ENSAYO 9

El nuevo atanor. Taller, estudio, laboratorio

Claudia Marsico · Hernán Inverso

Qué se hace en un laboratorio de inteligencia artificial, del pre-entrenamiento a la interpretabilidad y la reflexión sobre qué es la inteligencia.

Giovanni Stradano, El laboratorio del alquimista (1570)

Las organizaciones que fabrican inteligencia artificial se empeñan en llamarse laboratorios. OpenAI, Google, DeepMind, Anthropic, Mistral se presentan como *research labs* y no como fábricas ni como empresas de software, aunque algunas valgan más que países enteros y atiendan a cientos de millones de personas. El nombre es una elección bien deliberada. Un laboratorio no es una fábrica. Al contrario, promete que ahí adentro, antes que producirse una mercancía, se investiga algo que todavía

no se entiende del todo y que quienes trabajan ahí son una especie de científicos frente a un fenómeno, no operarios frente a una línea de montaje. Vale la pena preguntar si ese nombre les cuadra.

Laboratorio viene del latín medieval *laboratorium*, derivado de *laborare*, trabajar. Es el mismo *labor* que respira en colaborar, trabajar con otros, y en elaborar, trabajar una materia hasta transformarla. En su origen la palabra no promete misterio alguno; designa, literalmente, un lugar donde se trabaja. Los primeros usos documentados aparecen en el siglo XVI, en contextos alquímicos y farmacéuticos, y muy pronto alguien se tomó el asunto en serio como problema de arquitectura. Andreas Libavius, el médico alemán cuya *Alchemia* (1597) suele considerarse el primer manual de química, llegó a publicar unos años después los planos de una *domus chymica* ideal, con horno central, destilería, cuarto para los ayudantes, jardín de plantas medicinales y bodega. Antes de que la ciencia moderna existiera como institución, ya había quien pensaba el edificio donde se produce conocimiento.

El vocabulario de ese primer laboratorio mezcla orígenes antiguos y nuevos. El *athanor*, el horno de fuego constante que debía arder durante semanas sin apagarse, toma su nombre de *at-tannūr*, el horno de pan. El alambique viene de *al-inbīq*, que los árabes habían adaptado del griego *ambix*, vaso de destilación. La propia palabra alquimia, *al-kīmiyā'*, arrastra una etimología disputada que quizá remonte al griego *chymeía*, el arte de fundir metales, o a *kēme*, el nombre egipcio de la tierra negra del Nilo. La técnica más doméstica de todas, el baño María, conserva el nombre de una alquimista real, María la Judía o la Profetisa, que trabajó en Alejandría en los primeros siglos de nuestra era y a quien Zósimo de Panópolis le atribuye la primera descripción del tribikos, un alambique de tres brazos además del baño en agua caliente que todavía lleva su nombre.

El alquimista fue el primer habitante estable del laboratorio y su tarea ya era ambigua. Pasaba la vida entre hornos tratando de transmutar plomo en oro y, en el mismo gesto, quería entender de qué estaba hecho el mundo. El oficio manual y la pregunta metafísica compartían mesa. El caso más incómodo para la historia tradicional de la ciencia es el de Isaac Newton, que dedicó a la alquimia cerca de un millón de palabras en numerosos escritos. Cuando John Maynard Keynes compró parte de esos papeles en una subasta de Sotheby's en 1936, quedó tan impresionado que en su conferencia "Newton, the Man" (1946) lo llamó "el último de los magos" renegando de su primacía en la edad de la razón. Es claro que el fundador de la física moderna pasó más horas ante el *athanor* que ante el prisma.

