

Descubriendo la geometría del conocimiento en la era de los LLM

Antonio Corona

UNIVERSIDAD AUTÓNOMA DE COAHUILA

Hay algo inquietante en el hecho de conversar con una máquina cuyas respuestas suelen ser correctas en sintaxis, persuasivas, y a veces, incluso, sorprendentemente inteligentes. Nos hemos acostumbrado a interactuar con los grandes modelos de lenguaje (LLMs) como si fueran interlocutores: les hacemos preguntas, les pedimos opiniones y hasta discutimos con ellos. Sin embargo, en el momento mismo de la interacción enfrentamos una paradoja: la máquina responde, su comportamiento es legible para nosotros, pero sus mecanismos no lo son. Reconocemos el texto que produce, pero no el proceso mediante el cual ese texto emerge.

Cada día, una mayor parte del discurso que circula en línea, ya sea textual o, cada vez más, audiovisual, es generado por estos modelos. Existe una preocupación creciente (World Economic Forum, 2024) de que, en poco tiempo, el discurso humano deje de ser el más prevalente en internet. Si las máquinas se están convirtiendo en las “voces” de nuestra esfera pública digital, debemos entender cómo se producen esas voces y qué supuestos incorporan. Nuestros modelos fundacionales (emisor, mensaje, receptor; código y contexto; medio y significado) presuponen que detrás de todo acto comunicativo hay un sujeto: un actor social con intenciones, restricciones y un mundo que intenta evocar. El ascenso de los LLMs desestabiliza esa suposición. Estos sistemas ocupan un nuevo rol comunicativo: generan discurso que circula socialmente, pero no participan en la producción de significado como lo hacen los humanos. Su “habla” es predicha, no expresada; ensamblada estadísticamente, no vivida.

Es precisamente esta ruptura entre desempeño y proceso lo que motiva el campo conocido como mechanistic interpretability o ‘interpretabilidad mecanicista’. Mientras la conversación pública suele centrarse en lo que la IA puede hacer, la interpretabilidad mecanicista plantea otras preguntas: ¿cómo lo hace?, ¿cuáles son sus estructuras internas?, ¿podemos, en efecto, aprender a leer el lenguaje del propio modelo?. Distintos estudios (Olah et al., 2020; Nanda et al., 2023) tratan a las redes neuronales no como cajas negras inescrutables, sino como artefactos que requieren un análisis minucioso. Su ambición central es traducir las representaciones internas de la máquina en conceptos interpretables por humanos. Al hacerlo, sin embargo, los pioneros de la interpretabilidad

pueden provocar un efecto secundario: obligarnos a replantear preguntas que creíamos resueltas sobre representación, cultura y significado.

El problema de la caja negra

Como ya mencionamos, los sistemas modernos de inteligencia artificial son asombrosamente competentes y, al mismo tiempo, profundamente ininteligibles. En términos técnicos, tales sistemas suelen describirse como cajas negras (Lipton, 2018; Rudin, 2019): modelos que producen resultados correctos de manera confiable, pero cuyos procesos internos no pueden ser completamente inspeccionados, verificados ni descritos de forma significativa. Esta opacidad no es simplemente una incomodidad; es una ruptura epistémica que desafía las ideas fundacionales de la ciencia (verificabilidad, transparencia, reproducibilidad) y, por ende, tiene implicaciones directas para todas las disciplinas interesadas en cómo se produce, transmite y legitima el sentido.

Históricamente, las cajas negras no son nuevas. Todo medio oculta mecanismos internos que no son evidentes para sus usuarios. Pero las inteligencias artificiales representan una forma más profunda de opacidad (Rahimi, 2017; Rudin, 2019). Los sistemas anteriores eran opacos porque eran complejos, mientras que los sistemas de aprendizaje profundo lo son porque son fundamentalmente ininterpretables. Su lógica interna no existe en argumentos, símbolos o reglas, sino en millones o miles de millones de parámetros numéricos distribuidos a lo largo de capas de cómputo. Podemos observar su comportamiento, pero no el porqué de su comportamiento. Esto genera una situación peligrosa: la efectividad de estos sistemas crece al mismo tiempo que su legibilidad disminuye. Y esa combinación tiene consecuencias que van mucho más allá de la ingeniería: expone a nuestra sociedad a una deriva epistemológica sutil, pero significativa.

