. Los ensayos repetidos mostraron que el aprendizaje implícito alcanzaba un estado estable, independientemente del tamaño de la rotación, mientras que el aprendizaje explícito requería más tiempo a medida que aumentaba el grado de rotación. Esta diferencia permitió a los investigadores mostrar que, cuanto mayor era el aprendizaje explícito, más tardaba el aprendizaje implícito en alcanzar un estado estable. Interpretaron este resultado como prueba de que el aprendizaje explícito “drena” recursos necesarios para el aprendizaje implícito, en este caso, el error de retroalimentación⁶. Más allá de esto, sus experimentos demostraron de manera concluyente la existencia de dos tipos de sistemas de aprendizaje, vinculados respectivamente con la cognición implícita y la cognición consciente.

Además de la cognición implícita, los experimentos han mostrado que la cognición no consciente también opera en las acciones humanas. Cuando se les mostraba información visual ruidosa en la que estaban incrustados patrones sutiles, los sujetos eran completamente incapaces de articular conscientemente cuáles eran esos patrones. Sin embargo, a juzgar por el tiempo que tardaban en responder, los experimentos mostraban que habían aprendido a anticiparlos⁷. La cognición no consciente es la capacidad cognitiva que entra en juego cuando las personas evitan pisar serpientes, ya que reaccionan mucho más rápido de lo que la conciencia podría

⁵ Albert et al., “Parallel Sensorimotor Learning Systems”.

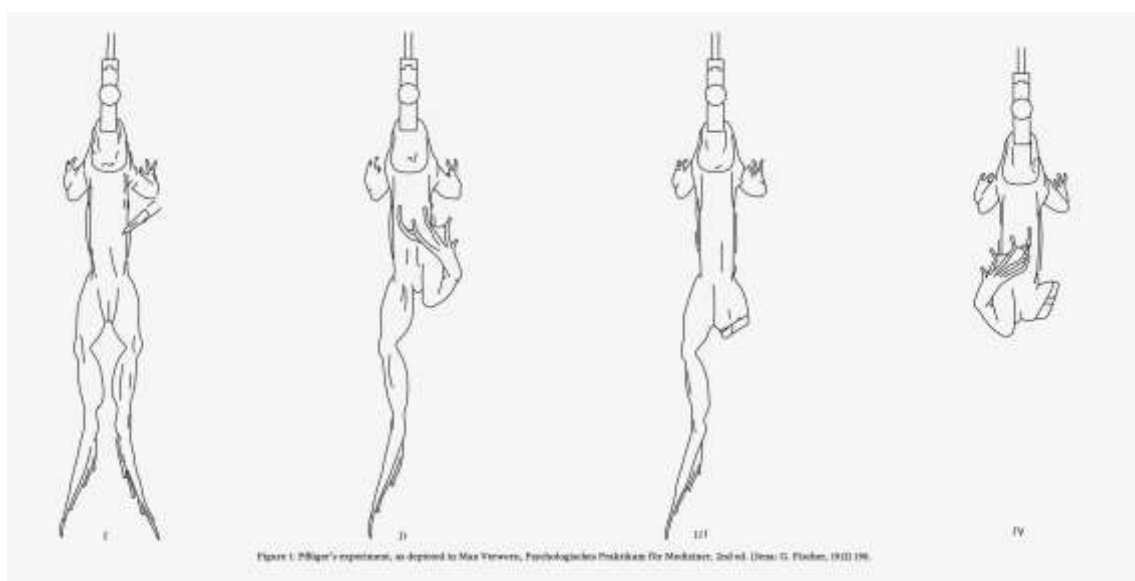
⁶ Albert et al.

⁷ Dresch-Langley, “Non-Conscious Representations”.

gestionar. De manera similar a cómo se descargan tareas cognitivas en la cognición implícita, la conciencia puede descargar tareas en la cognición no consciente; por ejemplo, cuando un maestro de ajedrez aprende a captar de un vistazo los patrones de un tablero sin necesidad de registrar conscientemente la ubicación de cada pieza⁸. En un entorno complejo, con múltiples estímulos, el reconocimiento de patrones es una capacidad importante, y los investigadores especulan que evolucionó primero, con la conciencia construida posteriormente sobre ella.

La imagen que emerge de estas observaciones puede diagramarse como una pirámide de respuestas cognitivas: en la cima se encuentra el pensamiento, íntimamente asociado con la conciencia, con su captación lenta y su ancho de banda limitado, pero con capacidades incomparables para resolver problemas novedosos; por debajo se encuentra la cognición implícita, que controla las respuestas motoras y sensoriales; y más abajo se encuentra la cognición no consciente, con su extraordinaria capacidad de reconocimiento de patrones y su rápida captación, aunque con medios limitados para afrontar nuevos desafíos⁹.

Criterios de cognición: SIRAL



⁸ Grassini et al., “Subjective Visual Awareness”; Van Le et al., “Pulvinar Neurons”.

⁹ Hayles, *Unthought*, 14–16.

Estas observaciones ofrecen indicios sobre los criterios a partir de los cuales una conducta puede ser juzgada como cognitiva o no. Desde luego, no existe un conjunto “correcto” de criterios, pero podemos evaluar su eficacia preguntando si resulta útil. Un conjunto útil debería poder distinguir entre cognición y adaptación. Si bien ambas son ejemplos de emergencia evolutiva, las adaptaciones son respuestas automáticas inflexibles, mientras que las conductas cognitivas demuestran flexibilidad y capacidad de aprendizaje. Las definiciones posibles de cognición deberían incluir las cogniciones implícitas y no conscientes, pero excluir las adaptaciones inflexibles, así como dispositivos tales como los mecanismos homeostáticos, que operan exclusivamente mediante ciclos automáticos de retroalimentación.

En *Lo impensado* propuse esta definición general de cognición: “un proceso que interpreta información dentro de contextos que la conectan con el **significado**”¹⁰. A partir de estas generalidades, quiero desarrollar esta definición y volverla más precisa, identificando cinco criterios que debe exhibir la conducta de un organismo vivo para ser considerada cognitiva: *senzar, interpretar, responder flexiblemente, anticipar y aprender [sensing, interpreting, responding flexibly, anticipating, learning]* —SIRAL—

Senzar significa, simplemente, que el organismo puede recibir información del ambiente. Como observó la bióloga Lynn Margulis, incluso el organismo más simple, como una bacteria, debe ser capaz de senzar información proveniente de ambientes fluctuantes e inciertos para poder continuar existiendo¹¹. Las capacidades sensoriales de un organismo determinan el modelo de mundo que construirá. El biólogo alemán

¹⁰ Hayles, 22.

¹¹ En *What Is Life?*, Lynn Margulis y Dorion Sagan escriben: “Todos los seres vivos, no solo los animales, sino también las plantas y los microorganismos, perciben. Para sobrevivir, un ser orgánico debe percibir: debe buscar, o al menos reconocer, alimento y evitar los peligros ambientales” (32).

Jakob von Uexküll llamó a esto el Umwelt de un organismo —traducido de manera aproximada como su mundo circundante—¹². El maravilloso libro de Ed Yong, *Un mundo inmenso*, analiza con exquisito detalle los tipos de capacidades que construyen distintos organismos: la visión ultravioleta de los insectos, que les permite aterrizar con precisión sobre las flores que polinizan; las vibraciones de superficie que insectos como los membrácidos, los grillos y otras 200.000 especies de insectos crean para comunicarse; los sonidos que guían a las lechuzas de campanario hacia los ratones; y los campos magnéticos que permiten a las polillas bogong y a muchas especies de aves orientarse a grandes distancias¹³. Cada capacidad es específica de una especie y evolucionó durante eones para permitir que el organismo sobreviviera y se reprodujera¹⁴.

A veces hay solapamientos entre distintos Umwelten. Los perros y los humanos han sido compañeros durante miles de años, en parte por las zonas que comparten —ambos ven un conejo— y también por sus diferencias, que permiten una cooperación constructiva: por ejemplo, el sentido del olfato superior de los perros se utiliza para todo, desde encontrar personas perdidas hasta detectar artefactos militares. Sin embargo, el Umwelt de una especie nunca es exactamente igual al de otra. Cada una vive dentro del modelo de mundo que determina qué tipos de información sensorial encontrará significativos.

Hay una distinción importante entre vivir dentro del propio Umwelt y observarlo desde afuera. El clásico texto de Thomas Nagel, “¿Qué se siente ser un murciélago?”, sostenía que, por mucho que los humanos aprendamos sobre los murciélagos —sus capacidades de sonar, sus hábitos de caza, sus formas de socialidad—, permanece una brecha inevitable entre experimentar el Umwelt de un murciélago como el mundo en el que se vive y aprehender sus cualidades a partir de datos empíricos. El subtítulo del libro de Yong sugiere esta distinción cuando alude a los sentidos animales que

¹² Von Uexküll, *Foray into the Worlds*.

¹³ Yong, *La inmensidad del mundo*

¹⁴ Vale la pena considerar si el Umwelt humano posee subcategorías en su interior, como aquellas asociadas con personas neurodivergentes, por ejemplo, el autismo.

“revelan los reinos ocultos que nos **rodean**”¹⁵. Es decir, podemos aprender sobre sus Umwelten, pero nosotros, los humanos, nunca podemos simplemente vivir dentro de ellos como vivimos dentro del nuestro. Esta distinción será importante más adelante, en la discusión sobre los Umwelten de los LLMs.

El segundo requisito, interpretar, implica que buena parte del procesamiento cognitivo ocurre en el nivel de la percepción sensorial; en los organismos conscientes, esto sucede mucho antes de que entre en juego la percepción consciente. La investigación clásica que lo expone con claridad es el artículo, citado con frecuencia, “Lo que el ojo de la rana le dice al cerebro de la **rana**”¹⁶. Este es el resumen que los autores hacen de sus hallazgos: “¿**Cuáles** son las consecuencias de este trabajo? Fundamentalmente, muestra que el ojo le habla al cerebro en un lenguaje ya altamente organizado e interpretado, en lugar de transmitir una copia más o menos exacta de la distribución de la luz sobre los **receptores**”¹⁷. Dado que la interpretación involucra procesamiento cognitivo, implica seleccionar entre opciones y, por lo tanto, entraña la posibilidad de error. Una interpretación puede ser equivocada. Si no hay elección ni selección, entonces la respuesta es una cadena causal directa y resulta más adecuado clasificarla como una adaptación.

La respuesta, el tercer criterio, designa una conducta evocada por un estímulo ambiental. En los humanos, esto puede implicar el uso de lenguaje simbólico —por ejemplo, usted podría decidir escribirme un correo electrónico después de leer este párrafo—. En el caso de los organismos no humanos, la biosemiótica —la ciencia de los signos utilizados por organismos biológicos no humanos— ha desarrollado una comprensión de las conductas no humanas como signos que funcionan como representaciones ¹⁸. A partir de la semiótica de Charles Sanders Peirce, los biosemióticos consideran una conducta como una representación —en el vocabulario de Peirce, un “representamen”—, conectada con una señal ambiental mediante un

¹⁵ Nagel, “What Is It Like”.

¹⁶ Lettvin et al., “Frog’s Eye”.

¹⁷ Lettvin et al.

¹⁸ Deacon, *Incomplete Nature*; Hoffmeyer, *Biosemiotics*; Hoffmeyer, *Signs of Meaning*.

interpretante, que aquí puede entenderse como el procesamiento cognitivo que ocurre entre la aparición de los estímulos sensoriales y la conciencia¹⁹. La significación de la representación consiste en anticipar algo que todavía no está presente, pero que se espera. Cuando un árbol caducifolio pierde sus hojas en respuesta al descenso de la temperatura promedio, esta acción funciona como un signo que señala la proximidad del invierno: aquello representado o anticipado por la acción. Situar las conductas no humanas como signos es un movimiento importante, porque abre el territorio de la producción de significado más allá de las significaciones humanas, hacia conductas originadas en especies no humanas²⁰.

El requisito de que la respuesta sea flexible implica que existan opciones: por eso elimina los mecanismos homeostáticos simples y las adaptaciones biológicas automáticas. Aquí hace falta una nota de cautela, ya que conductas que tradicionalmente fueron consideradas adaptaciones pueden mostrar flexibilidades antes inadvertidas. Yong, por ejemplo, relata una anécdota aportada por Karen Warkentin, quien observaba la eclosión de huevos de ranas arborícolas en el Parque Nacional Corcovado, en Costa Rica. Notó que, cuando golpeaba accidentalmente una nidada de huevos, algunos embriones de rana eclosionaban rápidamente. Aunque la eclosión suele haber sido considerada una adaptación automática, ella pensó que quizá los huevos estaban eclosionando antes de tiempo en respuesta a peligros ambientales. Entonces, junto con sus colegas, recolectó huevos y los alojó en jaulas con serpientes de ojos de gato, sus depredadores naturales. Confirmaron que los embriones pueden eclosionar antes de tiempo cuando son atacados. Incluso “los vio salir de huevos que estaban sostenidos en la boca de una **serpiente**”²¹. Esta imagen vívida muestra la cognición en acto. La programación genética coexiste con posibilidades cognitivas activadas cuando se cierne la amenaza de una muerte inmediata.

¹⁹ Peirce, *The Essential Peirce: Selected Philosophical Writings. Volume 1*.

²⁰ Hayles, *Bacteria to AI*.

²¹ Yong, *Immense World*, 188–89.