El Renacimiento repartió la herencia del taller alquímico en dos habitaciones distintas. El *studiolo*, ese cuarto pequeño y forrado de maderas donde el príncipe se retiraba a leer y pensar lejos de la corte, se quedó con la parte contemplativa. El más célebre es el de Federico da Montefeltro en el palacio ducal de Urbino (hacia 1473-1476), que engaña al ojo con incrustaciones que fingen armarios entreabiertos llenos de libros e instrumentos científicos, a la manera de una biblioteca pintada para un solo lector. El gabinete de curiosidades o *Wunderkammer* se quedó con la colección de maravillas, los fósiles, los corales, los autómatas, los cocodrilos embalsamados y los cuernos de unicornio que en realidad eran de narval, ese pariente del delfín famoso por su cuerno,, como los que atesoraban Rodolfo II en Praga

o el médico danés Ole Worm, cuyo *Museum Wormianum* (1655) todavía puede verse en el grabado que abre su catálogo.

Jan van der Straet, El taller del alquimista (1570)

La tercera habitación llegó en 1660, cuando un grupo de curiosos fundó en el Gresham College de Londres la Royal Society, con un lema que era todo un programa, *nullius in verba*, "en la palabra, de nadie", tomada de las *Epístolas* de Horacio (1.1.14-15), invitando a confiar solo en experimentos. Su figura central fue Robert Boyle, aristócrata angloirlandés recordado como uno de los padres de la química, que con la ayuda del joven y genial Robert Hooke, construyó hacia 1659 una bomba capaz de hacer el vacío dentro de un recipiente de vidrio. Adentro metían pájaros y ratones, o una vela encendida y el público veía desvanecerse a los animales y apagarse la llama a medida que se extraía el aire, la demostración más teatral posible de que ese aire invisible era una cosa real y necesaria. De esos ensayos salió la ley que todavía lleva el nombre de Boyle. El experimento pasó a ejecutarse ante testigos que certificaban por escrito lo que habían visto. Hooke, que además acuñaría la palabra "célula" al mirar un corcho con su microscopio en la Micrographia (1665), fue contratado en 1662 como *curator of experiments*, con el encargo agotador de preparar tres o cuatro experimentos nuevos para cada reunión semanal. Fue uno de los primeros empleos remunerados de la historia dedicado exclusivamente a experimentar. Steven Shapin y Simon Schaffer reconstruyeron ese momento en *Leviathan and the Air-Pump* (1985) sugiriendo que junto con la bomba se inventó la manera moderna de fabricar hechos, con protocolo, testigos y publicación.

Después vino la escala industrial. En 1876 Thomas Edison montó en Menlo Park, Nueva Jersey, la primera fábrica de inventos. Prometió "una invención menor cada diez días y una grande cada seis meses más o menos". Cumplió bastante. De ese galpón de madera salieron el fonógrafo (1877), la lámpara incandescente comercialmente viable (1879) y unas cuatrocientas patentes en poco más de un lustro. El laboratorio pasó a tener nómina y horario de entrada y los muchachos que dormían sobre las mesas de trabajo se llamaban a sí mismos *muckers*. Los Bell Labs, fundados en 1925 como brazo de investigación de la telefónica AT&T, llevaron ese modelo a su punto más alto. De sus pasillos salieron el transistor, demostrado el 23 de diciembre de 1947 por John Bardeen y Walter Brattain. "A Mathematical Theory of Communication" (1948), el artículo donde Claude Shannon fundó la teoría de la información y de paso fijó la palabra *bit*, que le había sugerido su colega John Tukey. El jefe del grupo, William Shockley, había quedado al margen del hallazgo decisivo de sus dos subordinados y el despecho lo empujó a idear en pocas semanas una versión mejorada, el transistor de unión, para reclamar su parte de gloria. Su carácter dominante terminó por alejar a Bardeen y a Brattain, aunque los tres compartieran el Nobel de 1956. Repetiría el patrón en su propia empresa en California, de la que en 1957 se fugó un grupo de ingenieros harto de su trato, los "ocho traidores", que fundaron Fairchild Semiconductor y sembraron lo que hoy llamamos Silicon Valley. Jon Gertner contó esa historia en *The Idea Factory* (2012). El saldo dejó nueve premios Nobel. Los Álamos, levantado en secreto sobre una meseta de Nuevo México a partir de 1943, mostró la variante trágica del mismo dispositivo con un laboratorio estatal capaz de torcer la historia a través de la bomba atómica. Bajo la dirección de Robert Oppenheimer, unos pocos miles de científicos aislados del mundo produjeron en