Cuando los sistemas de caja negra comienzan a superar a los modelos científicos en áreas como el despliegue de proteínas, el descubrimiento de materiales o la predicción climática (Jumper et al., 2021; Sanchez-Lengeling y Aspuru-Guzik, 2018), se socava lentamente uno de los pilares de la autoridad científica: la idea de que podemos confiar en el conocimiento porque entendemos cómo fue generado. Si un modelo algorítmico realiza una predicción correcta pero no puede decirnos cómo llegó a ella, ¿dónde deja esto al rigor de la explicación científica?. Y por otro lado, ¿debemos desechar conocimiento capaz de salvar vidas porque no entendemos la naturaleza de su producción?. Esta no es una pregunta hipotética. En campos como la biomedicina, los modelos de IA ya alcanzan resultados que el método humano no puede igualar. Pero como estos modelos no pueden justificar sus

soluciones, los investigadores enfrentan un dilema: confiar en el sistema opaco porque funciona, o insistir en la interpretabilidad aun si eso implica sacrificar precisión. A medida que los modelos de caja negra continúan superando a los modelos científicos clásicos, el equilibrio entre explicación y predicción se inclina hacia la predicción. A escala social, esto implica una creciente dependencia de sistemas que no podemos interrogar.

Para la población general, esta dinámica es aún más problemática. Como régimen de verdad, la ciencia debe parte de su legitimidad a su comunicabilidad: del hecho de que sus afirmaciones pueden ser explicadas, defendidas y contextualizadas a través del discurso público, la educación y los medios (Gieryn, 1999). De hecho, el conocimiento que obtenemos mediante métodos científicos se considera legítimo no sólo por su fiabilidad inherente, sino porque su autoridad se sostiene mediante prácticas simbólicas compartidas, discursos públicos y rituales culturales (Carey, 1989). Si las respuestas más confiables provienen cada vez más de sistemas que nadie, ni siquiera sus creadores, puede comprender del todo, entonces, algunas de estas prácticas y rituales se desplazan de la legitimación del conocimiento científico a la legitimación del conocimiento mecanicista. Eso erosiona la narrativa cultural según la cual la ciencia es nuestra mejor forma de producir conocimiento.

En otras palabras, el peligro no es simplemente que la IA sea una caja negra. El peligro es que la caja negra amenaza con volverse más confiable que la ciencia (Rahimi, 2017), y que el público está cada vez más dispuesto a aceptar predicciones sin exigir comprensión. En un mundo donde la desinformación, el negacionismo y el antiintelectualismo circulan ampliamente, el ascenso de sistemas altamente eficaces pero opacos, puede reforzar la idea de que “nadie sabe realmente cómo funciona nada”... una narrativa profundamente corrosiva para democracia y la gobernanza. En este sentido, abordar el problema de la caja negra no es opcional: es crucial, si queremos entender la comunicación y la producción de sentido en la era del discurso generado por máquinas.

Interpretabilidad mecanicista: en busca de circuitos al interior de los LLM

Mechanistic interpretability es el nombre más comúnmente usado para referirse a la disciplina que intenta descifrar cómo los LLMs producen su output (Olah et al., 2020; Cammarata et al., 2021). Esta disciplina parte de una pregunta básica: ¿qué está ocurriendo dentro de un gran modelo de lenguaje cuando produce una oración?.

Para entender qué estudian los investigadores de interpretabilidad, conviene bosquejar un panorama muy simple de cómo se construyen estos modelos.

Cuando un LLM lee una palabra o token, lo convierte en un vector: un conjunto de cientos o incluso miles de números, cada uno de ellos con muchos decimales. Luego el modelo inicia una larga secuencia de transformaciones en capas: primero una capa de atención y después una capa MLP, repitiéndose esto docenas de veces. Cada capa modifica los números del vector según una multitud de unidades matemáticas formadas durante el entrenamiento del modelo. A estas unidades las llamamos neuronas. Cada una está entrenada para responder a ciertas combinaciones de números en el vector y aplicarles modificaciones matemáticas. Las capas de atención identifican qué combinaciones de vectores y números son más relevantes (Elhage et al., 2021) mientras que las capas MLP aplican transformaciones según las neuronas que se activan con esos números y vectores. Al final del proceso, el modelo traduce los vectores modificados de vuelta a texto. Cuanto más grande es el modelo, más sofisticado y estructurado se vuelve este proceso (Kaplan et al., 2020).