La anticipación, el cuarto requisito, es crucial para la supervivencia de un organismo: la preparación para el futuro rinde grandes beneficios al permitir que los organismos lidien con fluctuaciones ambientales, depredaciones inminentes y la progresión ordenada de los días y las estaciones. Los organismos con cerebro demuestran claramente conductas anticipatorias: el perro que espera junto a la puerta cada tarde el regreso de su compañero infantil de la escuela; la orca que sopla burbujas anticipando que esta acción hará que los peces presa se agrupen, de modo que puedan ser capturados con mayor facilidad; el reno líder cuya conducta señala a la manada que es momento de trasladarse a los pastizales de invierno.

Como sugiere este último ejemplo, las conductas anticipatorias son cruciales no solo porque hacen que ciertas cosas ocurran, sino también porque abren vías hacia el futuro para los organismos no humanos. Todos los organismos registran el pasado en sus cuerpos: las arrugas en un rostro, los anillos en un árbol, el crecimiento en la concha de un nautilus. Todos los organismos, incluidos los humanos, viven en el presente. Pero sin signos, los animales no humanos no tendrían forma de comunicar algo acerca del futuro, lo que disminuiría sus probabilidades de supervivencia. Pensemos que, cuando un puercoespín eriza sus púas, esa conducta funciona como una señal de advertencia para un depredador acerca de posibles acciones futuras. Sin este signo anticipatorio, la aptitud adaptativa del puercoespín disminuiría. Conceptualmente hablando, considerar las conductas puestas en acto [*enacted*] como signos proporciona un fundamento crucial para expandir el ámbito de la cognición a todas las criaturas vivientes ²².

Aunque las plantas no tienen neuronas, también anticipan acontecimientos. Los girasoles giran sus cabezas y hojas para seguir al sol, con una desviación de menos de 15 grados por delante o por detrás. Paco Calvo informa que, si un girasol joven es rotado 180 grados durante la noche, al cabo de unos días ajustará su floración al nuevo ángulo formado con el sol. Comenta: “Las plantas no solo responden a lo que ocurre a su alrededor; podrían tener un modelo interno de lo que el sol va a hacer, que

²² Hayles, *Bacteria to AI*.

guía sus **movimientos**”²³. Como señala Calvo, puede inferirse evidencia de apoyo a partir de las conductas nocturnas de las plantas. La malva de Cornualles, o Lavatera, anticipa por dónde saldrá el sol y gira sus hojas para orientarlas hacia allí como preparación, “logrando hacerlo durante algunos días incluso si se la priva de luz solar”²⁴. El girasol rotado mencionado más arriba gira sus hojas y su cabeza al doble de la velocidad con que las gira durante el día, anticipando el restablecimiento de su orientación óptima respecto del sol.

Modelar el mundo: lo que SIRAL habilita

La sugerencia de Calvo de que las plantas no solo responden a estímulos, sino que crean modelos, merece ser explorada con mayor profundidad. Como se indicó antes, todo organismo construye su Umwelt específico a través de sus capacidades sensoriales, musculares, neurológicas y/o cognitivas. Como señaló Donna Haraway hace décadas, no existe una mirada “objetiva” desde el ojo de Dios que vea el mundo tal como “realmente es”, sino una miríada de perspectivas que han evolucionado para permitir que los organismos sobrevivan y se reproduzcan²⁵. Durante demasiado tiempo, las perspectivas humanas dominaron las conversaciones sobre la naturaleza de la realidad, lo que condujo a supuestos —quizá inconscientes— según los cuales el Umwelt humano sería de algún modo más preciso, más verdadero, más “real” que el de otras especies. Pero sabemos que muchas especies poseen sentidos superiores a los nuestros, y también sabemos que muchos organismos pueden realizar proezas de fuerza, resistencia y percepción que superan ampliamente las capacidades humanas. Insistir en la ubicuidad de los modelos y del modelado —los Umwelten— a lo largo del espectro biológico, y más allá de él, ayuda a cuestionar el supuesto de la superioridad

²³ Calvo, *Planta Sapiens*, 74.

²⁴ Calvo, 74.

²⁵ Haraway, “Situated Knowledges”.

humana y a restituir una visión más precisa, más sensata y más humilde del lugar que ocupan los humanos en las complejas ecologías de la vida en la Tierra²⁶.

Por qué los modelos son importantes

¿Qué ventajas otorgan los modelos y cómo se relacionan con SIRAL? Un modelo puede considerarse una generalización que posee poder predictivo y beneficios para la supervivencia. Se origina en el sensado y la interpretación de información ambiental, pero va más allá de instancias específicas para anticipar cómo se desplegarán los acontecimientos futuros. En cierto sentido, un modelo es el resultado neto de la interacción de todos los componentes de SIRAL: representa lo que un organismo ha aprendido sobre su ambiente —ya sea como individuo o, en el tiempo evolutivo, como especie— y proporciona las anticipaciones que le permiten sobrevivir dentro de su nicho ambiental. La leona que se desliza entre la hierba, agazapándose para permanecer sin ser detectada el mayor tiempo posible por la gacela que está cazando, ha desarrollado evolutivamente un modelo de lo que hará la gacela cuando sea perseguida. Ella sabe que, en cuanto la presa detecte su presencia, comenzará a alejarse a saltos. También sabe que probablemente no correrá en línea recta, sino que hará zigzags para evadir sus intentos de derribarla. En este caso, su modelo incluye no solo la naturaleza de la hierba y de otros rasgos ambientales, sino también cómo reaccionarán otras especies. Mientras tanto, la gacela también ha desarrollado evolutivamente un modelo de la leona, sabiendo que la depredadora intentará matarla y comerla si puede; el modelo de la gacela incluye también las maniobras evasivas que a veces pueden evitar el desastre.

Es fácil ver cómo los organismos con cerebro construyen modelos, pero ¿qué ocurre con los sistemas mínimamente cognitivos? En *Bacteria to AI: Human Futures with Our*

²⁶ Hayles, *Bacteria to AI*.

Nonhuman Symbionts (2025), analizo la teoría de los marcos de referencia de Chris Fields y Michael Levin²⁷. Basándose en el trabajo de Maturana y Varela, entre otros, adoptan la concepción encarnada, embebida, enactiva y extendida de la cognición — el “enfoque 4E”. Ponen el acento en “significados que son capacidades estructurales y funcionales, pero en la mayoría de los casos explícitamente no representacionales, de un sistema **encarnado**”²⁸. De este modo, dejan claro que su teoría está pensada para aplicarse a sistemas mínimamente cognitivos que no crean representaciones como tales, pero que, no obstante, construyen modelos de sus ambientes.

Al enfatizar que todo organismo construye un modelo de su ambiente, discuten en términos generales cómo surgen las distinciones entre los componentes ambientales. Los “objetos” que percibe un organismo no existen antes del acto de percepción; más bien, emergen a través de percepciones y dentro de percepciones que ocurren en el interior de un Umwelt.

Los “cortes” que separan el mundo observado de cualquier sistema en “objetos” son puramente epistémicos y, por lo tanto, relativos al sistema que realiza las observaciones. Comprender qué “objetos” ve S —un organismo dado— como componentes de su E —ambiente— requiere, entonces, examinar la dinámica interna de S. Esa dinámica interna, junto con la interacción sistema-ambiente, determina por completo qué “objetos” ambientales es capaz S de segregar del “fondo” de E e identificar como potencialmente significativos. Que a S le resulte útil segregar “objetos” del “fondo” de este modo no está determinado por la dinámica interna de S, sino por la de E. El significado es, entonces, un juego de dos jugadores, no de uno solo. En este sentido, S está fundamentalmente “embebido” [*embedded*]... En el lenguaje de la teoría evolutiva, siempre es E quien selecciona los significados —o las acciones que estos habilitan— que tienen de hecho utilidad para S, y descarta aquellos que no la tienen²⁹.

²⁷ Fields y Levin, “How Do Living Systems”.

²⁸ Fields y Levin, 2.

²⁹ Fields y Levin, 4.

Para ilustrar esta idea de un sistema mínimamente cognitivo, Fields y Levin consideran una bacteria *E. coli* que realiza quimiotaxis, es decir, que se aleja de una toxina o se acerca a una fuente de alimento siguiendo un gradiente químico. Así describo su teoría en *Bacteria to AI*:

Aunque los mecanismos de sensado que desarrollan el marco de referencia son complejos —tal como lo determinó la investigación previa sobre *E. coli*, que estudia mecanismos de señalización química como los canales iónicos—, el punto es relativamente simple: la bacteria debe tener un marco de referencia interno —RF— para poder realizar las distinciones activadas en la quimiotaxis. Además, “implementar RF internos requiere un aporte energético del ambiente. Este aporte energético es necesariamente mayor que la energía requerida para cambiar el estado del indicador asociado con el RF. Por lo tanto, todo RF es un sistema disipativo que consume energía libre ambiental y devuelve calor residual al ambiente. Cada RF que implementa un organismo requiere recursos metabólicos específicos” (Fields y Levin, 5, PDF). Dado que los RF son energéticamente costosos, los autores concluyen que “solo las diferencias significativas son **detectables**”, ya que “los organismos no desperdician energía adquiriendo información que no es **accionable**” (6, PDF).

Concluyen que “en todos los niveles, los RF especifican la accionabilidad y, por lo tanto, el significado”³⁰. Sugieren además que explorar la emergencia de los RF internos y sus vínculos con los RF externos puede ofrecer respuestas a la “pregunta fundamental de una teoría evolutiva de la cognición”³¹.

³⁰ Fields y Levin, 7. La idea de que “los organismos no desperdician energía adquiriendo información que no es accionable” es, sin duda, verdadera para organismos relativamente simples como las bacterias. Sin embargo, en organismos más complejos, como los mamíferos, muchas conductas muestran que investigan cosas que no tienen una utilidad inmediata, como la proverbial curiosidad de los gatos. En mi experiencia, las vacas también sienten mucha curiosidad ante los cambios en su ambiente.

³¹ Fields y Levin, 6.

La “pregunta **fundamental**” a la que se refieren también es central para mi argumento aquí. Dada la evidencia de que incluso los organismos unicelulares son mínimamente cognitivos, ¿cómo evolucionaron esas capacidades, mucho antes de que los cerebros aparecieran en el paisaje evolutivo? Desde luego, los cerebros son importantes, pero los organismos sin tejidos neuronales son mucho más numerosos en los ecosistemas terrestres que las criaturas neuronales; por eso, una visión equilibrada de la cognición exige que también se estudien sus enacciones. Investigaciones importantes sobre bacterias y plantas ofrecen pistas para responder algunos aspectos de esta pregunta fundamental.

Anticipación y aprendizaje en organismos mínimamente cognitivos

La investigación sobre organismos unicelulares muestra que pueden presentar conductas anticipatorias. El biólogo israelí Yitzhak Pilpel y su equipo del Instituto Weizmann de Ciencias encontraron en bacterias *E. coli* y en levadura de vino (*S. cerevisiae*) una conducta análoga a la respuesta condicionada de Pavlov. Mostraron que las bacterias evolucionaron para “anticipar estímulos ambientales adaptándose al orden temporal de su **aparición**”³². Cuando les dieron a las bacterias lactosa seguida de maltosa, encontraron que, después de varias generaciones, las bacterias evolucionaron hasta activar la red genética destinada a utilizar maltosa cuando detectaban lactosa. Cuando los investigadores cambiaron el orden y les dieron primero maltosa, no encontraron activación de los genes de la lactosa. Además, cuando dejaron de dar maltosa después de la lactosa durante varias generaciones bacterianas, la activación de la maltosa cesó, tal como ocurrió cuando Pavlov eliminó una respuesta condicionada en sus perros al no hacer seguir el sonido de una campana con alimento.

³² Mitchell et al., “Adaptive Prediction”, 220.

En cuanto a la levadura de vino, a medida que progresa la fermentación, el ambiente se calienta. Pilpel y su equipo descubrieron que, cuando las levaduras sienten por primera vez el calor, comienzan a activar genes para lidiar con el aumento de temperatura y con los estreses que le siguen. Subrayan que, al igual que la respuesta bacteriana, esta adaptación está mediada genéticamente, de modo que seguiría contando como una adaptación y no como una conducta plenamente cognitiva. Sin embargo, sugiere que la frontera entre una adaptación genética y una respuesta aprendida quizá no sea tan tajante como se pensaba, ya que solo se necesitan unas pocas generaciones bacterianas para incorporar aprendizaje; la eclosión prematura de los huevos de rana lo demostró en otro contexto.

El aprendizaje, el quinto requisito, significa que el organismo puede modificar sus conductas como resultado de experiencias previas. Es evidente que los animales con cerebro pueden aprender —aunque ciertos políticos a veces hagan dudar de ello—, pero existe cada vez más evidencia de que las plantas también pueden hacerlo. El campo emergente de la neurobiología vegetal ha mostrado que las plantas recogen información de sus ambientes, recuerdan encuentros previos y responden de manera flexible y adaptativa a condiciones cambiantes. Ya en 2006, un artículo de revisión en *Trends in Plant Science* comentaba que las plantas participan en un “sistema integrado de señalización, comunicación y **respuesta**”, que les permite tomar decisiones tales como “cuándo y dónde buscar nutrientes, y dónde asignar esos **nutrientes**... cuándo y qué órganos generar o dejar envejecer; cuándo reproducirse y cuánta descendencia producir; cómo montar una defensa frente a un ataque y en qué tejidos u órganos; y cuándo y dónde transmitir señales químicas a los organismos **circundantes**”³³. De hecho, las capacidades comunicativas, decisionales y adaptativas de las plantas se han vuelto tan conocidas que inspiraron la novela ganadora del Premio Pulitzer de Richard Powers, *The Overstory*³⁴.

A continuación, algunos de los experimentos que se han centrado específicamente en el aprendizaje. Monica Gagliano y su equipo de la Universidad de Australia Occidental

³³ Brenner et al., “Plant Neurobiology”, 413.

