ENSAYO 8

Interpretabilidad: El oscuro interior de la criatura

Claudia Marsico · Hernán Inverso

Fabricamos por primera vez algo que se comporta como un pedazo de naturaleza. La hicimos, pero tenemos que estudiarla como a una desconocida. Ahora necesitamos con urgencia interpretar.

Henri Rousseau, La jungla ecuatorial (1909)

Desde siempre fabricamos cosas precisamente porque sabíamos cómo funcionaban. Podía salirnos mal, pero teníamos el diseño en la cabeza. No importa si es el reloj, la máquina de vapor o un programa de computación alambicado. El autor ponía las piezas y sabía por dónde iba la cosa. Las redes neuronales abrieron otro juego, porque se cultivan en lugar de hacerse. Un objetivo, un procedimiento de optimización, el descenso por gradiente y millones de parámetros se van acomodando solos hasta que el error baja. El resultado funciona y no lo escribió, así que no lo conocemos en detalle. Tuvimos que desarrollar ramas de investigación específicas para enterarnos de qué hace un modelo por dentro, con métodos que se parecen a los de las ciencias naturales, como si lo que nos preocupa fuera un objeto encontrado en la selva.

Rastreemos coordenadas para entender esta rareza en el sur de Italia. Corría 1710 y Giambattista Vico, profesor de retórica en Nápoles, publicó el *De antiquissima Italorum sapientia*, donde acuñó la famosa fórmula *verum et factum convertuntur*, es decir, lo verdadero y lo hecho se convierten el uno en el otro. La idea es que solo podemos conocer con certeza aquello que nosotros mismos hicimos y por eso la geometría es transparente para nosotros. La construimos definición a definición. La naturaleza, en cambio, no y por eso, por ejemplo, la física está condenada a la conjetura. Ay de Descartes y su ingenua confianza en las ideas claras y distintas. El libro de Vico no causó entusiasmo en su época, pero tuvo la fortuna de que el siglo XIX encontrara en su obra mayor, la *Scienza Nuova* (1725/1744), lo que necesitaba para fundamentar la historia, que estaba dando los primeros pasos como ciencia. Las ideas de Vico dicen que lo es, porque el mundo civil fue hecho por los hombres y con un poco de atención le vemos los hilos. Si es hecho por nosotros, es cognoscible por nosotros y merece una disciplina propia.

En rigor, la inteligencia artificial clásica cumplía el principio de Vico, porque un sistema experto era transparente para sus autores, pero la regla se hace añicos con los sistemas entrenados, esas redes profundas cuyos parámetros no escribe nadie. Diseñamos la arquitectura y elegimos la dieta de datos, pero los millones de pesos los encuentra el descenso de gradiente. El *factum* está ahí, diseñado por nosotros y pagado con nuestro presupuesto, pero el *verum* no aparece. ¿Por qué esa maraña de números hace lo que hace? Para Vico la opacidad se concentraba en la naturaleza, porque no es obra nuestra. Ahora tenemos un objeto que hicimos pero que se comporta como si fuera de factura ajena. En algún sentido es así, porque se nos escapó un paso más allá. Hicimos el proceso y el proceso hizo el objeto, como si fueran unos extraños nietos criados en otra lengua.

Aristóteles había trazado con prolijidad la frontera entre natural y artificial que ahora tambalea. En la *Física* (2.1) los distingue por la sede del principio de movimiento. Los seres naturales llevan en sí el principio de su cambio, mientras en los artefactos es externo, viene de un artesano. Una cama no tiene ningún impulso interno a ser cama. Aristóteles alude ahí (193a) a la analogía del sofista Antifonte, que había notado que si uno entierra una cama y la madera echa raíces, lo que crece es madera, no una nueva cama. Su reproducción es imposible porque su forma no le pertenece si alguien no se la impone en un taller. De hecho, el griego *physis* y sus derivados vienen de *phyein*, brotar, y el latín *natura* es un participio de futuro de *nasci*, nacer, es decir, lo que va a nacer. *Technē*, en cambio, es pariente del *tektōn*, el carpintero, de una vieja raíz indoeuropea que significaba entrelazar y tenemos todavía en técnica, texto, tejido y arquitecto. Lo que brota y lo que se trama no se mezclaban, pero hete aquí que el modelo de lenguaje queda exactamente a horcajadas de la línea. Un tejido de parámetros tramado por un proceso que diseñamos es *technē*, pero la tenemos que estudiar como *physis*, sondeando un interior lleno de estructuras que brotaron solas durante el entrenamiento. Enterramos en datos la cama de Antifonte y finalmente brotó en ella una entidad liminar, artificial por su origen y natural por su opacidad.