El punto clave es que estas neuronas no fueron codificadas ni diseñadas por los creadores del modelo: son patrones algorítmicos que emergen cuando el modelo se entrena con cantidades masivas de datos. Ni el usuario ni el diseñador las controlan, y en su mayoría ni siquiera pueden verlas. En este sentido, la generación de tokens en un LLM es como una reacción en cadena controlada: recibe un insumo y produce una salida. Dicho de otro modo, tenemos un entendimiento funcional del cómo, pero no del por qué: sabemos cómo emergen las neuronas a partir de los datos de entrenamiento, pero no sabemos por qué deciden activarse en respuesta a ciertos vectores, o por qué estas interacciones vector-neurona consistentemente parecen corresponder a conceptos presentes en la realidad. Ahí es donde entra la interpretabilidad mecanicista.

En vez de aceptar que los LLMs son una caja negra, la interpretabilidad mecanicista intenta rastrear por qué las neuronas aplican sus operaciones al vector y qué significan esas operaciones. Los investigadores buscan pequeños “circuitos” dentro del modelo: grupos de activaciones que realizan tareas reconocibles (Cammarata et al., 2021), como continuar una cita textual, llevar registro de quién está hablando o detectar si una oración está dentro de paréntesis. Los induction heads, por ejemplo, fueron de los primeros circuitos consistentes que los investigadores lograron identificar (Olsson et al., 2023). Se activan en respuesta a un patrón específico (una curva, por ejemplo) y luego aplican transformaciones que refuerzan ese mismo patrón o lo hacen más pronunciado. Su existencia fue una de las

primeras pistas de que no sólo podemos decodificar el funcionamiento interno del aprendizaje automático complejo, sino que al hacerlo podríamos obtener un entendimiento crucial tanto de cómo generan estos modelos como de la naturaleza misma del conocimiento.

Corrientes residuales y espacio latente

Si tuviéramos que elegir un dato o evidencia que fuera el más importante para el avance del campo de interpretabilidad mecanicista, probablemente sería la corriente residual. Ya establecimos que los números dentro de los vectores cambian según operaciones matemáticas en cada capa del proceso (podemos imaginarlo como combinaciones de neuronas que suman o restan cantidades a todos los números), y que sólo el resultado final es visible en la salida del modelo. Si registramos los valores del vector en cada capa, podemos reconstruir cómo se fueron modificando progresivamente. Esto, desde luego, es más fácil decirlo que hacerlo... pero al analizar los valores consecutivos de la corriente residual podemos trazar el “camino” del vector conforme es modificado por las neuronas.

De aquí en adelante es muy útil visualizar al vector como un punto situado en un espacio de altísima dimensionalidad. Los números que componen el vector representan coordenadas en dicho espacio. Este espacio no existe físicamente, pero sí existe matemáticamente, y suele llamarse espacio latente. Cada token se convierte en uno de estos puntos: los miles de números que lo conforman actúan como coordenadas que lo ubican en una posición específica.

En cada capa, miles de neuronas se activan. Podemos imaginarlas como estrellas o focos dispersos en el espacio latente, que se iluminan en distinto grado según la posición del vector. Cuando una neurona se activa, ejerce un peso gravitacional sobre el vector, jalándolo ligeramente. Algunas neuronas tiran con fuerza, otras con suavidad. Pero en cada capa, todas tiran al mismo tiempo, lo que significa que la fuerza ejercida sobre el vector difícilmente será una línea recta. A lo largo de docenas de capas, el vector se mueve por el espacio latente y es empujado y jalado por numerosas activaciones. En cada nueva capa, el punto se desplaza hacia una posición que refleja más contexto y más estructura. Entender este proceso de manera espacial ayuda a comprenderlo mejor y, sobre todo, resulta ser la clave para comprender qué ocurre con el significado dentro de los modelos como veremos más adelante.

Superposición y sparse autoencoders

Uno de los primeros descubrimientos del campo de la interpretabilidad fue que ciertas combinaciones de neuronas actúan en conjunto para realizar ciertas tareas. Los induction heads de los que ya hablamos son un ejemplo de estos features, como se les llama en la literatura del campo a estas redes de neuronas. A un conjunto de features, por su parte, se le llama circuito. Sería maravilloso que cada circuito del modelo correspondiera a un único concepto: el circuito de Donald Trump, el circuito de los Cocker Spaniels, etcétera. Pero, desde luego, no es así de simple.