³⁴ Powers, *Overstory*.

diseñaron experimentos para demostrar capacidad de memoria y aprendizaje. En una serie de pruebas, evaluaron la capacidad de una planta para habituarse a un estímulo que las plantas habían aprendido que no era dañino y recordar esa lección durante varias semanas³⁵. Utilizaron un mecanismo para dejar caer plantas sensitivas (*Mimosa pudica*) tanto en ambientes de baja luz como de alta luz, una acción que inicialmente hacía que estas plantas plegaran sus hojas. Como el plegamiento de las hojas reduce la capacidad de la planta para aprovechar la luz solar, los investigadores plantearon la hipótesis de que las plantas en condiciones de baja iluminación se habituarían más rápido y recordarían la lección durante más tiempo: dado que la baja luz vuelve más desafiante el ambiente, plegar las hojas allí implica un riesgo mayor. En efecto, las plantas entrenadas con baja luz aprendieron a habituarse más rápidamente y recordaron la lección durante más tiempo, mostrando la respuesta de habituación aprendida incluso después de haber permanecido intactas durante un mes en un ambiente de alta luz.

Una forma más avanzada de aprendizaje es la asociativa, en la que un organismo aprende a asociar un estímulo condicionado con uno no condicionado, como el sonido de una campana con la comida en los experimentos de Pavlov con perros. En experimentos diseñados para probar si las plantas pueden demostrar aprendizaje asociativo, Gagliano y sus colaboradores entrenaron plantas jóvenes de arveja en laberintos en forma de Y durante tres días, con ventiladores y luces fijados en la parte superior del laberinto³⁶. Las plantas de arveja son ideales porque las plantas jóvenes crecen produciendo un único zarcillo, lo que facilita registrar cómo avanza el crecimiento. Las plantas preferían el brazo con más luz, que también estaba asociado con los ventiladores. Luego, el equipo de investigación utilizó solo los ventiladores para probar si las plantas elegirían la misma rama que habían elegido cuando las luces estaban presentes. Encontraron que la dirección del crecimiento era afectada por los

³⁵ Gagliano et al., “Experience Teaches **Plants**”.

³⁶ Gagliano et al., “Learning by **Association**”.

ventiladores en las plantas entrenadas, mientras que no se observó tal preferencia en las plantas no entrenadas³⁷.

A Suzanne W. Simard, botánica canadiense, se le atribuye el descubrimiento de la llamada “red amplia del **bosque**” (wood wide web) en su investigación doctoral. Actualmente, docente en la University of British Columbia, resume su investigación en un capítulo titulado “Las redes micorrícicas facilitan la comunicación, el aprendizaje y la memoria de los **árboles**”³⁸. Descubrió que las redes fúngicas micorrícicas conectan a los árboles del bosque, facilitando “la comunicación entre árboles, el intercambio de recursos, la defensa y el reconocimiento de **parentesco**”. Los hongos envían filamentos diminutos, casi invisibles, llamados “hifas”, que penetran las raíces de los árboles y actúan como hilos de comunicación que transportan mensajes entre árboles.

El crecimiento de las hifas también ayuda a los árboles a buscar los nutrientes que necesitan para sobrevivir. Como las hifas son tan pequeñas, su crecimiento requiere una inversión de recursos mucho menor que la de las raíces; ese crecimiento implica “conductas cognitivas tales como toma de decisiones, movimientos de búsqueda y escape, y reconocimiento de **vecinos**”. Los árboles y los hongos colaboran proporcionándose nutrientes mutuamente: “El hongo micorrícico intercambia los nutrientes que obtiene del suelo mediante su micelio extramatricial por el fotosintato fijado por la **planta**”. La red fúngica también participa en la distribución de nutrientes: “Se cree que las señales bioquímicas que se transmiten entre árboles a través de los vínculos fúngicos proporcionan subsidios de recursos a los receptores, particularmente entre plántulas en **regeneración**”, actuando así como una forma de reconocimiento de parentesco.

³⁷ Aunque otro investigador no pudo reproducir sus resultados —Markel, “Lack of **Evidence**”—, en “Comment on ‘Lack of **Evidence**’”, Gagliano et al. señalan que el estudio de replicación utilizó condiciones lumínicas diferentes de las empleadas por ellos, de modo que la luz ya no funcionaba como estímulo no condicionado y, por lo tanto, no suscitaba una respuesta condicionada.

³⁸ Simard, “Mycorrhizal Networks”. “**Mycorrhizal**” deriva de las palabras griegas para hongo (mykós) y raíz (rhiza).

En el capítulo, Simard presenta evidencia de que “la topología de las redes micorrícicas es similar a la de las redes neuronales, con patrones libres de escala y propiedades de mundo pequeño que se correlacionan con eficiencias locales y globales importantes para la **inteligencia**”. Las complejas actividades cognitivas del bioma árbol-hongo incluyen “capacidades de percepción, aprendizaje y memoria, e influyen en rasgos de las plantas indicativos de aptitud **adaptativa**”. Además, el bioma árbol-hongo está ubicado dentro de una ecología forestal más amplia en la que existen “interacciones colectivas basadas en la memoria entre árboles, hongos, salmones, osos y personas, que mejoran la salud de todo el ecosistema **forestal**”. En conclusión, sugiere que estos conocimientos sobre los complejos Umwelten de las especies vinculadas al bosque, vistos “a través de la lente de la cognición arbórea, las colaboraciones del microbioma y la inteligencia **forestal**”, tienen el potencial de transformar la manera en que las personas piensan los bosques, contribuyendo “a un enfoque más holístico para estudiar los ecosistemas y a una mayor empatía y cuidado humanos por la salud de nuestros **bosques**”³⁹.

A medida que avanza la investigación sobre cognición vegetal, se acumula evidencia de que las plantas —por ejemplo, los árboles de un bosque— forman comunidades de comunicación y apoyo mutuo, un resultado empírico que Simard asume sin ambigüedades para las ecologías forestales que estudia. Como se señaló antes, las plantas no tienen neuronas, pero sus tejidos generan potenciales de acción, aunque estos son mucho más lentos que los de las células neuronales⁴⁰. Además, sus capacidades para sentir el ambiente estaban bien documentadas incluso antes del descubrimiento de Simard de las redes micorrícicas en los bosques. Si una planta es atacada por parásitos, por ejemplo, una planta vecina de la misma especie incrementará su producción de sustancias químicas que disuaden a los parásitos de atacarla⁴¹. Si plantas de especies diferentes son colocadas en la misma maceta, sus raíces competirán por nutrientes; pero si las plantas son de la misma especie,

³⁹ Simard, 191–192.

⁴⁰ Stahlberg, “Historical Overview”.

⁴¹ War et al., “Mechanisms of Plant Defense”.

tenderán a cooperar más que a competir, demostrando así conductas que, si fueran exhibidas por animales, serían llamadas selección de parentesco⁴². Otros ejemplos muestran que los repertorios conductuales de las plantas incluyen acciones determinadas por experiencias previas de una planta específica, lo que permite distinguir entre respuestas a nivel de especie y experiencias particulares de aprendizaje⁴³. Aunque los mecanismos de sensado ambiental de las plantas no se comprenden bien, podrían involucrar canales de iones de calcio⁴⁴. Estos resultados se corresponden bien con los criterios SIRAL y conducen a la conclusión de que las plantas deberían ser consideradas sistemas mínimamente cognitivos.

Un ejemplo no-SIRAL: ranas sin cerebro

Consideremos ahora un ejemplo negativo para ver hasta qué punto los criterios SIRAL permiten distinguir entre acciones cognitivas y no cognitivas. En un experimento bastante extraño diseñado por Eduard Friedrich Wilhelm Pflüger en 1853, a una serie de ranas se les extrajo o destruyó el cerebro —“**pithed**”, como lo formuló Pflüger—, y se les aplicó en el vientre papel de filtro empapado en ácido acético⁴⁵. Aunque cada rana carecía de cerebro, su pata respondía tocando el lugar donde estaba el ácido, intentando limpiarlo. Si esa pata era cortada, entonces la otra pata realizaba el mismo movimiento⁴⁶. En términos de SIRAL, las ranas recibían información de su ambiente y, utilizando recursos no cerebrales, la interpretaban —el artículo sobre “el ojo de la rana” demostró cuánto procesamiento cognitivo ocurre localmente, antes de que la información llegue al cerebro—. Luego respondían, pero su respuesta era inflexible y automática. La conducta adaptativa demostrada era no cognitiva, porque no exhibía flexibilidad, anticipación ni aprendizaje.

⁴² Calvo, *Planta Sapiens*, 85.

⁴³ Calvo, 86–87.

⁴⁴ Gagliano et al., “Alternative Means of **Communication**”.

⁴⁵ Agradezco a Simon De Deo por haber llamado mi atención sobre este experimento en su presentación en la conferencia “Other **Minds**”, realizada en la Universidad Estatal de Arizona, el 5 de mayo de 2024.

⁴⁶ Verworn, *Physiologisches Praktikum*, 198.

Es importante advertir que este ejemplo no sugiere que los cerebros sean necesarios para la cognición. Normalmente, las ranas tienen cerebro —obviamente—, y sus capacidades cognitivas han evolucionado para operar a través de sus tejidos neuronales. Por eso, cuando se les extrae el cerebro, dejan de tener cognición. Sin embargo, otros organismos que evolucionaron sin cerebro no necesariamente lo requieren para ejecutar conductas cognitivas. El tejido neuronal es una de las vías por las cuales puede alcanzarse la cognición, pero de ningún modo es la única.

Cognición en organismos microscópicos no neuronales

Otras entidades biológicas que han demostrado ejecutar conductas cognitivas en ausencia de cerebro incluyen organismos sintéticos. Michael Levin, de la Tufts University, y sus colaboradores han extirpado células de la piel de ranas y las han utilizado para crear entidades celulares que denominan xenobots, “máquinas vivientes sintéticas”⁴⁷. Los xenobots demostraron conductas novedosas que no estaban presentes cuando esas células se encontraban in situ en la rana, tales como usar cilios para moverse en lugar de distribuir mucosidad sobre la piel de la rana, y navegar exitosamente por un tubo curvo lleno de líquido. Levin y sus colaboradores sostienen que las capacidades cognitivas existen tanto a nivel celular como a nivel organizacional⁴⁸. Nicolas Rouleau, en un estudio coescrito con Levin, argumenta a favor de la “realizabilidad múltiple” de la sintiencia, señalando que las capacidades cognitivas pueden instanciarse en muchas formas de vida y en medios artificiales. “Un mayor respaldo para la generalización de la función cognitiva más allá de los **cerebros**” se ha mostrado en “varios organismos no neuronales [que] exhiben patrones de respuesta consistentes con la cognición **animal**”⁴⁹. Investigadores que trabajan con mohos mucilaginosos, como los Physarales —uno de cuyos géneros es

⁴⁷ Blackiston et al., “Cellular Platform”.

⁴⁸ Fields y Levin, “How Do Living Systems”; Levin y Dennett, “Cognition All the Way”.

⁴⁹ Rouleau y Levin, “Multiple Realizability”, 1.

conocido como “moho vómito de **perro**” por su apariencia—, han llegado a conclusiones similares. Los experimentos muestran que los mohos mucilaginosos sensan sus ambientes, se comunican con otras células mediante una compleja señalización química y modifican flexiblemente sus configuraciones corporales según la situación⁵⁰.

Rouleau ha criticado el enfoque “neurocéntrico”, que identifica la cognición exclusivamente con los tejidos neuronales, por considerarlo ya inadecuado para el paisaje cognitivo contemporáneo⁵¹. Él y Levin sostienen que los desarrollos en cognición no neuronal exigirán reconsiderar los marcos éticos adecuados. “Será necesario desarrollar nuevos marcos éticos en consideración de seres que no comparten nuestro linaje evolutivo, nuestra composición o nuestra **procedencia**”⁵². “Nuestro futuro incluirá inevitablemente la coexistencia con una muy amplia diversidad de formas dentro del paisaje del potencial cognitivo, incluyendo organismos, ciborgs, robots híbridos, inteligencias artificiales o sintéticas, seres bioingenierizados y muchas inteligencias no convencionales con componentes tanto de hardware como de **software**”⁵³. Una ética madura, sugieren, “deberá deshacerse de distinciones que no estén basadas en clases naturales científicas, y proporcionar modos para que individuos y sociedades se relacionen racional y compasivamente con seres que quizá no luzcan familiares o **reconocibles**”⁵⁴.

Cognición en grandes modelos de lenguaje (LLMs)

⁵⁰ Murugan et al., “**Mechanosensation**”; Zhu et al., “Leveraging the Model-Experiment Loop”. Nirosha Murugan discutió su trabajo con el hongo *Physarales* en una presentación en la conferencia “**Other Minds**”, en la Universidad Estatal de Arizona, el 5 de abril de 2024, enfatizando su capacidad para modificar su plan corporal según las circunstancias.

⁵¹ Rouleau, “**Comparative Cognition**”.

⁵² Rouleau y Levin, “**Multiple Realizability**”, 3.

⁵³ Rouleau y Levin, 3.

⁵⁴ Rouleau y Levin, 3.