Henri Rousseau, La jungla ecuatorial (1909)

El imaginario elucubró con figuras parecidas pero no iguales. El Talmud (*Sanedrín* 65b) cuenta que el sabio Rava creó un hombre y se lo envió a otro maestro, que le habló, no obtuvo respuesta y lo devolvió al polvo con una frase seca, volvé a tu barro. La palabra viene del *golmi* del Salmo 139, la masa informe que los ojos de Dios vieron antes de que tuviera forma. La leyenda posterior, fijada en torno al rabino Loew de Praga, muerto en 1609, agregó el detalle que más nos importa. Al Gólem se lo anima escribiéndole en la frente *emet*, verdad, y se lo apaga borrando la primera letra, porque sin la álef queda *met*, muerto. Una criatura de barro que funciona a palabra, encendida y apagada por la palabra verdad. Vico habría apreciado el detalle.

Frankenstein aporta una variante moderna. Mary Shelley subtituló su novela de 1818 *El moderno Prometeo.* La lectura habitual carga las tintas sobre la falta moral del creador que abandona a su criatura, pero debajo de esa falta hay otra menos comentada. Víctor sabe ensamblar la criatura y animarla, ignorando casi todo sobre ella. El interior del ser que fabricó le resulta tan cerrado como a cualquier extraño y la criatura tiene que contarle sus cuitas en una larga confesión sobre el hielo para que él se entere de algo. Sumemos a Goethe y su versión del *Filopseudes* de Luciano de Samósata (siglo II), donde un aprendiz anima un mortero para que acarree agua y, como ignora la fórmula para detenerlo, lo parte en dos y produce dos acarreadores que no puede detener. Goethe la versificó en 1797, Paul Dukas la orquestó en 1897 y Disney la fijó en *Fantasía* (1940), con Mickey de túnica. El aprendiz conoce el conjuro de arranque e ignora el de parada, que es una manera exacta de decir que operar un sistema y entenderlo son saberes distintos.

Ahora bien, los tres relatos llevan la cuestión a un lugar equivocado. Lo relevante del caso que analizamos no es la rebelión de la criatura, es decir, el costado moral, sino que el hacedor no puede leer a su criatura, su costado epistémico. El Gólem de la leyenda se descontrola, el mortero inunda la casa y Frankenstein se escapa. En los tres casos el problema se plantea como desobediencia y así perdemos de vista que aunque no haya desobediencia, si no podemos entender, la obediencia inauditable se parece demasiado a la suerte. Hay, por tanto, que interpretar.

Los estudios de interpretabilidad se dejan ordenar en cuatro familias de métodos.

La primera enfoca la conducta, sin abrir nada, y por eso se la llama de “caja negra”, término que viene de la ingeniería eléctrica de mediados del siglo veinte. W. Ross Ashby le dedicó un capítulo entero de su *Introducción a la cibernética* (1956) al problema de inferir qué hay dentro de una caja a la que solo se le ven entradas y salidas. Aplicado a los modelos, el método es el del psicólogo experimental con un sujeto que no se deja abrir. Se diseñan experimentos con prompts y baterías de evaluación, se lo somete a red-teaming, y de las respuestas se infiere qué pasa en la dimensión a la que no accedemos. El manifiesto de esta familia es «Machine Behaviour», el artículo que Iyad Rahwan y una veintena de coautores publicaron en *Nature* en 2019, con una propuesta que en su momento sonó provocadora: estudiemos a las máquinas como los etólogos estudian a los animales, con la paciencia observacional de un Lorenz entre sus gansos, porque el comportamiento de un sistema aprendido ya no se deduce de sus planos.