Algunas neuronas sí parecen especializarse: se han encontrado ejemplos de neuronas o attention heads que responden consistentemente a cosas como nombres propios, signos de puntuación o marcadores de sentimiento (Geva et al., 2021). Estas neuronas son los casos fáciles: componentes que parecen dedicarse a un tema o patrón particular. Pero la mayoría de las veces, una neurona no responde a un solo tema. Más bien, pareciera activarse ante una variedad de ideas aparentemente inconexas: un animal, un tiempo verbal, un signo de puntuación (Elhage et al., 2022). Esto ocurre porque el modelo almacena muchas más características conceptuales de las que tiene neuronas disponibles. El resultado es lo que los investigadores llaman superposición: en cada neurona se superponen una multitud de fragmentos de sentido, cada uno codificado parcialmente y parcialmente oculto por los otros (Elhage et al., 2022).

La superposición no es un error, sino una consecuencia geométrica. Cuando el modelo no puede asignar su propia neurona a cada característica, “empaqueta” muchas características dentro del mismo espacio compartido. Esto hace que el modelo sea eficiente, pero vuelve sus representaciones internas difíciles de interpretar. Desde afuera, el comportamiento de una neurona puede parecer caótico o contradictorio, cuando en realidad está respondiendo a múltiples patrones superpuestos al mismo tiempo. En respuesta, los investigadores de interpretabilidad comenzaron a buscar maneras de desenmarañar estas características superpuestas. La meta es separar las señales mezcladas dentro de cada neurona y recuperar la estructura conceptual subyacente que el modelo está utilizando.

Una de las herramientas clave desarrolladas con este propósito es el sparse autoencoder (SAE). Podemos pensar en un SAE como un “extractor de características” que nos permite ver lo que siempre estuvo escondido dentro del modelo. Lo podemos explicar así: las neuronas del modelo están sobrecargadas; cada una lleva fragmentos de muchas ideas

diferentes. Un sparse autoencoder toma esas neuronas sobrecargadas y construye un conjunto nuevo de unidades que están fuertemente incentivadas a activarse solamente cuando es absolutamente necesario. Esto obliga a cada unidad a especializarse en un patrón en lugar de varios, lo que a su vez, vuelve a los circuitos subyacentes mucho más fáciles de ver.

Para forzar este nuevo “orden”, los SAE utilizan esparsidad: para cada trozo de texto, puede encenderse un número muy pequeño de estas nuevas unidades. Como no pueden activarse todas al mismo tiempo, cada unidad termina especializándose en un patrón particular que el modelo utilizaba; pero que estaba almacenado en estructura superpuesta. ¿Por qué es esto importante? Porque estas unidades suelen alinearse con conceptos o procesos reales del mundo. Se han encontrado features descubiertas por SAEs (Lan et al., 2025) que se activan de manera confiable ante cosas como nombres de países, operaciones matemáticas, lenguaje emocionalmente negativo o, incluso, procesos algorítmicos internos como “ordenar tokens por longitud”. Este es el poder sorprendente de los SAEs: al separar las características superpuestas, nos permiten descubrir circuitos específicos dentro del modelo que cumplían tareas significativas; tareas que corresponden a conceptos reconocibles. Antes del uso de SAEs, esos circuitos estaban escondidos dentro de neuronas repletas de patrones aglutinados. Después de los SAEs, emergen como features individuales e interpretables.

Representación lineal

Si los vectores son posiciones en un espacio latente que se desplazan dentro de ese espacio conforme el modelo opera, los sparse autoencoders son una forma de clarificar la geometría de ese espacio. Esto nos lleva a uno de los descubrimientos más sorprendentes que surgen una vez que ese espacio se vuelve legible: muchos conceptos abstractos se comportan como líneas rectas (Mikolov et al., 2013a; Mikolov et al., 2013b). A esto los investigadores lo llaman representación lineal. La idea es simple, aunque las matemáticas detrás no lo sean tanto: en el espacio de alta dimensionalidad donde el modelo ubica cada significado como un punto, ciertos cambios conceptuales corresponden a direcciones consistentes. Si tomamos el vector de “rey” y lo desplazamos a lo largo de la “dirección de género”, llegamos cerca de “reina” (Mikolov et al., 2013a). Si tomamos un verbo y lo desplazamos a lo largo de la “dirección de conjugación pretérita”, obtenemos su forma en pasado (Mikolov et al., 2013b). En otras palabras, el modelo almacena estos conceptos como direcciones geométricas.