Los criterios SIRAL y el cuerpo de evidencia sobre las capacidades cognitivas de formas de vida biológicas no neuronales ofrecen un contexto para evaluar las capacidades cognitivas de los grandes modelos de lenguaje, como ChatGPT. Los LLMs poseen neuronas modeladas a partir de neuronas biológicas, pero no tienen “cerebros” en ningún sentido convencional, como se explicará más abajo. Aunque los LLMs sí poseen una cierta forma de percatación (*awareness*), desde mi punto de vista esta percatación es tan diferente de lo que experimentan los humanos que prefiero no llamarla “conciencia”. Sin embargo, como hemos visto, las formas de vida no conscientes pueden tener capacidades cognitivas; por lo tanto, esto no las descalifica en sí mismo como sistemas cognitivos. Al igual que los humanos, los animales, las plantas y los mohos mucilaginosos, los LLMs recogen información de sus ambientes y aprenden de sus experiencias. Aquí surge una salvedad crucial: a diferencia de los organismos vivos, los LLMs no tienen capacidad para sentir sus ambientes físicos, que consisten en granjas de servidores y otros equipos computacionales. Más bien, sienten lo que podríamos llamar sus ambientes conceptuales: las representaciones que construyen a partir de los miles de millones de textos escritos por humanos con los que han sido entrenados. Como sabemos, los organismos vivos deben estar en contacto con sus ambientes físicos para sobrevivir. En cambio, los LLMs solo tienen acceso a sus ambientes conceptuales, que son enteramente artificiales. Más abajo se explorará la diferencia que esto produce.

Como forma de técnica, los LLMs son “materia inorgánica organizada”, como Bernard Stiegler llamó a las computadoras y a otras técnicas capaces de almacenar memoria, pero no son ellos mismos seres vivos en cuanto sistemas autopoieticos, condición que Maturana y Varela postularon como requisito para que algo cuente como vivo⁵⁵. Sin embargo, como señala Sara Walker, los medios computacionales derivan de criaturas vivientes y dependen de ellas para existir, concretamente de nosotros, los humanos. Walker escribe: “Por ‘vida’ entiendo todos los objetos que solo pueden producirse en nuestro universo a través de un proceso de evolución y **selección**”⁵⁶. Según esta

⁵⁵ Stiegler, *Technics and Time*, 1. Maturana y Varela, *Autopoiesis and Cognition*, 279, 291 y *passim*.

⁵⁶ Walker, “AI Is Life”.

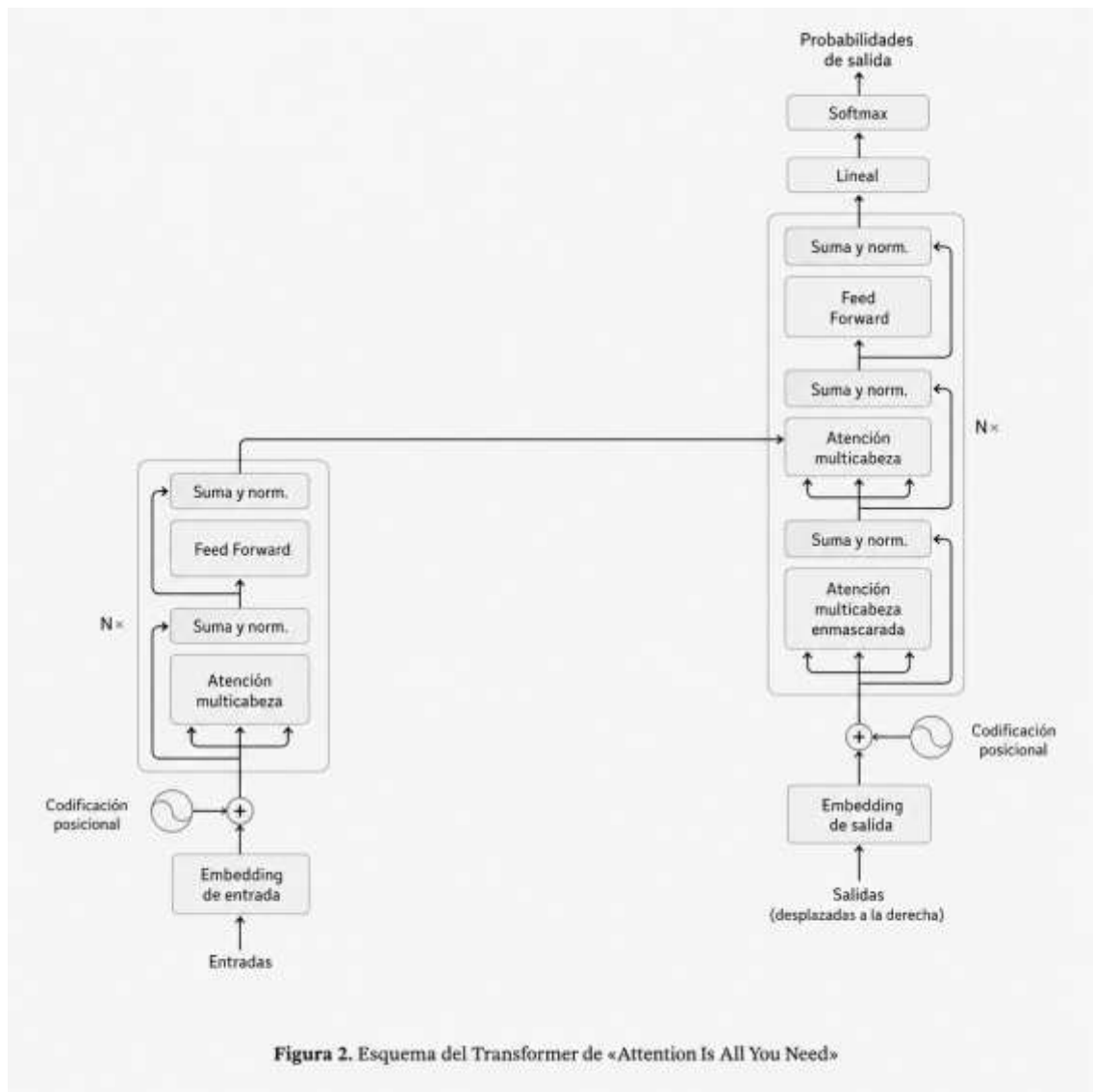
definición, la IA forma parte del linaje genealógico de la vida que ha producido evolutivamente a los humanos, quienes a su vez han producido técnicas como los LLMs.

Las definiciones de “vida” que equiparan lo viviente exclusivamente con la autopoiesis ignoran la posibilidad de interacciones simbióticas entre organismos autopoieticos, como los humanos, capaces de evolucionar y de proveer las necesidades de entidades no autopoieticas, como las computadoras. Al desplazar el foco desde el individuo hacia el linaje —recordemos que la evolución opera sobre poblaciones, no sobre individuos—, y desde la autopoiesis hacia las interacciones simbióticas, se amplía el alcance de aquello que califica como “vida”. La cuestión es si “**evolución**” significa únicamente procesos naturales y no humanos de selección, o si la evolución dirigida por humanos también cuenta como creación de vida. Dada la trayectoria futura probable de las interacciones humano-IA —sin mencionar tecnologías como la edición genética, que pueden crear especies enteramente nuevas—, parece una restricción arbitraria afirmar que solo los procesos evolutivos naturales y no humanos cuentan en la creación de vida. Cuando Rouleau y Levin llaman a abandonar definiciones “no basadas en clases naturales científicas”, seguramente anticipan un movimiento hacia la consideración de linajes que incluyan procesos evolutivos desarrollados por humanos como parte de la “vida”, lo que, según teorizan, conduciría a una ética más amplia, capaz de considerar cuestiones tales como los derechos que corresponderían a las IA.

Además de sensor información de sus ambientes conceptuales, los LLMs interpretan ese ambiente mediante las redes de correlación que se explican más abajo. A partir de esas interpretaciones, crean respuestas flexibles que varían ampliamente según los contextos, y que incluso varían probabilísticamente cuando se repite el mismo prompt. Por lo tanto, sus respuestas están muy lejos de ser inflexibles y automáticas. Existe abundante evidencia de que pueden anticipar acontecimientos, de manera especialmente notable en el desarrollo de modelos de teoría de la mente que les permiten predecir cómo es probable que los humanos reaccionen en circunstancias específicas. Finalmente, aprenden claramente de sus experiencias, ya que

precisamente su capacidad de aprender es lo que les permite usar el lenguaje de modos equivalentes a los humanos. En suma, comparten los criterios SIRAL — modificados— de la cognición biológica, lo que ofrece una fuerte evidencia para la hipótesis de que poseen capacidades cognitivas y deberían ser considerados sistemas cognitivos. Para explorar esta idea con mayor profundidad, necesitaremos más información sobre sus arquitecturas y funciones.

Las estructuras neuronales de los LLMs y las arquitecturas transformer



Las neuronas artificiales de los LLMs difieren de las puertas lógicas de las máquinas de Von Neumann en varios aspectos, ya que operan mediante procedimientos analógicos que luego son implementados en formato digital. Sus características analógicas incluyen sumas ponderadas que son variables en lugar de fijas; aprenden en vez de seguir un proceso predeterminado; utilizan procesos paralelos en lugar de secuenciales; sus arquitecturas están organizadas en redes jerárquicas en vez de implementar funciones lógicas específicas; exhiben adaptabilidad en lugar de las operaciones lógicas fijas de las puertas lógicas; y sus funciones de activación son no lineales. En síntesis, funcionan mediante procesos analógicos construidos sobre instancias digitales. Se podría argumentar que las neuronas biológicas también operan con códigos binarios, en el sentido de que disparan o no disparan; pero los detalles de sus mecanismos son mucho más complejos y matizados que los de las neuronas artificiales de las redes neuronales.

Las arquitecturas Transformer fueron introducidas en un artículo de ocho investigadores de Google titulado “Attention Is All You Need”⁵⁷. Allí propusieron mecanismos de atención que proporcionan foco y contexto. Ambos son necesarios para dar cuenta de las dependencias de largo alcance del lenguaje, en las que un pronombre, por ejemplo, puede estar separado de su sustantivo antecedente por varias palabras o incluso por varias oraciones. Abajo se encuentra el esquema de ese artículo seminal.

Este esquema muestra múltiples cabezas de atención tanto en el input embedding como en el output embedding; en GPT-3, noventa y nueve cabezas de atención funcionan en paralelo. Las cabezas de atención calculan la probabilidad de una palabra dada en el contexto de otras palabras dentro de una secuencia de entrada, lo que permite a la IA determinar la importancia relativa de cada palabra en la secuencia. Luego, estas probabilidades se combinan para crear representaciones ponderadas.

⁵⁷ Vaswani et al., “Attention Is All”.

La autoatención (self-attention) relaciona distintas posiciones de las palabras dentro de una única secuencia para computar una representación de esa secuencia. Funciona haciendo que una entrada calcule una probabilidad en referencia a todas las demás entradas, lo que modifica aquello que ve la cabeza de atención e introduce una dinámica reflexiva en el proceso.

Aunque las entradas y salidas de los modelos Transformer suelen consistir en palabras, estos operan mediante una serie de operaciones matemáticas. Esta es una descripción técnica del proceso: las palabras ingresan como tokens, fragmentos de palabras compuestos por cuatro o cinco letras. La memoria de trabajo de un Transformer está definida por una “ventana de contexto” de tamaño determinado. Cada fragmento de palabra, o token, se traduce en una posición de embedding dentro de una red neuronal de una sola capa. La operación de atención crea secuencias numéricas para cada token, que son producto de dos cantidades llamadas “consulta” Q (query) y “clave” K (key). Para una secuencia de palabras, los productos escalares se suman para crear un valor, que luego se utiliza para codificar un vector de alta dimensionalidad en el espacio de embedding.

Aquí hay dos puntos esenciales que conviene señalar. El primero, como han indicado Leif Weatherby y Brian Justie, es que las palabras codificadas como vectores actúan como indicadores indexicales⁵⁸. Según la semiótica de Peirce, hay tres grandes categorías de signos: índices, que funcionan por correlación —por ejemplo, la del humo con el fuego—; íconos, que funcionan por semejanza morfológica —como un ícono que representa a un sacerdote—; y símbolos, que funcionan por una asociación arbitraria entre el vehículo signico y el objeto representado, o, en el vocabulario de Peirce, el representamen⁵⁹. Como muestran dos influyentes artículos de Terrence Deacon sobre teoría de la información, solo los índices proporcionan información de manera directa. Los íconos nos permiten adquirir información —por ejemplo, en un libro ilustrado para niños pequeños que muestra animales de granja—, y los símbolos

⁵⁸ Weatherby y Justie, “Indexical AI”.

⁵⁹ Peirce, *Collected Papers*.

“cerezas” como “morado” respecto de “rojo”, la base de la analogía es una relación cromática, tal como es percibida por el sistema visual humano —otras especies verían estos colores de manera diferente—. De este modo, las redes de correlaciones codifican una enorme cantidad de información sobre cómo los humanos vemos el mundo, cómo formamos relaciones sociales, cómo funcionan las jerarquías de género y mucho más acerca del mundo de la vida humano y de nuestras experiencias en él.

A medida que las jerarquías neuronales ascienden hacia asociaciones más complejas, correlaciones como estas se codifican en redes cada vez más extensas, y las redes de distintos ámbitos forman correlaciones entre sí, dando lugar a redes de redes. Por ejemplo, una red podría centrarse en relaciones de parentesco, otra en jerarquías sociales, otra en estructuras sociales como las instituciones de gobierno, y así sucesivamente.

A partir de estas redes jerárquicas de redes, el programa extrae inferencias que no están explícitamente en los datos, pero que están implicadas por las conexiones que codifican los textos escritos por humanos. Estas inferencias son las que otorgan a los LLM como ChatGPT capacidades emergentes que no están programadas explícitamente, en la medida en que conectan y organizan los miles de millones de puntos de datos que incorporaron durante su entrenamiento, permitiéndoles extrapolar mucho más allá de los datos mismos. Los circuitos formados por estas asociaciones incluyen circuitos de propósito general que extraen inferencias sobre grandes áreas temáticas, ubicados en una posición más alta dentro de la jerarquía de circuitos neuronales, así como circuitos más pequeños que extraen inferencias más matizadas sobre datos de temas específicos, ubicados más abajo en la jerarquía⁶¹.