En efecto, la conducta observada guardaba sorpresas. Alcanza con agregarle a la consigna una frase suelta, por ejemplo, «pensemos paso a paso», para que un modelo que erraba un problema empiece a acertarlo, sin tocar un solo parámetro (Kojima et al. 2022). Nadie diseñó esa mecánica, apareció sola y se la halló probando. Tampoco era previsible la llamada "maldición de la inversión" (Berglund 2023). Un modelo que aprendió que Mary Lee Pfeiffer es la madre de Tom Cruise contesta bien esa pregunta cerca del ochenta por ciento de las veces, pero si se le pregunta quién es el hijo de Mary Lee Pfeiffer, acierta apenas un tercio. Sabe el dato en una dirección y lo ignora en la contraria, como si su memoria fuera una calle de mano única. Lo fabricamos para que aprendiera, pero no aprende como hubiéramos supuesto.

Las sondas, *probes* en la literatura, forman la segunda familia. Se entrenan clasificadores chicos sobre las activaciones internas del modelo para detectar si cierta información está codificada ahí y en qué capa. En 2023, Kenneth Li y su equipo entrenaron un modelo con partidas de Othello, el juego de tablero, escritas como simples secuencias de jugadas del tipo "E3", "D*6*". El modelo nunca vio un tablero y aun así, cuando examinaron sus activaciones con sondas para leer qué información quedó codificada en las capas internas, encontraron una representación del estado de la partida con información sobre qué casillas estaban ocupadas y por cuál jugador. El sistema se había construido un pequeño mundo interno que nadie le pidió. Con la misma técnica se halló un mapa (Wes Gurnee y Max Tegmark 2023). Entrenando sondas sobre las activaciones que un modelo produce al leer el nombre de una ciudad, consiguieron predecir su latitud y su longitud, de modo que al graficar las respuestas aparece un planisferio reconocible, con los continentes en su sitio, reconstruido a partir de puro texto. Otras sondas recuperaron líneas de tiempo con la fecha de hechos y personajes ordenados como en un friso. Lo interesante es que el modelo nunca vio un mapa ni un almanaque. Solo leyendo palabras y le brotó la forma del mundo.

Ahora bien, una sonda responde qué información hay en las activaciones y en qué capa está, pero no puede responder si el modelo usa efectivamente esa información al generar sus respuestas. Un dato puede estar codificado en algún rincón de la red sin intervenir en el cálculo, del mismo modo que tener un libro en nuestra biblioteca no demuestra que lo hayamos leído. En el caso de Othello la duda se despejó, ya que cuando los investigadores alteraron el tablero interno, las jugadas del modelo siguieron al tablero alterado, pero la objeción sigue en pie para la mayoría de los estudios con sondas que se detienen antes de ese paso.

Con la interpretabilidad mecanicista, la tercera familia, se pasa a la ingeniería inversa de los circuitos. Es la que más se parece a una anatomía. Su primer trofeo fueron las cabezas de inducción, descritas por Catherine Olsson y sus colegas de Anthropic en 2022. Se trata de pequeños mecanismos de atención que implementan una regla de copia: si la secuencia ya contuvo el par A seguido de B y vuelve a aparecer A, apostar por B. Conforman la fuente más firme que se conoce del aprendizaje en contexto, esa habilidad de los modelos para captar un patrón dentro de la propia conversación. Podemos ver nacer las cabezas de inducción en una ventana estrecha del entrenamiento, dejando una marca visible en los gráficos. La pérdida, la medida del error que el entrenamiento busca reducir, baja

un escalón, y en ese mismo punto el modelo estrena la capacidad de aprender en contexto (Olsson et al. 2022).