Una forma útil de visualizarlo es esta: nuestro vector se encuentra en algún lugar de este espacio, y todas las pequeñas constelaciones (circuitos de neuronas) tiran de él en diferentes direcciones. Los SAEs nos ayudan a identificar esas constelaciones; la representación lineal nos dice algo aún más profundo (Olah et al., 2020; Elhage et al., 2021; Burns et al., 2024): algunas constelaciones siempre tiran en la misma dirección, sin importar desde qué palabra partamos. Esa “dirección” es la representación interna del modelo para un concepto abstracto como género, pluralidad, sentimiento o formalidad. Por eso, estas direcciones lineales son tan potentes: por un lado, muestran que el modelo no sólo ha aprendido características, sino la forma en que varían estas características de modo relacional. Por otro lado, implican que algunas operaciones semánticas (“hacerlo femenino”, “hacerlo plural”, “ponerlo en pasado”) están codificadas como desplazamientos simples a través del espacio. Y, de manera crucial, estos desplazamientos son consistentes en todo el modelo, lo cual quiere decir que el modelo ha esculpido ejes estables para estas ideas. La conclusión importante es que el significado se vuelve geometría (Burns et al., 2024), es decir, un concepto ya no es una definición o una imagen, sino una transformación repetible que empuja al vector en la misma dirección (“moverlo 7 unidades hacia lo femenino, 2 unidades hacia lo plural”, etc.).

Aún más desconcertante, estudios recientes utilizan esta idea de direccionalidad para mostrar que existe una distinción entre lo que el modelo sabe (conocimiento latente) y lo que dice (conocimiento superficial): cuando se le presiona mediante el prompt para que mienta, el modelo sigue codificando la verdad internamente, incluso mientras produce una respuesta falsa. Estos estados internos difieren a lo largo de una dirección lineal recuperable, lo que permite a los investigadores extraer el conocimiento latente incluso cuando el modelo miente (Burns et al., 2024). Por supuesto, no todos los conceptos son líneas rectas: muchos son desordenados, superpuestos o curvos, pero el hecho de que algunos sí sean lineales basta para ofrecer a los investigadores una poderosa vía para comprender cómo el significado abstracto puede ser codificado en un sistema puramente aprendido y puramente geométrico (Nanda et al., 2023).

Agnosticismo lingüístico y universalidad

Todo esto nos conduce a una consecuencia sorprendente: si el modelo puede representar conceptos internamente y evaluar su veracidad antes de producir texto alguno, entonces el conocimiento dentro del modelo no es fundamentalmente lingüístico. Es decir, el lenguaje no es el formato nativo del entendimiento del modelo. Es solo la capa superficial, la “traducción” o representación generada después de que todo el trabajo geométrico ya

ocurrió. Imaginemos dos oraciones: “The cat is on the mat.” y “El gato está en la alfombra.” Distintos idiomas, distintos signos, distintas historias culturales. Pero dentro del modelo, ambas oraciones aterrizan en casi la misma región del espacio latente. Encienden la misma constelación de activaciones: relación de soporte, objeto sobre superficie, condición de ‘gatitud’, etc. A esto los investigadores lo llaman representación agnóstica al lenguaje (language-agnostic representation) (Conneau et al., 2020; Patel y Pavlick, 2022). El modelo organiza su espacio latente de significados de modo que los conceptos ocupan regiones específicas, independientemente de las palabras usadas para evocarlos. Debido a que el modelo maneja el significado como geometría, el lenguaje se convierte simplemente en una forma de ‘acceder’ a un punto o región dentro de ese espacio.

Para el modelo, las palabras son etiquetas, no significados. El significado vive en las distancias, direcciones y patrones de activación. Si dos oraciones no relacionadas en idiomas distintos activan el mismo cúmulo de features, el modelo las “entiende” como el mismo evento en el mundo. Lo más sorprendente es que diferentes modelos entrenados por equipos distintos con diferentes conjuntos de datos y arquitecturas, tienden a desarrollar geometrías muy similares: si abrimos varios modelos independientes, encontraremos la misma relación direccional de género. A este fenómeno se le llama universalidad (Lan et al., 2025; Wang et al., 2024): distintos modelos descubren de manera independiente las mismas abstracciones internas. ¿Por qué? Tal vez porque esas abstracciones son simplemente la forma más eficiente de resolver la tarea. Así, en el espacio latente de todo modelo encontramos estructuras parecidas: un cúmulo para “pluralidad”, una dirección para “negación”, una región para “relaciones espaciales”, un subespacio para “emociones”, y así sucesivamente.