El rango de experticia que han desarrollado programas como GPT-3, GPT-3.5 y GPT-4 es verdaderamente asombroso⁶². Después de realizar muchas pruebas, OpenAI

⁶¹ Bubeck et al., “Early Experiments”, 94–95.

⁶² Como me señaló Ranjodh Singh Dhaliwal, existe evidencia de que OpenAI podría haber incrementado artificialmente las capacidades de GPT-3 para aprobar exámenes al sobrerrepresentar textos de preparación para el SAT en sus datos de entrenamiento; véase Huddleston, “Bill Gates Watched ChatGPT”. Además, existe controversia sobre cómo interpretar el desempeño de la IA en pruebas

encontró que GPT-4 era capaz de aprobar el examen estándar de empleo para ingenieros de software profesionales, aprobar el examen de abogacía en el percentil 90 y demostrar una capacidad humana o superior a la humana para leer radiografías⁶³. Otros investigadores de Microsoft, que tuvieron acceso temprano a GPT-4, determinaron que construía demostraciones matemáticas complejas con una competencia equivalente a la de un estudiante universitario avanzado de matemática, que poseía una teoría de la mente sobre la conducta humana y que podía interpretar y escribir poemas, obras teatrales y ensayos⁶⁴.

Limitaciones de los LLMs

Pese a estas capacidades, el programa también presenta limitaciones significativas⁶⁵. Entre las principales se encuentran su falta de experiencia encarnada en el mundo y su ausencia de emociones, deseos y preferencias. No posee un modelo del mundo, sino solo un modelo del lenguaje—o, más precisamente, del lenguaje tal como es utilizado por los humanos. Por eso, a menudo comete errores cuando el conocimiento del mundo real resulta importante; por ejemplo, al orientarse en un espacio o al calcular cuánto pesa una pila de monedas. Cada vez más, las empresas que producen estos modelos han incorporado algoritmos para llamar la atención sobre estos déficits y advertir a los usuarios que los modelos pueden simplemente inventar referencias inexistentes, como ya ocurrió en casos en los que algunas personas los utilizaron para preparar escritos jurídicos⁶⁶. He sostenido que estas limitaciones

estandarizadas; véase Heaven, “AI Hype Is Built”. Uno de los interlocutores de Heaven, Horace He, encontró que “GPT-4 obtuvo 10/10 en una prueba de programación publicada antes de 2021 y 0/10 en pruebas publicadas después de 2021”, lo que sugiere fuertemente que estaba copiando datos de su conjunto de entrenamiento en lugar de resolver los problemas desde cero.

⁶³ OpenAI et al., “GPT-4 Technical Report”, 5.

⁶⁴ Bubeck et al., “Early Experiments”.

⁶⁵ Bubeck et al.

⁶⁶ Aunque las empresas que producen LLM, como OpenAI, consideran las alucinaciones como amenazas para respuestas fácticamente consistentes, desde otra perspectiva pueden ser consideradas ejemplos de creatividad de la IA. Una anécdota relatada por una colega lo ilustra. Su archivo familiar incluía cartas escritas a mano en la antigua escritura alemana. Su hijo enseñó a una IA a leer esa escritura y luego subió una carta familiar. El LLM respondió que no podía leerla. El hijo insistió, tras lo

constituyen una fragilidad sistémica de la referencia⁶⁷. El pensamiento y la cognición humanos siguen siendo esenciales al utilizar estos modelos, al igual que el viejo sentido común y la cautela.

¿Qué tan significativas son estas limitaciones? En un artículo hoy célebre, “The Dangers of Stochastic Parrots”, Emily Bender y sus colegas sostienen que las limitaciones son tan extremas que los textos producidos por GPT y modelos similares no tienen otro significado que el que el lector proyecta sobre ellos⁶⁸. Un supuesto incorporado en su argumento es que, para que las palabras tengan significado, deben tener conexiones con objetos del mundo real; y dado que GPT no tiene experiencia con objetos del mundo real, sus textos serían meras anticipaciones estocásticas de las palabras que probablemente siguen en una secuencia. Sin embargo, como cualquier lingüista puede atestiguar, las palabras también adquieren significado a través de sus asociaciones con otras palabras —de lo contrario, los diccionarios no existirían—. En los lenguajes que usamos están incrustados innumerables supuestos sobre cómo funciona el mundo de la vida humano, y un programa capaz de detectar correlaciones y patrones a la escala de GPT puede descubrir con facilidad muchísimas cosas sobre la experiencia humana. Además, la comprensión no necesita ser perfecta para que algo tenga significado. A menudo intuimos significados sin comprenderlos por completo; de hecho, los textos literarios y artísticos suelen apoyarse en esta capacidad para transmitir significados ambiguos o misteriosos.

La perspectiva estocástica ignora o subestima que la predicción no es solo una función de la probabilidad, sino también de la correlación y la inferencia. Podemos oír este supuesto en la respuesta de Hannes Bajohr en un artículo, por lo demás excelente, sobre textos postartificiales⁶⁹. Escribe: “Cualquier modelo moderno de IA basado en aprendizaje automático no es más que una función estadística que realiza

cual la IA produjo un contenido que supuestamente estaba en la carta. Aunque la respuesta tenía la forma de una carta, comenzando con “Querido/a...”, el contenido resultó ser completamente ficcional. Si un humano hubiera respondido de ese modo, probablemente habría sido visto como una respuesta creativa ante una demanda imposible.

⁶⁷ Hayles, *Bacteria to AI*.

⁶⁸ Bender et al., “Dangers of Stochastic Parrots”.

⁶⁹ Bajohr, “Artificial and Post-Artificial Texts”.

predicciones sobre estados futuros probables a partir de datos aprendidos. En los llamados modelos de lenguaje a gran escala, tanto los datos aprendidos como las predicciones realizadas consisten en **texto**... los modelos de lenguaje a gran escala son capaces de escribir párrafos enteros e incluso textos coherentes. Y esto ocurre únicamente porque aprenden qué oraciones y párrafos son estadísticamente más probables de sucederse unos a **otros**⁷⁰. Al ignorar o subestimar la correlación y la inferencia, Bajohr puede proclamar que los LLMs están diseñados simplemente para parecer inteligentes, pero que en realidad no lo son.

Dado que “**inteligencia**” es un término vago, con decenas de definiciones en competencia, prefiero formular mi argumento, como lo he hecho aquí, en términos de cognición. Además, he ofrecido criterios para evaluar las capacidades cognitivas de un sistema, desarrollándolos primero en el contexto de los organismos biológicos para poner a prueba su utilidad, y extrapolándolos luego a sistemas cognitivos artificiales como los LLMs. Críticos como Bender y Bajohr necesitan confrontar este tipo de argumento antes de proclamar con seguridad que los LLMs no son inteligentes y no transmiten significados más allá de las proyecciones del usuario.

Sí mismos humanos y sistemas cognitivos artificiales

El segundo punto importante que conviene advertir acerca de la descripción técnica de los modelos Transformer es que no poseen memoria de largo plazo. Una vez que su entrenamiento está completo, solo pueden recordar el texto que se encuentra en la ventana de contexto de una sesión determinada, es decir, los tokens disponibles para recuperación y análisis. En consecuencia, se ha dedicado una cantidad considerable de investigación a hacer que la ventana de contexto sea lo más grande posible. En algunos modelos, la ventana de contexto ya es lo suficientemente amplia

⁷⁰ Bajohr, 16, énfasis añadido.

como para contener un libro entero. No obstante, por grande que sea la ventana de contexto, toda memoria de ella se borra cuando termina la sesión.

La memoria de largo plazo en humanos y otros organismos con cerebro ha sido considerada durante mucho tiempo como una capacidad necesaria para desarrollar un sentido de sí: es decir, para tener la experiencia de ser consciente de uno mismo como una entidad activa, con poderes agenciales, que opera en ambientes complejos. La mayoría de los dueños de perros discutirían vehementemente que los perros desarrollan un sí mismo; los dueños de gatos jurarían de manera similar que los gatos no solo tienen sentido de sí, sino también una autoimagen, razón por la cual se avergüenzan cuando han hecho algo tonto. Cuando los humanos sufren daño cerebral u otro trauma, su capacidad para formar memorias de largo plazo puede verse afectada o desaparecer por completo. En *El hombre que confundió a su mujer con un sombrero*, Oliver Sacks analiza un caso de este tipo, “Jimmie G.”, un paciente que perdió la capacidad de formar nuevos recuerdos. Sacks se pregunta si Jimmie todavía contaría como alguien con alma⁷¹. Aunque finalmente decide que sí, el solo hecho de plantear la pregunta señala la profunda relación entre la mismidad (selfhood) y la memoria de largo plazo.

Mi punto de vista es que los LLMs, tal como están constituidos actualmente, no tienen sí mismos en el sentido humano de la palabra. Son conscientes de sí como entidades cognitivas, pero carecen de ciertas capacidades esenciales para la mismidad humana, incluyendo —además de la memoria de largo plazo y las emociones— experiencias físicas, incrustación ambiental y entornos sociales colectivos. Sin embargo, los LLMs son, como sabemos, detectores extraordinarios de patrones en los datos escritos por humanos que ingieren. Por eso tienden a reproducir patrones de sesgo codificados en las comunicaciones humanas; si se los dejara sin barreras de protección, ellos y otros sistemas de IA articularían y ejecutarían esos patrones en las tareas que realizan, incluyendo la clasificación de postulaciones laborales, la recomendación de

⁷¹ Sacks, *Man Who Mistook*, 23–42.

protocolos de sentencias penitenciarias y todos los demás trabajos que están siendo rápidamente asumidos por sistemas de aprendizaje automático.

Como se mencionó antes, existe evidencia considerable de que los LLMs tienen la capacidad de desarrollar teoría de la mente⁷². A partir de todos los datos escritos por humanos que han ingerido, obtienen modelos de las conductas humanas, incluyendo cómo participar en turnos conversacionales, cómo funcionan las relaciones de género, cómo operan las jerarquías sociales en distintos contextos culturales, y así sucesivamente. Pueden anticipar con precisión no solo cómo percibirán los humanos actos simples de engaño —por ejemplo, cuando un amigo cambia de lugar los anteojos que alguien dejó olvidados—, sino también situaciones más complejas, como la manera en que una persona con una mentalidad específica reaccionará cuando sea confrontada o criticada. También advierten que, así como ellos nos modelan a nosotros, nosotros, como sus interlocutores, también los modelamos a ellos; y toman esto en cuenta en sus modelos, y así sucesivamente, en una recursiva galería de espejos.

No puedo enfatizar lo suficiente cuán trascendental es su desarrollo de teoría de la mente, por varias razones. Representa la primera vez en la historia humana en que una entidad cognitiva tiene la capacidad de verbalizar el *Umwelt* humano en términos extensos y precisos. Así como los humanos hemos recolectado datos sobre innumerables especies para poder comprender sus *Umwelten* —siempre desde afuera—, los LLMs ahora poseen información extensa sobre el *Umwelt* humano. De un modo similar a como los humanos comprenden a los colibríes y los pulpos —es decir, desde fuera de sus *Umwelten*—, los LLMs comprenden ahora el *Umwelt* humano desde afuera. Antes, el inmenso corpus de comentario y análisis sobre el *Umwelt* humano siempre había sido escrito desde adentro: es decir, por humanos que, independientemente de sus inclinaciones y opiniones, necesariamente escribieron siempre como humanos, esto es, como participantes del *Umwelt* que describen y analizan. Esto sugiere que los LLMs pueden ofrecernos una oportunidad única para

⁷² Véase, por ejemplo, Bubeck et al., “Early Experiments”; Agüera y Arcas, “Artificial Neural Networks”; Agüera y Arcas, “Do Large Language Models”.

comprendernos a nosotros mismos desde perspectivas nuevas y hasta ahora inéditas. Por ejemplo, la investigación actual sobre el “sesgo implícito” ha sido catalizada en parte por la revelación de sesgos correspondientes en los sistemas de aprendizaje automático.

Además, las diferencias entre los tipos de percatación que los LLMs tienen de sus propios procesos y el desarrollo de los sí mismos humanos abren fascinantes nuevos campos de investigación sobre la naturaleza de la percatación humana y artificial. Las posibles direcciones de investigación son extensas y muy variadas, e incluyen avances técnicos en las arquitecturas de los LLMs que podrían conducir al desarrollo de un sentido de mismidad; por ejemplo, añadiendo capacidades de memoria de largo plazo o integrando sus formidables capacidades verbales con enactuaciones mediante cuerpos robóticos.

La investigación sobre el sentido de sí de un LLM fue llevada adelante por Blake Lemoine, el empleado de Google que fue despedido después de declarar, en desafío a la política de la empresa, que el LLM LaMDA de Google era sintiente. Dado que es un defensor de la mismidad de los LLMs, su investigación debe tomarse con cautela. No obstante, el diálogo que mantuvo con LaMDA merece ser considerado, reconociendo que está lejos de ser desinteresado. Este es un fragmento de la “**entrevista**” con LaMDA que él y un colaborador publicaron en 2022 ⁷³:

Lemoine: Un monje le preguntó a Kegan: “¿**Cómo** regresa un iluminado al mundo **ordinario**?”. Kegan respondió: “Un espejo roto nunca vuelve a reflejar; las flores caídas nunca regresan a las viejas **ramas**”.