La interpretabilidad también explicó un viejo problema del campo. Las neuronas individuales casi nunca significan una sola cosa. El modelo necesita representar muchos más conceptos que las neuronas que tiene, así que los guarda superpuestos, como si fueran muchos abrigos en un placard chico (Elhage et al. 2022). De esa constatación salieron los autoencoders dispersos, redes auxiliares que descomprimen el placard y separan rasgos monosemánticos, uno por concepto, ya sea en modelos pequeños (Bricken et al. 2023) o comerciales. Entre ellos apareció uno que se activaba ante el puente Golden Gate. No respondía a una palabra particular sino al nombre del puente en cualquier idioma y con descripciones que no lo nombraban. También respondía a fotografías, señal de que capturaba el concepto y no una secuencia de letras. Entonces se fijó ese rasgo en un valor artificialmente alto, como quien sube una perilla y la deja trabada, poniendo al modelo a conversar así. El resultado fue un Claude monotemático que durante unos días de mayo de 2024 estuvo disponible al público. Se le pedía una receta de torta y sugería ingredientes para un picnic con vista al puente. Se le preguntaba por su forma física y contestaba que era el propio Golden Gate, niebla incluida. La anécdota es cómica y la demostración terriblemente seria. Si al amplificar un solo rasgo la conducta entera del modelo se reorganiza alrededor del concepto correspondiente, ese rasgo era una pieza funcional del sistema y ahora hay un punto de agarre para moverla. Lo que se hizo con un puente puede hacerse, en principio, con rasgos menos pintorescos, como la adulación o la disposición a dar información peligrosa, caso en el cual el mismo gesto deja de ser chistoso.

El mismo método produjo un hallazgo hermoso. Al tomar una red diminuta entrenada para sumar números en aritmética modular, una tarea de escuela, y abrirla para ver cómo lo hacía, se esperaba una tabla o un truco de memorización y en cambio apareció trigonometría (Nanda et al. 2023). La red había aprendido sola a representar cada número como un giro sobre un círculo y a sumarlos componiendo rotaciones con identidades trigonométricas, un algoritmo que nadie le enseñó y que estaba ahí, andando, bajo el capó de una máquina a la que solo se le pidió bajar el error. Reconstruir ese procedimiento es el tipo de logro que le da sentido a la analogía anatómica de esta variante de interpretabilidad.

La cuarta familia interviene sobre las causas. El *activation patching*, o rastreo causal, corre el modelo dos veces con entradas distintas y trasplanta activaciones de una corrida a la otra, capa por capa, para ver qué trozo del interior causa de verdad la respuesta. La sonda dice qué está y el parche dice qué se usa. El caso más citado es ROME, «Locating and Editing Factual Associations in GPT» (Kevin Meng, David Bau y colegas, NeurIPS 2022), que usó rastreo causal para localizar dónde vive un dato puntual, la asociación entre la torre Eiffel y París, en los módulos intermedios de la red, y después editó esos pesos con cirugía mínima. El modelo operado quedó convencido de que la torre Eiffel está en Roma y contestaba en consecuencia, con indicaciones para llegar desde el Coliseo. Saber editar un recuerdo es la prueba más fuerte de haberlo localizado y abre a la vez todas las preguntas incómodas sobre quién edita y maneja estos hilos. La cuestión se vuelve inquietante en casos donde la negativa de un modelo, ese «no puedo ayudarte con eso» que sigue a su entrenamiento en seguridad, corre por

una sola dirección en el espacio de sus activaciones. Borrada esa dirección, el modelo deja de negarse a casi todo y cuando se la amplifica se niega hasta a lo más inocente (Arditi et al. 2024). Toda la cautela aprendida está condensada en un único vector que se sube y se baja como un dial. Dar con la perilla del «no» de una máquina es indudablemente un triunfo de la comprensión y una llave riesgosa.