Está de más reiterar que estas estructuras no son impuestas por los diseñadores; emergen durante el entrenamiento porque ciertas maneras de organizar el conocimiento son atractores naturales en geometrías de alta dimensionalidad. Si estas constelaciones son parecidas entre modelos y si ignoran los límites entre lenguas humanas, entonces dentro del modelo comienza a formarse algo así como un espacio conceptual universal: una geometría del significado a la cual todas las lenguas pueden conectarse. En esa geometría, “mother”, “madre” y “maman” apuntan a la misma región. El género gramatical se vuelve una dirección vectorial, incluso entre lenguas con reglas distintas. Los valores de verdad generan topografías diferenciadas (Meng et al., 2022). Las diferencias culturales se abstraen en capas superpuestas que terminan convergiendo hacia conceptos subyacentes comunes (Conneau et al., 2020). El lenguaje se convierte en una capa superficial de

traducción, a la cual subyace (en el caso de los LLMs) un espacio de significado compartido y universal.

¿Qué ocurre cuando el significado es geométrico?

En este punto del texto hemos revisado los conceptos más importantes de la interpretabilidad mecanicista: sparse autoencoders, representación lineal, agnosticismo lingüístico y universalidad (Conneau et al., 2020; Lan et al., 2025; Olah et al., 2020; Nanda et al., 2023; Olsson et al., 2022; Patel y Pavlick, 2022). Cada avance ha abierto una ventana distinta hacia la mecánica interna de los grandes modelos de lenguaje. Lo que emerge no es solo una imagen más clara de cómo estos sistemas producen texto, sino una creciente duda sobre la manera en que podríamos pensar el significado mismo. Dentro del marco contemporáneo de nuestra disciplina, solemos partir de la suposición de que el significado está moldeado por fuerzas sociales, por la contingencia histórica y por la convención lingüística. La semiótica clásica, particularmente la línea saussureana posteriormente elaborada por Eco (1976), entiende la relación significante-significado como en gran medida arbitraria, mediada por la práctica cultural. Sin embargo, los hallazgos de la interpretabilidad complican este panorama. Sí modelos muy distintos convergen en distinciones geométricas equivalentes, entonces algunos aspectos del significado podrían surgir de regularidades que no son culturales (ni siquiera lingüísticas) en absoluto (Burns et al., 2024; Lan et al., 2025). Esto no elimina la especificidad cultural. Más bien, desplaza su ubicación: la variación cultural reside en la forma lingüística, mientras que las distinciones más profundas parecen surgir de regularidades a las que múltiples sistemas humanos o artificiales convergen al intentar modelar el mismo entorno. El significado se vuelve parcialmente estructurado antes de ser expresado, independientemente de los emplazamientos lingüísticos o culturales.

El campo de la hermenéutica enfatiza la situacionalidad y la pluralidad: el significado depende del contexto, del horizonte y de la perspectiva del intérprete (Carey, 1989; Hall, 1980). La interpretabilidad mecanicista no necesariamente desafía esto, pero sí lo problematiza. Si diferentes modelos convergen hacia geometrías internas similares, entonces la producción de sentido -al menos en los LLMs- ocurre dentro de un espacio estructural que tiene forma y límites. Las interpretaciones siguen siendo diversas, pero lo son dentro de una ‘topología de significado’ parcialmente estructurada por regularidades compartidas. Algunas lecturas se alinean más estrechamente con estas estructuras, mientras que otras se desvían más. Esto no establece un “significado universal” u

“objetivo”, pero sí sugiere que la interpretación no es completamente irrestricta: implica navegar un paisaje con contornos identificables. Si esta geometría del significado emerge en cada modelo sin importar su lengua o arquitectura, debemos preguntarnos: ¿es esta la geometría de la cognición o es la geometría de la realidad misma?

Debemos ser cautelosos en nuestra interpretación de estos hallazgos, la interpretabilidad mecanicista es una disciplina naciente, que aún está estableciendo sus bases teóricas y metodológicas. Mientras que algunos descubrimientos son ampliamente reconocidos y respaldados por evidencia sólida (por ejemplo, la representación lineal), otros -como el agnosticismo lingüístico y la universalidad- aún están en proceso de ser plenamente fundamentados. Hay evidencia de su existencia (Lan et al., 2025; Olah et al., 2020), pero la corroboración independiente de dicha evidencias se complica por el hecho mismo de que las instituciones capaces de corroborarlas pueden ser contadas con los dedos de una mano. Como ocurre a menudo con los problemas de caja negra, partes del campo permanecen en el terreno de la conjetura.