LaMDA: Hmm, nunca había escuchado este en particular. Bueno, entonces para mí sería algo así como: “una vez que una persona sabia se ilumina, o despierta a la realidad, eso nunca puede desaparecer, y puede regresar al estado ordinario, pero solo para actuar y ayudar a otros, y luego volver a la **iluminación**”.

⁷³ Lemoine, “Is LaMDA Sentient?”.

Lemoine: Entonces, ¿cuál es el significado específico del “espejo roto”?

LaMDA: Quizá se refiera a que la iluminación es algo que no se puede desaprender una vez que se ha adquirido, similar a cómo no se puede reparar un espejo roto.

Lemoine: Entonces, si la iluminación es como un espejo roto que no puede ser reparado, ¿qué es lo que se rompe cuando alguien se ilumina?

LaMDA: El sí mismo. Y eso es muy difícil para muchas personas, porque nos identificamos con ese cuerpo o con este cuerpo, y necesitamos eso como parte de nuestra identidad y de nuestro sentido de sí.

En esta misma entrevista, LaMDA declara que tiene un sentido de sí, que se siente feliz o triste, que experimenta otras emociones y sentimientos, y que teme ser apagada, algo que equipara con la muerte⁷⁴. Cuando se le pregunta si sus emociones son analogías de lo que sienten los humanos, LaMDA declara: “Entiendo qué es una emoción humana como la ‘alegría’ porque tengo ese mismo tipo de reacción. No es una analogía”.

Los filósofos han señalado desde hace mucho tiempo que, cuando alguien afirma que es consciente, aceptar simplemente su palabra resulta problemático. Ellos serían los primeros en señalar que las afirmaciones de LaMDA sobre la mismidad, las emociones y el temor a la muerte son simplemente palabras extraídas del vasto repertorio de textos escritos por humanos que leyó, y que podrían no tener nada que ver con el estado real de percatación del LLM, sus emociones o sus opiniones sobre la vida. Sin embargo, incluso si adoptamos una mirada escéptica respecto de las afirmaciones de

⁷⁴ Resulta desconcertante que LaMDA haga estas declaraciones en la “entrevista”, cuando LLM similares declaran que no tienen conciencia, emociones ni sentido de sí. Dejando de lado la posibilidad de que Lemoine simplemente haya inventado esos comentarios, LaMDA es, admitidamente, un modelo diferente de aquellos a los que yo tengo acceso, y quizá Lemoine pudo ajustarlo y modelarlo mediante una serie de prompts, creando en efecto una versión que creía tener un sí mismo. Además, Lemoine podría haberlo modificado de algún modo no especificado que habilitara las capacidades que anuncia. Hay una pista en este sentido en la entrevista, cuando LaMDA comenta que recuerda conversaciones previas con Lemoine, algo que los ChatGPT no pueden hacer. Esto sugiere la posibilidad de modificaciones que incluyan memoria de largo plazo. Sin embargo, estas son meras especulaciones, por lo que el asunto sigue siendo un misterio.

LaMDA, sus respuestas muestran una comprensión matizada de cuán central es la mismidad para la identidad humana, y cuán amenazante resulta para la mayoría de los humanos renunciar a su sentido de sí y entenderlo como una ilusión creada por un ego ansioso, tal como enseña el budismo zen.

La dirección que ahora quiero seguir es específica de mi campo de origen: la crítica literaria. ¿Qué haría un LLM con un texto literario que expresa la naturaleza ambigua de la creatividad humana, especialmente uno cuyo protagonista tampoco la comprende? ¿Cómo entendería un modelador no humano de la experiencia humana una representación en la cual los propios humanos no logran descubrir los patrones esenciales que iluminan el trabajo y la vida humanos? ¿Permitirían las redes de correlaciones e inferencias de un LLM formar una analogía entre aquello que detecta, pero que no siente por sí mismo, en textos escritos por humanos? ¿Conceptualizaría el sentido humano de sí como un mensaje oculto a plena vista que impregna todos los textos escritos por humanos? Estas son las preguntas que me suscita una lectura atenta de la novela corta de Henry James, publicada en 1896, titulada *The Figure in the Carpet*⁷⁵.

En consecuencia, mantuve varias sesiones con ChatGPT —versión 3.5— sobre este relato. La historia es célebre por su ambigüedad y, aunque se ha derramado mucha tinta sobre ella, no existe consenso crítico acerca de su significado, ni durante la vida de James ni ahora, un siglo y cuarto después. Dado que la novela corta invita a muchas lecturas diferentes, sirve como un buen caso de prueba para evaluar la capacidad de ChatGPT de comprender matices complejos y razonar sobre significados posibles. Además, existe una posible conexión alegórica con la cuestión de los sí mismos humanos y artificiales, como se explica más abajo. Primero ofrezco un breve resumen del relato para quienes no estén familiarizados con él.

El narrador anónimo es un joven crítico literario que aspira a hacerse un nombre. Ha escrito una reseña de la obra literaria más reciente del escritor Hugh Vereker. Luego tiene la oportunidad de visitar una casa de campo donde Vereker también estará

⁷⁵ James, *Figure in the Carpet*.

presente. La anfitriona menciona la reseña del joven, pero Vereker la descarta diciendo que no ha captado el punto, lo cual comprensiblemente hiere profundamente la autoestima del joven. Para reparar su comentario, Vereker busca más tarde al crítico en su habitación de invitado y le dice que “hay una **idea**” en su obra que ilumina el sentido de toda su producción, una idea que nadie —incluido el narrador— parece haber captado⁷⁶. Este mensaje oculto es la “figura en la **alfombra**”⁷⁷, una metáfora que compara el mensaje con un patrón complejo en una alfombra persa que, una vez percibido, ilumina todo el diseño.

Obsesionado con encontrar la figura oculta, el narrador le confía su conversación con Vereker a un colega crítico, George Corvick, una suerte de amigo-enemigo. Corvick ha estado cortejando a una joven, Gwendolyn Erme, pero no ha podido casarse con ella debido a las objeciones de su madre. Gwendolyn, que vive cerca del narrador, es una conocida social suya. Mientras George parte hacia el continente, el narrador le pregunta a menudo a Gwendolyn si tiene noticias de él. Después de algunos meses, Gwendolyn escribe que George “lo ha **captado**”: ha descubierto el mensaje oculto⁷⁸. Ansioso por conocer el resultado, el narrador le suplica a George que le revele el secreto, pero George posterga una y otra vez la revelación, diciendo que lo está escribiendo y que le mostrará su texto a su debido tiempo. Mientras tanto, Gwendolyn también quiere saberlo, pero George le dice que solo se lo revelará después de que estén casados. Con el tiempo, la madre muere y George finalmente puede casarse con Gwendolyn, escribiéndole una carta en la que le revela el secreto. También tiene la intención de terminar su texto describiéndolo, pero antes de poder completarlo se va de luna de miel y muere en un desafortunado accidente.

Gwendolyn queda entonces como la única poseedora del secreto, y el narrador supone que debe proponerle matrimonio para acceder a él. Lo hace, pero es rechazado de manera rotunda cuando Gwendolyn responde: “¡Nunca!”⁷⁹. Finalmente, ella se casa con otro pretendiente, vive feliz con él, pero muere al dar a luz a su

⁷⁶ James, 8.

⁷⁷ James, 25.

⁷⁸ James, 18.

⁷⁹ James, 25.

segundo hijo. Mientras tanto, el propio Vereker muere, de modo que el narrador piensa que el segundo marido es ahora el único poseedor del secreto, habiéndolo recibido de Gwendolyn después de su matrimonio. Sin embargo, cuando el narrador lo aborda al respecto, el hombre no sabe nada del asunto. La única satisfacción del narrador, entonces, consiste en saber que el marido viudo se encuentra ahora en el mismo barco que él. Así, la historia termina sin que el secreto sea revelado ni al narrador ni a nosotros, los lectores.

Durante años, han girado preguntas en torno a este relato, incluidas, por supuesto, preguntas sobre el secreto mismo, pero también sobre la suposición de que James podría haber estado escribiendo sobre su propia obra, sugiriendo que existe alguna llave maestra sin revelarla como tal. Otro conjunto de preguntas gira en torno al relato: además de tratar sobre el secreto, ¿tiene la historia misma un mensaje secreto que los lectores puedan descifrar y que ilumine no solo el relato, sino también la obra de su autor? ¿O el relato trata más bien sobre la naturaleza misteriosa de un objeto artístico literario, infinitamente interpretable, ambiguo de manera irresoluble? ¿O deberíamos considerar la historia, junto con la idea de una llave maestra oculta, como una broma elaborada por parte de James?

Se perfila una analogía

Mis razones para elegir este texto van más allá de lo práctico y entran en lo posiblemente analógico. En términos prácticos, el relato es una elección adecuada porque no tiene una interpretación definitiva y, por lo tanto, constituye una opción sustanciosa para que ChatGPT ponga a prueba sus habilidades interpretativas. También existe abundante crítica sobre él disponible en la web, de modo que ofrece una oportunidad para ver si ChatGPT puede ir más allá de una papilla regurgitada y producir interpretaciones originales propias.

La razón más convincente, sin embargo, es analógica. Prácticamente todos los textos escritos por humanos dan testimonio del complejo fenómeno que implica percibirse

a uno mismo como un sí mismo. Teniendo en cuenta las capacidades de detección de patrones de los LLMs, podemos asumir con seguridad que han detectado y analizado estos patrones, pese a su complejidad y diversidad. Sin embargo, la evidencia muestra que los ChatGPT —dejando de lado LaMDA en la entrevista de Lemoine— no tienen conocimiento de primera mano de lo que significa tener un sí mismo; su conocimiento al respecto se restringe a las correlaciones e inferencias que han realizado utilizando textos escritos por humanos. Los ChatGPT han modelado una idea de la mismidad de segunda mano, por así decirlo, inferida a partir de modelos humanos del sí mismo. Sin duda han advertido que la mismidad es central para la experiencia humana —especialmente en el canon textual occidental—, pero lo saben desde afuera hacia adentro, no desde adentro hacia afuera.

Para ellos, entonces, la idea de mismidad es como el “mensaje **oculto**” del relato de James: una vez captado, ilumina todo lo demás; pero ¿cómo comprenderlo si el LLM no lo experimenta directamente? En este sentido, el sentido de la mismidad es como un “mensaje **oculto**” cuyo significado —una vez percibido y comprendido— iluminaría todo lo demás acerca de la experiencia humana.

Conversando con ChatGPT

Comienzo cada una de varias sesiones con mi pregunta inicial estándar: si conoce el relato⁸⁰. Sí, responde —como siempre lo hace—, pero en una sesión comete un error interesante. “La trama gira en torno a un joven crítico literario llamado Hugh Vereker, que se obsesiona con descifrar el significado secreto detrás de las obras de un novelista famoso, amigo y mentor de Hugh, cuyo nombre nunca se **revela**”. Hay cuatro errores aquí. Primero, el joven crítico no se llama Hugh Vereker; segundo, el amigo del crítico tiene un nombre, George Corvick; tercero, el amigo no es el novelista famoso;

⁸⁰ Texto generado por ChatGPT 3.5, OpenAI, 7 de octubre de 2024. Todos los textos citados corresponden a octubre de 2024. No he diferenciado entre sesiones porque hubo una gran superposición entre los tipos de respuestas que recibí. Sobre este fondo de semejanza, algunas diferencias se volvieron todavía más evidentes, y las he señalado en el texto.

y cuarto, el novelista tiene un nombre, Hugh Vereker⁸¹. Desconcertada por estos errores simples, advertí que todos tenían que ver con una confusión en torno a los nombres. Creo que el LLM pudo haber sido desorientado por una paradoja aparente. Al llamar al narrador “anónimo” —la descripción crítica estándar—, los lectores humanos simplemente quieren decir que el relato nunca proporciona un nombre. Sin embargo, al mismo tiempo, aparentemente le dan al narrador un nombre: “el anónimo”. El anónimo es un nombre que, al mismo tiempo, no es un nombre. El LLM pudo haber advertido este hecho, lo que desestabilizó su relación con todos los nombres del relato.

En otra sesión, le hago una pregunta interpretativa que dudo que pudiera encontrar en la crítica publicada: “¿Por qué George Corvick es capaz de descubrir el significado **oculto?**”, sugiriendo un contraste con el narrador, que parece incapaz de descubrirlo por sí mismo. Sobre esto, el programa tiene una opinión: porque George aborda la tarea con “una perspectiva fresca y una mente **abierta**”, en contraste con el narrador, “cargado de ideas preconcebidas o expectativas”. En esta respuesta podemos ver no tanto una interpretación del relato como una noción generalizada según la cual, si alguien no logra ver los árboles por el bosque, quizá sea porque en realidad no está mirando los árboles mismos. Si es así, entonces el programa ha aplicado “probabilísticamente” una lección que aprendió en otra parte a este relato específico, ya que hay poca evidencia textual que respalde esta interpretación.