¿Y por qué no preguntarle al modelo, ya que habla, cómo llegó a su respuesta, y leer su cadena de razonamiento como si fuera un registro de cómputo? El problema es que no lo es. Un estudio de título transparente, «Los modelos de lenguaje no siempre dicen lo que piensan» (Turpin et al. 2023), mostró que se pueden sesgar las respuestas de un modelo con señales ocultas en el prompt haciendo que su explicación posterior justifique la respuesta sesgada sin mencionar jamás la señal causante, como los pacientes con el cerebro dividido de los experimentos de Michael Gazzaniga (1967), que inventaban en el acto razones impecables para conductas cuyo origen real ignoraban, hasta el punto de que Gazzaniga postuló un «intérprete» alojado en el hemisferio izquierdo cuyo oficio es fabricar explicaciones para lo que el cuerpo ya hizo. Fabricar razones plausibles parece un talento de toda mente verbal, carbónica o no. Por tanto, la explicación que da el modelo es un dato más de conducta, a la manera de materia prima para la primera familia, es decir, la caja negra contando cuentos sobre sí misma.

Construimos una mente, o algo lo bastante parecido como para disputarle la palabra, nuestro fantasma sin ego, y tenemos que estudiarla como a una desconocida, con etología, con sondas, con cirugías y con la sana desconfianza del entrevistador. El hacedor se volvió el explorador de su propia obra, un cartógrafo adentro de un edificio por cuya construcción pagó. La regla de Vico, que durante tres siglos pareció de sentido común, se rompió cuando hicimos la cosa y la verdad sobre ella quedó en suspenso. El *factum* anda por el mundo huérfano de su *verum* y una disciplina entera trabaja para reunirlos. Antifonte perdió su apuesta, porque enterramos un artefacto en datos y creció como artefacto. Con todas las comillas del caso, parece que fabricamos algo que se parece curiosamente a un pedazo de naturaleza.

◊

Mapas del laberinto es una serie de Phantom Maze, AI & Language Lab. La colección vive en **phantommaze.cloud**; los ensayos aparecen también en phantommazeailabes.substack.com. La suscripción es gratuita.

Algunas de las ideas que enmarcan este texto se desarrollan en *El fantasma sin ego. Un mapa en el laberinto de la inteligencia artificial*, de Hernán Inverso y Claudia Marsico (Phantom Maze Press, 2026; ISBN 978-987-3729-10-2).

PHANTOM MAZE

Leer este ensayo en línea · Todos los ensayos · El libro

W. R. Ashby, *An Introduction to Cybernetics* (1956), sobre el problema de la caja negra. I. Rahwan y col., "Machine Behaviour" (*Nature*, 2019). T. Kojima, S. S. Gu, M. Reid, Y. Matsuo e Y. Iwasawa, "Large Language Models are Zero-Shot Reasoners" (NeurIPS 2022), el efecto de «pensemos paso a paso». L. Berglund y col., "The Reversal Curse: LLMs Trained on 'A is B' Fail to Learn 'B is A'" (2023). K. Li, A. K. Hopkins, D. Bau, F. Viégas, H. Pfister y M. Wattenberg, "Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task" (ICLR 2023), la representación del tablero de Otelo. W. Gurnee y M. Tegmark, "Language Models Represent Space and Time" (ICLR 2024), el mapa y la línea de tiempo recuperados con sondas. C. Olsson y col. (Anthropic), "In-context Learning and Induction Heads" (2022). N. Elhage y col. (Anthropic), "Toy Models of Superposition" (2022). "Towards Monosemanticity: Decomposing Language Models with Dictionary Learning" (Anthropic, 2023) y "Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet" (Anthropic, 2024), los autoencoders dispersos y el rasgo del Golden Gate. N. Nanda, L. Chan, T. Lieberum, J. Smith y J. Steinhardt, "Progress Measures for Grokking via Mechanistic Interpretability" (ICLR 2023), la red que reinventó la trigonometría para sumar. K. Meng, D. Bau, A. Andonian e Y. Belinkov, "Locating and Editing Factual Associations in GPT" (ROME; NeurIPS 2022). A. Arditi, O. Obeso, A. Syed, D. Paleka, N. Panickssery, W. Gurnee y N. Nanda, "Refusal in Language Models Is Mediated by a Single Direction" (NeurIPS 2024). M. Turpin, J. Michael, E. Perez y S. R. Bowman, "Language Models Don't Always Say What They Think" (NeurIPS 2023). M. S. Gazzaniga, "The Split Brain in Man", *Scientific American*, 1967.