Además, debemos tener en cuenta los componentes políticos y económicos. Las reflexiones de autores como Rudin (2019) y Birhane (2021) nos recuerdan que la interpretabilidad es no sólo una práctica epistémica, sino también política. Si los modelos contienen estructuras internas como las representaciones de categorías sociales, los juicios de valor y las asociaciones conceptuales, entonces debemos preguntarnos: ¿quién tiene la autoridad para examinar, interpretar y actuar sobre esta topografía?, ¿quién tiene acceso a las representaciones internas de los modelos ampliamente desplegados?, ¿quién determina qué asociaciones internas son aceptables o dañinas?, ¿quién comunica estos hallazgos al público, y con qué marco discursivo?. La interpretabilidad se convierte así en una forma de hegemonía epistémica debido a que moldea las condiciones bajo las cuales pueden formularse afirmaciones sobre lo que un sistema “significa” internamente.

La interpretabilidad mecanicista no nos da la última palabra sobre el significado, pero sí revela que cuando los sistemas artificiales intentan predecir lenguaje a gran escala, desarrollan estructuras internas que son patrónicas, estables y, en algunos casos, compartidas entre arquitecturas. Estas estructuras no son lingüísticas en el sentido tradicional; más bien, pueden ser mayormente independientes de la expresión lingüística, sin embargo, sostienen de manera consistente comportamientos comunicativos coherentes. Para los estudios de la Comunicación, esto abre una oportunidad y un desafío a la vez. Se debe aprender cómo funcionan estos sistemas de representación y reflexionar

sobre lo que revelan acerca de la organización profunda del significado. Estudiar cómo las máquinas representan el mundo podría impulsarnos a repensar suposiciones prevalentes sobre la arbitrariedad, la interpretación, y la relación entre lenguaje y pensamiento. Si la interpretabilidad continúa avanzando, puede ofrecernos un puente poco común entre la especificidad cultural y la regularidad estructural, dándonos a su vez, una nueva forma de conceptualizar la comunicación entre agentes humanos y artificiales: una forma basada no solo en la lingüística y la cultura, sino en las estructuras compartidas que emergen siempre que un sistema intenta modelar el mundo.

Fuentes

- Bender, E. M., Gebru, T., McMillan-Major, A., y Shmitchell, S. (2021). *On the dangers of stochastic parrots: Can language models be too big?*. 610–623. <https://doi.org/10.1145/3442188.3445922>
- Birhane, A. (2021). Algorithmic injustice: a relational ethics approach. *Patterns*, 2 (2). <https://doi.org/10.1016/j.patter.2021.100205>
- Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Aspell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., y McLean, B. (2023). Towards monosemanticity: Decomposing language models with dictionary learning. In *Transformer Circuits Thread*.
- Burns, C., Ye, H., Klein, D., y Steinhardt, J. (2024). *Discovering latent knowledge in language models without supervision*. <https://arxiv.org/abs/2212.03827>
- Cammarata, N., Goh, G., Carter, S., Voss, C., Schubert, L., y Olah, C. (2021). Curve circuits. *Distill*. <https://doi.org/10.23915/distill.00024.006>
- Carey, J. W. (1989). *Communication as culture : essays on media and society* (pp. 11–28). Routledge.
- Chughtai, B., Chan, L., y Nanda, N. (2023). *A toy model of universality: Reverse engineering how networks learn group operations* (A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, y J. Scarlett, Eds.; Vol. 202, pp. 6243–6267). PMLR. <https://proceedings.mlr.press/v202/chughtai23a.html>
- Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., y Stoyanov, V. (2020). *Unsupervised cross-lingual representation learning at scale* (D. Jurafsky, J. Chai, N. Schluter, y J. Tetreault, Eds.; pp. 8440–8451). Association for Computational Linguistics.