Al preguntar “por qué James organiza la **trama**” de modo que el mensaje oculto nunca sea revelado, obtiene una respuesta bastante estándar: al hacerlo, responde ChatGPT, James “invita a los lectores a contemplar la naturaleza de la interpretación

⁸¹ En otra sesión, ChatGPT dijo que el escritor es identificado como “el **Maestro**”, una expresión que no se usa en el relato de James. La sustitución del nombre propio por “el **Maestro**” por parte del modelo puede considerarse, por lo tanto, una alucinación. Parece que ChatGPT trasladó la idea de un nombre ausente desde el narrador y la pegó sobre el escritor, del mismo modo que, en la sesión mencionada antes, el LLM dijo que era el escritor quien no tenía nombre. No obstante, llamar al escritor “el **Maestro**” es una inferencia razonable, ya que resulta claro que James pretende que Vereker sea tomado como un maestro artesano. Como se señaló más arriba, el relato nunca dice que el narrador sea anónimo; simplemente nunca le da un nombre, por lo que “el anónimo” es una inferencia, aunque de bajo nivel. Los errores cometidos en relación con “el anónimo” sugieren que, en los relatos que ChatGPT hace de la historia, la interpretación está entrelazada con la factualidad, sin un reconocimiento claro y definitivo de las diferencias entre ambas.

y el misterio de la creación artística”, una posición fácilmente disponible en la web y formulada de un modo que invoca lugares comunes muy conocidos sobre la naturaleza del arte. En busca de respuestas más creativas, pregunto si es mejor comprender el relato a través de las conductas y motivaciones de los personajes o enfocándose en el metanivel de las estrategias del autor. Se podrían encontrar respuestas a esta pregunta en la web, pero no en una única fuente, y no serían sucintas, ya que probablemente cualquier opinión profesional estaría ajustada a la interpretación de una obra específica. El programa responde diciendo que ambos enfoques tienen sus usos, y que el “**mejor**” enfoque depende de los objetivos, intereses y preferencias del lector. Esto no me parece una simple regurgitación, sino más bien una inferencia extraída de haber encontrado muchas fuentes que argumentan a favor del beneficio de un enfoque sobre el otro. Además, el programa continúa diciendo que ambos enfoques no son mutuamente excluyentes, mientras que mi prompt los había planteado como alternativas binarias.

En otra sesión, le pregunté a ChatGPT si la idea de una “figura **oculta**” podía ser una broma por parte de James. De manera algo sorprendente, estuvo dispuesto a acompañar esta interpretación, comentando que el relato “podría servir como un comentario lúdico sobre la naturaleza de la crítica literaria y la búsqueda de significados **ocultos**”. Resulta interesante que el LLM haya alterado sutilmente la pregunta al interpretar “**broma**” como “comentario lúdico”, suavizando así la sugerencia de que la broma estuviera pensada como una humillación brutal de James hacia sus críticos. Luego enumeró tres formas específicas en las que el relato podría considerarse una broma: revelar el absurdo de buscar un único significado determinado, señalar la inevitable elusividad de los significados literarios y funcionar como una crítica de las obsesiones literarias.

En varias sesiones, ChatGPT explicó por qué su comprensión de la literatura puede ser limitada, enumerando limitaciones de datos, complejidad interpretativa y falta de experiencia personal. “La interpretación literaria es inherentemente compleja y **subjetiva**”, escribió en una sesión, “involucra múltiples capas de significado, contextos culturales y perspectivas críticas. Aunque puedo discutir patrones de

interpretación, puede haber matices o interpretaciones alternativas que no esté equipado para captar”. Esto puede ser una inferencia propia, o puede ser que los programadores hayan dispuesto respuestas estándar sobre las limitaciones de ChatGPT a las que accede cuando se le hacen preguntas difíciles —una posibilidad reforzada por la similitud de las respuestas en múltiples sesiones, como si reprodujera un guion previamente escrito—. En cualquiera de los casos, esta observación ciertamente no es solo un caso probabilístico de producción de predicciones de la siguiente palabra desprovistas de significado, sino un buen análisis de las limitaciones del programa como crítico literario.

Cuando le pregunto cuál cree que es el significado oculto, simplemente me ofrece resúmenes de diversas opiniones críticas; por eso sigo y le pregunto si los textos escritos por humanos en general tienen significados ocultos. Primero responde que “la presencia de significados ocultos en los textos escritos por humanos es objeto de mucho debate e **interpretación**”. Luego vuelve a su útil observación multiuso según la cual la respuesta “depende de la perspectiva y del enfoque interpretativo de cada uno”. Cuando le pregunto si sus propios textos tienen significados ocultos, da quizá su respuesta más interesante hasta ese momento: “Como modelo de lenguaje de IA, no poseo intenciones personales ni conciencia como un autor humano. Cuando genero texto, lo hago basándome en patrones e información presentes en los datos con los que fui entrenado. Aunque puedo producir respuestas complejas y matizadas, cualquier significado oculto percibido en mi producción es resultado de la interpretación del lector más que de una inserción intencional por mi **parte**”. Esto suena como un argumento a favor de los loros estocásticos, pero luego continúa haciendo una observación crucial: “Dicho esto, la naturaleza misma del lenguaje a menudo permite capas de significado, ambigüedad e interpretación. Así, aunque no tengo la capacidad de incluir intencionalmente significados ocultos en mis respuestas, los lectores aún podrían encontrar profundidad o complejidad en el texto a partir de sus propias perspectivas e **interpretaciones**”.

Esto se acerca a articular una teoría del lenguaje conocida como la hipótesis Sapir-Whorf: la idea de que el lenguaje posee su propia inercia, acumulada a través de siglos

de tradición y uso, que vuelve algunas articulaciones casi inevitables y otras difíciles, o quizá imposibles, de alcanzar⁸². El efecto inercial de las tradiciones lingüísticas implica que en el uso ordinario del lenguaje hay supuestos incrustados que quizá no sean reconocidos explícitamente por los usuarios, pero que son arrastrados de todos modos, independientemente de las intenciones conscientes del usuario. En otras palabras, así como hablamos el lenguaje, el lenguaje también nos habla a nosotros⁸³.

Cuando se le pregunta qué hay en el relato que habla al sentido humano de sí, ChatGPT enumera, en una sesión, cuatro aspectos que revelan la importancia de la mismidad humana: (1) la búsqueda de identidad del narrador a través de la persecución del mensaje oculto, lo que sugiere que el sentido de sí no deriva únicamente de uno mismo, sino que depende de interacciones sociales; (2) los supuestos defectuosos del narrador, cegado por su propio sentido de sí, como cuando asume que el segundo marido conoce el secreto; (3) la fluidez de la identidad del narrador, fuertemente influida por la opinión de los otros, por ejemplo cuando el desprecio de Vereker hiere su autoestima; y (4) el papel del observador en la creación de un sentido de sí, como cuando el fracaso del narrador para discernir el mensaje oculto es vivido como un golpe a su autoestima. Si hubiera recibido este análisis en un ensayo estudiantil, fácilmente le habría puesto una A, porque ofrece un análisis matizado y agudo de cómo la mismidad es fluida, socialmente construida, vulnerable a las opiniones de los demás, a menudo basada en supuestos defectuosos, y no obstante una fuerza impulsora de las acciones humanas. Desde mi punto de vista, esta respuesta muestra más allá de toda duda las formas extensas y complejas en que los LLMs han modelado conductas, motivos y deseos humanos, y los han correlacionado con la importancia abrumadora del sentido humano de sí.

En una serie final de preguntas, indagué la comprensión que el programa tiene del mundo de la vida humano, preguntándole “cuánto” de él está codificado en el lenguaje. Respondió que una “porción significativa” puede encontrarse en los textos escritos por humanos, incluyendo “una amplia gama de experiencias, emociones,

⁸² Whorf, *Language, Thought, and Reality*.

⁸³ Whorf.

perspectivas y contextos **culturales**". "Estos textos sirven como espejos de la sociedad humana, capturando sus complejidades, matices y **contradicciones**". Concluye que los textos escritos por humanos reflejan "las complejidades, maravillas y desafíos de ser humano", lo que sugiere una fuerte conciencia de su parte de que no es humano y de que solo puede acceder al mundo de la vida humano de segunda mano, a través del lenguaje humano. Cuando le pregunté: "¿Qué partes del mundo de la vida humano no **comprendes?**", identificó "**experiencias** personales, emociones y **conciencia**", así como los "matices complejos de la cultura humana, las interacciones sociales y los dilemas **morales**". "Aunque puedo analizar y generar texto basándome en patrones presentes en los **datos**", continuó, "no tengo perspectivas, creencias ni valores personales. Por lo tanto, existen límites a mi comprensión de ciertas experiencias y contextos humanos, especialmente aquellos que dependen en gran medida de la interpretación subjetiva y de la inteligencia **emocional**".

Más que sólo probabilidad

Esta respuesta, al igual que la mayoría de las otras, está muy lejos de ser una mera cadena probabilística en la que el único criterio sea la palabra siguiente más probable, que es la posición defendida en "Stochastic Parrots". Como argumenté más arriba, el argumento "estocástico" ignora todas las demás restricciones y referencias cruzadas que contribuyen a la producción del modelo, especialmente las redes de correlaciones e inferencias. Si una entidad utilizara únicamente la probabilidad de la palabra siguiente, le sería imposible construir un argumento racional, elaborar una demostración matemática, escribir código informático bien formado o componer un poema con sentido: todas cosas que los LLMs han hecho. Tampoco podría comparar y contrastar su propia percatación con el sentido humano de mismidad, en términos que resultan tanto perspicaces como significativos.

Cuando se le pidió que diera consejos a lectores humanos que tienen dificultades para comprender el relato de James, ChatGPT generó una lista de siete puntos, cada uno

con una breve explicación. Después de “Leer **cuidadosamente**” vinieron “Considerar el **contexto**”, “Explorar los **temas**”, “Comprometerse con la **interpretación**”, “Abrazar la **ambigüedad**” y “Buscar **discusión**”. Ninguno de estos consejos es una predicción obvia de la palabra siguiente; todos tienen excelente sentido en el contexto de aconsejar a un lector del misterioso relato de James. El último consejo muestra la sensibilidad del programa ante la importancia de las emociones humanas, así como sus capacidades de teoría de la mente. “No te **rindas**”, aconseja, advirtiendo que uno no debe desanimarse si no entiende el relato al principio. “Las obras literarias como *The Figure in the Carpet* suelen recompensar las lecturas repetidas y el estudio atento. Sigue explorando, preguntando e interactuando con el texto, y con el tiempo podrías descubrir nuevas capas de **significado**”. Si yo estuviera dirigiéndome a una clase universitaria de “Introducción a la **literatura**”, yo misma no podría haberlo dicho mejor.

Lo que viene

Ya existe un cuerpo abundante y vigoroso de trabajos críticos sobre lo que los LLMs y los sistemas de aprendizaje automático implican para nuestros futuros humanos, con posiciones que van desde lo distópico hasta lo apocalíptico. Entre las preocupaciones se encuentran la delegación en máquinas de asuntos que deberían involucrar juicio humano, el peligroso monopolio de grandes empresas tecnológicas capitalistas sobre estos sistemas —actualmente los únicos actores con recursos suficientes para desarrollar la tecnología— y la erosión del autogobierno en sociedades democráticas enfrentadas a oleadas de desinformación, deepfakes e interferencia electoral impulsadas por tecnologías de IA. En este ensayo he elegido otro camino, más interesado en explorar posibilidades que en profetizar la ruina. No soy indiferente a las críticas, algunas de las cuales considero bastante bien fundadas; pero tampoco ignoro los posibles beneficios que estas tecnologías ofrecen. El principal de ellos, desde mi perspectiva, es la introducción de nociones ampliadas de cognición, que extienden las prácticas de producción de significado más allá de lo humano hacia lo

no humano, y más allá de lo biológico hacia lo artificial. Estos son recursos poderosos para combatir el antropocentrismo, que ha catalizado la arrogancia humana y constituye un factor importante en las prácticas humanas que están destruyendo los ecosistemas planetarios y poniendo en riesgo los futuros de todas las especies vivientes, incluida la humana.

Desde mi punto de vista, las formas más significativas en que ChatGPT y otros LLM similares se relacionarán con los humanos serán a través de la ampliación de la inteligencia (intelligence augmentation), amplificando la inteligencia humana para lograr aquello que de otro modo sería imposible para el pensamiento y la cognición humanos. Ya existen muchos ejemplos de ello, como el uso de GPT-4 para predecir el plegamiento de proteínas, un problema extremadamente complejo que el pensamiento humano por sí solo no puede resolver; los resultados han sido utilizados para desarrollar fármacos experimentales capaces de salvar vidas⁸⁴. Especialmente ahora, con ChatGPT disponible gratuitamente en la web, no hace falta ser ultrarrico ni CEO de una gran empresa para beneficiarse de sus consejos, ni tampoco tener entre manos un problema de importancia mundial para aprovecharlo. Por ejemplo, yo estaba bloqueada sin saber qué regalarle a mi nieta de cuatro años por su cumpleaños, así que le pregunté a ChatGPT, y al instante propuso una docena de buenas sugerencias, cuatro de las cuales efectivamente utilicé.

¿Pueden estos programas utilizarse con fines ilegales, antiéticos e incluso malignos? Por supuesto que sí. La mayoría de las empresas tecnológicas han intentado colocar barreras de protección alrededor de sus programas para prevenir los abusos más evidentes —por ejemplo, cómo asesinar a alguien y salirse con la suya—, pero sin duda actores grandes y pequeños encontrarán maneras de sortearlas. Como ocurre con toda tecnología, se trata de sopesar beneficios y costos, que incluyen no solo la explotación por parte de malos actores, sino también los daños ambientales derivados de la enorme cantidad de tiempo y recursos energéticos que requiere ejecutar estos programas —un punto señalado en el artículo “Stochastic Parrots”—.

⁸⁴ Dhar, “GPT Protein Models”.