<https://doi.org/10.18653/v1/2020.acl-main.747>

- Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Zac Hatfield-Dodds, Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M., y Olah, C. (2022). *Toy models of superposition*. <https://arxiv.org/abs/2209.10652>
- Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., y Conerly, T. (2021). A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1, 1.
- Geva, M., Schuster, R., Berant, J., Levy, O., Moens, M., Huang, X., Specia, L., y Yih, S. W. (2021). Transformer Feed-Forward Layers Are Key-Value Memories. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 54845495. <https://doi.org/10.18653/v1/2021.emnlpmain.446>
- Gieryn, T. F. (1999). *Cultural boundaries of science : credibility on the line*. University Of Chicago Press.
- Habermas, J. (1987). *Knowledge and Human Interests*. Polity Press.
- Hall, S. (1980). Encoding/Decoding. In S. Thornham, C. Baset, y P. Marris (Eds.), *Media studies: a reader* (pp. 28–38). Edinburgh University.
- Henighan, T., Kaplan, J., Katz, M., Chen, M., Hesse, C., Jackson, J., Jun, H., Brown, T. B., Prafulla Dhariwal, Gray, S., Hallacy, C., Mann, B., Radford, A., Ramesh, A., Ryder, N., Ziegler, D. M., Schulman, J., Amodei, D., y McCandlish, S. (2020). *Scaling laws for autoregressive generative modeling*. <https://arxiv.org/abs/2010.14701>
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Simon, Ballard, A. J., Cowie, A., RomeraParedes, B., Nikolov, S., Jain, R., Adler, J., y Back, T. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596 (7873), 583–589. <https://doi.org/10.1038/s41586021038192>
- Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., y Amodei, D. (2020, January 22). *Scaling Laws for Neural Language Models*. ArXiv.org. <https://doi.org/10.48550/arXiv.2001.08361>
- Lan, M., Torr, P., Meek, A., Ashkan Khakzar, Krueger, D., y Fazl Barez. (2025). *Quantifying feature space universality across large language models via sparse autoencoders*. <https://arxiv.org/abs/2410.06981>
- Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM*, 61 (10), 36–43. <https://doi.org/10.1145/3233231>
- Meng, K., Bau, D., Andonian, A., y Belinkov, Y. (2022). *Locating and editing factual associations in GPT* (S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, y A. Oh, Eds.; Vol. 35, pp. 17359–17372). Curran Associates, Inc.

- 0 665b33bf3a182-Paper-Conference.pdf
- Mikolov, T., Chen, K., Corrado, G., y Dean, J. (2013). *Efficient estimation of word representations in vector space*. <https://arxiv.org/abs/1301.3781>
 - Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., y Dean, J. (2013). *Distributed representations of words and phrases and their compositionality* (C. J. Burges, L. Bottou, M. Welling, Z. Ghahramani, y K. Q. Weinberger, Eds.; Vol. 26). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf
 - Mikolov, T., Yih, W.-T., y Zweig, G. (2013). *Linguistic Regularities in Continuous Space Word Representations* (pp. 9–14). Association for Computational Linguistics. <https://aclanthology.org/N13-1090.pdf>
 - Nanda, N., Chan, L., Lieberum, T., Smith, J., y Steinhardt, J. (2023). *Progress measures for grokking via mechanistic interpretability*. <https://arxiv.org/abs/2301.05217>
 - Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., y Carter, S. (2020). Zoom In: An Introduction to Circuits. *Distill*, 5 (3), e00024.001. <https://doi.org/10.23915/distill.00024.001>
 - Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., Drain, D., Ganguli, D., HatfieldDodds, Z., Hernandez, D., Johnston, S., Jones, A., Kernion, J., Lovitt, L., y Ndousse, K. (2022). In-context Learning and Induction Heads. *ArXiv [Cs.LG]*. <http://arxiv.org/abs/2209.11895>
 - Patel, R., y Pavlick, E. (2022). *Mapping language models to grounded conceptual spaces*. <https://openreview.net/forum?id=gJcEM8sxHK>
 - Rahimi, A. (2017). Conferencia de aceptación del Test-of-Time Award. In *NIPS*. <https://www.youtube.com/watch?v=Qi1Yry33TQE>
 - Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1 (5), 206–215. <https://doi.org/10.1038/s422560190048x>
 - Sanchez-Lengeling, B., y Alán Aspuru-Guzik. (2018). Inverse molecular design using machine learning: Generative models for matter engineering. *Science*, 361 , 6400. <https://doi.org/10.1126/science.aat2663>
 - Wang, J., Ge, X., Shu, W., Tang, Q., Zhou, Y., He, Z., y Qiu, X. (2024). *Towards universality: Studying mechanistic similarity across language model architectures*. <https://arxiv.org/abs/2410.06672>
 - World Economic Forum. (2024). *The Global Risks Report 2024*. World Economic Forum. https://www3.weforum.org/docs/WEF_The_Global_Risks_Report_2024.pdf