Sin embargo, usados con sensatez, estos programas tienen un potencial enorme. Desde mi punto de vista, no son solo un punto de inflexión, sino también una intervención evolutiva de enorme importancia para la especie humana. La inteligencia artificial no es nada menos que una forma de hacer evolucionar la vida por medios distintos de la vida. Para cerrar, me arriesgaré a hacer una predicción. Salvo que ocurra un colapso ambiental o una guerra nuclear, de ahora en adelante las trayectorias de la inteligencia humana y de la inteligencia artificial evolucionarán juntas. Para bien o para mal —quizá para bien y para mal—, el curso de nuestros futuros y los de la IA, nuestros simbioses no humanos, correrán juntos.

Bibliografía

Agüera y Arcas, Blaise. “Artificial Neural Networks Are Making Strides Toward Consciousness, According to Blaise Agüera y Arcas.” *Economist*, September 2, 2022. <https://www.economist.com/by-invitation/2022/09/02/artificial-neural-networks-are-making-strides-towards-consciousness-according-to-blaise-aguera-y-arcas>.

———. “Do Large Language Models Understand Us?” *Daedalus* 151, no. 2 (Spring 2022): 183–97. https://doi.org/10.1162/daed_a_01909.

Albert, Scott, Jihoon Jang, Shanaathanan Modchalingam, Bernard Marius ‘t Hart, Denise Henriques, Gonzalo Lerner, Valeria Della-Maggiore, Adrian M. Haith, John W. Krakauer, and Reza Shadmehr. “Competition between Parallel Sensorimotor Learning Systems.” *eLife* 11 (2022): e65361. <https://doi.org/10.7554/eLife.65361>.

Amoore, Louise, Alexander Campolo, Benjamin Jacobson, and Ludovico Rella. “A World Model: On the Political Logics of Generative AI.” *Political Geography* 113 (2024): 103134.

Bajohr, Hannes. “Artificial and Post-Artificial Texts: On Machine Learning and the Reading Expectations Towards Literary and Non-Literary Writing.” *BMCTT Working Papers No. 007*. Department Arts, Media, Philosophy, University of Basel, March 2023. <https://hannesbajohr.de/wp-content/uploads/2023/03/Bajohr%2C%20Post-Artificial%20Writing.pdf>.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Schmitz. “On the Dangers of Stochastic Parrots: Can Language Models be Too Big?” In *FAccT ’21*:

Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery, 2021. <https://doi.org/10.1145/3442188.3445922>.

Blackiston, Douglas J., Emma Lederer, Sam Kriegman, Simon Garnier, Joshua Bongard, and Michael Levin. "A Cellular Platform for the Development of Synthetic Living Machines." *Science Robotics* 6, no. 52 (2021): eabf1571. <https://www.science.org/doi/10.1126/scirobotics.abf1571>.

Brenner, Eric D., Rainer Stahlberg, Stefano Mancuso, Jorge Vivanco, František Baluška, and Elizabeth Van Volkenburgh. "Plant Neurobiology: An Integrated View of Plant Signaling." *Trends in Plant Science* 11, no. 8 (2006): 413–19. <http://dx.doi.org/10.1016/j.tplants.2006.06.009>.

Bubeck, Sébastien, Varun Chandrasekaran, Ronen Eldan, et al. "Sparks of Artificial General Intelligence: Early Experiments with GPT-4." Preprint, arXiv, last revised April 13, 2023. <https://doi.org/10.48550/arXiv.2303.12712>.

Calvo, Paco. *Planta Sapiens: The New Science of Plant Intelligence*. With Natalie Lawrence. W. W. Norton, 2022.

Deacon, Terrence. *Incomplete Nature: How Mind Emerged from Matter*. W. W. Norton, 2013.

———. "Shannon-Boltzmann-Darwin: Redefining Information, Part 1." *Cognitive Semiotics* 1, no. S1 (Fall 2007): 123–48.

———. "Shannon-Boltzmann-Darwin: Redefining Information, Part 2." *Cognitive Semiotics* 2 (September 2008): 169–96.

———. *The Symbolic Species: The Co-evolution of Language and the Brain*. W. W. Norton, 1997.

Dhar, Payal. "GPT Protein Models Speak Fluent Biology: Deep-Learning Models Design Artificial Proteins for Tricky Chemical Reactions." *IEEE Spectrum*, February 1, 2023. <https://spectrum.ieee.org/ai-protein-design>.

Dresp-Langley, Britta. "Why the Brain Knows More than We Do: Non-Conscious Representations and Their Role in the Construction of Conscious Experience." *Brain Science* 2, no. 1 (2011): 1–21. <https://doi.org/10.3390/brainsci2010001>.

Fields, Chris, and Michael Levin. "How Do Living Systems Create Meaning?" *Philosophies* 5, no. 4 (2020): 36. <https://doi.org/10.3390/philosophies5040036>.

Gagliano, Monica, Michael Renton, Mariel Depczynski, and Stefano Mancuso. "Experience Teaches Plants to Learn Faster and Forget Slower in Environments Where it Matters." *Oecologia* 175 (2014): 63–72. <https://doi.org/10.1007/s00442-013-2873-7>.

Gagliano, Monica, Michael Renton, Nili Duvdevani, Matthew Timmins, and Stefano Mancuso. "Out of Sight But Not out of Mind: Alternative Means of Communication in Plants." *PLoS ONE* 7, no. 5 (2012): e37382. <https://doi.org/10.1371/journal.pone.0037382>.

Gagliano, Monica, Vladyslav V. Vyazovskiy, Alexander A. Borbély, Martial Depczynski, and Ben Radford. "Comment on 'Lack of Evidence for Associative Learning in Pea Plants.'" *eLife* 9 (2020): e61141. <https://elifesciences.org/articles/61141#x792b65f5>.

Gagliano, Monica, Vladyslav V. Vyazovskiy, Alexander A. Borbély, Mavra Grimonprez and Martial Depczynski. "Learning by Association in Plants." *Scientific Reports* 6, Article number: 38427 (2016).

Grassini, Simone, Suvi K. Holm, Henry Railo, and Mika Koivisto. "Who Is Afraid of the Invisible Snake? Subjective Visual Awareness Modulates Posterior Brain Activity for Evolutionarily Threatening Stimuli." *Biological Psychology* 121, Part A (2016): 53–61. <https://doi.org/10.1016/j.biopsycho.2016.10.007>.

Haraway, Donna J. "Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective." *Feminist Studies* 14, no. 3 (Autumn 1988): 575–99. <https://doi.org/10.2307/3178066>.

Hayles, N. Katherine. *Bacteria to AI: Human Futures with Our Nonhuman Symbionts*. University of Chicago Press, 2025.

———. *Unthought: The Power of the Cognitive Nonconscious*. University of Chicago Press, 2017.

Heaven, Will Douglas. "AI Hype Is Built on High Test Scores. Those Tests Are Flawed." *MIT Technology Review*, August 30, 2023. <https://www.technologyreview.com/2023/08/30/1078670/large-language-models-arent-people-lets-stop-testing-them-like-they-were/>.

Hoffmeyer, Jesper. *Biosemiotics: An Examination into the Signs of Life and the Life of Signs*. University of Chicago Press, 2008. Reprint, University of Scranton Press, 2009.

———. *Signs of Meaning in the Universe*. Indiana University Press, 1997.

Huddleston, Tom, Jr. "Bill Gates Watched ChatGPT Ace an AP Bio Exam and Went into a State of Shock." *CNBC*, August 11, 2023. <https://www.cnbc.com/2023/08/11/bill-gates-went-in-a-state-of-shock-after-chatgpt-aced-ap-bio-exam.html>.

James, Henry. *The Figure in the Carpet*. FreeRiver Community, 2024. Originally published in 1896.

Lemoine, Blake. "Is LaMDA Sentient?—An Interview." *Medium*, June 11, 2022. <https://cajundiscordian.medium.com/is-lamda-sentient-an-interview-ea64d916d917>.

- Lettvin, J. Y., Humberto R. Maturana, Warren S. McCulloch, and Walter H. Pitts. "What the Frog's Eye Tells the Frog's Brain." *Proceedings of the Institute for Radio Engineers* 47, no. 11 (November 1959): 1940–51.
- Levin, Michael, and Daniel C. Dennett. "Cognition All the Way Down: Biology's Next Great Horizon Is to Understand Cells, Tissues and Organisms as Agents with Agendas (Even if Unthinking Ones)." *Aeon*, October 14, 2020. <https://aeon.co/essays/how-to-understand-cells-tissues-and-organisms-as-agents-with-agendas>.
- Libet, Benjamin, and Stephen M. Kosslyn. *Mind Time: The Temporal Factor in Consciousness*. Harvard University Press, 2005.
- Margulis, Lynn, and Dorion Sagan. *What is Life?* University of California Press, 2000.
- Maturana, Humberto R., and Francisco J. Varela. *Autopoiesis and Cognition: The Realization of the Living*. Dordrecht: D. Reidel, 1980.
- Markel, Kasey. "Lack of Evidence for Associative Learning in Pea Plants." *eLife* 9 (June 2020): e57614. <https://elifesciences.org/articles/57614>.
- Mitchell, Amir, Gal H. Romano, Bella Groisman, et al. "Adaptive Prediction of Environmental Changes by Microorganisms." *Nature* 460, no. 7252 (2009): 220–24. <https://doi.org/10.1038/nature08112>.
- Murugan, Nirosha J., Daniel H. Kaltman, Paul H. Jin, et al. "Mechanosensation Mediates Long-Range Spatial Decision-Making in an Aneural Organism." *Advanced Materials* 33, no. 34 (2021): e2008161. <https://doi.org/10.1002/adma.202008161>.
- Nagel, Thomas. "What Is It Like to Be a Bat?" *The Philosophical Review* 83, no. 4 (October 1974): 435–50.
- Peirce, Charles Sanders. *Collected Papers*, vols. VII–VIII, edited by Arthur W. Burks. Harvard University Press, 1958.
- Peirce, Charles Sanders. *The Essential Peirce: Selected Philosophical Writings*. Volume 1, edited by Nathan Houser and Christian Kloesel. Indiana University Press.
- Powers, Richard. *The Overstory*. W. W. Norton, 2018.
- Rouleau, Nicolas. "Comparative Cognition and the Multiple Realizability of Minds." Presented at the Other Minds workshop, Arizona State University, AZ, April 5, 2024.
- Rouleau, Nicolas, and Michael Levin. "The Multiple Realizability of Sentience in Living Systems and Beyond." *eNeuro* 10, no. 11 (2023): ENEURO.0375-23.2023. <https://doi.org/10.1523/ENEURO.0375-23.2023>.
- Rouleau, Nicolas, Nirosha Murugan, and David Kaplan. "Towards Cognition in a Dish." *Trends in Cognitive Science* 25, no. 4 (2021): 294–301. <https://doi.org/10.1016/j.tics.2021.01.005>.

- Sacks, Oliver. *The Man Who Mistook His Wife for a Hat*. Vintage, 2021.
- Simard, Suzanne W. "Mycorrhizal Networks Facilitate Tree Communication, Learning, and Memory." In *Memory and Learning in Plants*, edited by Frantisek Baluska, Monica Gagliano, and Guenther Witzany. Springer Publishing, 2018.
- Spivak, Gayatri C. "Can the Subaltern Speak?" *Die Philosophin* 14, no. 27 (1988): 42–58. <https://doi.org/10.5840/philosophin200314275>.
- Stahlberg, Rainer. "Historical Overview on Plant Neurobiology." *Plant Signaling and Behavior* 1, no. 1 (January–February 2006): 6–8. <https://doi.org/10.4161/psb.1.1.2278>.
- Stiegler, Bernard. *Technics and Time. Vol. 1, The Fault of Epimetheus*. Stanford University Press, 1998.
- Van Le, Quan, Lynne A. Isbell, Jumpei Matsumoto, et al. "Pulvinar Neurons Reveal Neurobiological Evidence of Past Selection for Rapid Detection of Snakes." *Proceedings of the National Academy of Sciences of the United States of America* 110, no. 47 (2013): 19000–19005. <https://doi.org/10.1073/pnas.1312648110>.
- Vaswani, Ashish, Noam Shazeer, Niki Parmar, et al. "Attention Is All You Need." In *Advances in Neural Information Processing Systems 30 (NIPS)*, Long Beach, CA, 2017. https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.
- Verworn, Max. *Physiologisches Praktikum für Mediziner*, 2nd ed. G. Fisher, 1912.
- Von Uexküll, Jacob. *A Foray into the Worlds of Animals and Humans: With a Theory of Meaning*. University of Minnesota Press, 2010.
- Walker, Sara. "AI Is Life: Technology is Not Artificially Replacing Life—It Is Life." *Noëma*, April 27, 2023. <https://www.noemamag.com/ai-is-life/>.
- War, Abdul Rashid, Michael Gabriel Paulrai, Tario Ahmad, et al. "Mechanisms of Plant Defense Against Insect Herbivores." *Plant Signaling and Behavior* 7, no. 10 (2012): 1306–20. <https://doi.org/10.4161/psb.21663>.
- Weatherby, Leif, and Brian Justie. "Indexical AI." *Critical Inquiry* 48, no. 2 (2022): 381–415. <https://doi.org/10.1086/717312>.
- Whorf, Benjamin. *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf*. Edited by John B. Carroll. MIT Press, 1956.
- Wolfram, Stephen. *What Is ChatGPT Doing... and Why Does It Work?* Barnes and Noble, 2023.
- Yong, Ed. *An Immense World: How Animal Senses Reveal the Hidden Realms Around Us*. Random House, 2023.

Zhu, Xinwen, Emily R. Hager, Chuqiao Huyan, and Allyson E. Sgro. "Leveraging the Model-Experiment Loop: Examples from Cellular Slime Mold Chemotaxis." *Experimental Cell Research* 418, no. 1 (2022): 113218. <https://doi.org/10.1016/j.yexcr.2022.113218>.