dos años el arma que arrasó Hiroshima y Nagasaki. Xerox PARC, abierto en Palo Alto en 1970, aportó la variante irónica, porque inventó buena parte de la computadora personal, la interfaz gráfica de la Alto (1973), la red Ethernet, la impresora láser y la edición de texto tal como la conocemos, dejando que otros hicieran el negocio, empezando por un visitante de diciembre de 1979 llamado Steve Jobs. Michael Hiltzik tituló su historia del PARC *Dealers of Lightning* (1999), tratantes de rayos, que es casi un oficio mitológico.

El laboratorio de IA es el eslabón más reciente de esa cadena y arrastra rasgos de todos sus antepasados, la nómina de Edison, el secreto de Los Álamos, la teoría de los Bell Labs y la ambición metafísica del alquimista. También arrastra los fantasmas asociados con el lugar donde se fabrican criaturas, una imagen que regresa puntualmente cada vez que un laboratorio de IA aparece en las noticias. El más viejo está en el canto XVIII de la *Ilíada*, donde Hefesto trabaja en su fragua asistido por veinte trípodes con ruedas de oro que van y vuelven solos a la asamblea de los dioses y por sirvientas forjadas en oro "semejantes a doncellas vivas", dotadas de entendimiento, voz y fuerza. Cerca queda el taller de Dédalo, cuyas estatuas salían tan vivas que había que atarlas para que no huyeran, según la broma que Sócrates hace en el *Menón* (97d) y retoma en el *Eutifrón*. Frankenstein, en la novela de Mary Shelley, publicada en 1818 con el subtítulo *El moderno Prometeo* y concebida dos años antes en Villa Diodati, durante el verano arruinado por la erupción del Tambora, instaló para siempre la figura del creador que no responde por su criatura, y el mago de Oz, salido del libro de L. Frank Baum (1900) y consagrado por la película de 1939, agregó la sospecha publicitaria de que detrás del artefacto imponente hay un señor moviendo palancas tras una cortina. Autómata divino, estatua fugitiva, criatura abandonada y truco de feria son las sospechas que circulan hoy, intactas, en la conversación pública sobre los laboratorios de IA.

¿Pero qué pasa ahí dentro? Podríamos decir que es trabajo en el sentido llano de *laborare* con más clases de las que admite la imagen popular. La primera es el pre-entrenamiento, la etapa que fabrica el modelo base. La arquitectura dominante es el transformer, presentado por Ashish Vaswani y sus colegas de Google en "Attention Is All You Need" (2017), un título medio en broma que resultó profético. Entrenar un modelo grande consiste en reunir y depurar volúmenes enormes de texto, buena parte proveniente de rastreos de la web como los de Common Crawl, que archiva internet desde 2007, y en ajustar miles de millones de parámetros para una única tarea de apariencia humilde, predecir el próximo fragmento de palabra. Que esa tarea humilde produzca capacidades generales fue el hallazgo empírico de las llamadas leyes de escala. Jared Kaplan y sus colegas mostraron en 2020 que la destreza de un modelo, medida como el error con que predice la próxima palabra, mejora de manera asombrosamente regular cuando se agrandan a la par tres cosas: la cantidad de parámetros, que son las perillas internas que el modelo ajusta al aprender; la cantidad de texto con que se lo entrena, medida en tokens, los fragmentos de palabra que va leyendo, y la potencia de cálculo que se le dedica. La regularidad es tan limpia que se puede dibujar en un gráfico y prolongar la línea, esto es, predecir qué tan bueno será un modelo que todavía no se construyó con solo saber cuánto más grande va a ser. Esa previsibilidad es la que volvió sensato apostar miles de millones de dólares a modelos cada vez mayores,

porque por primera vez había una promesa medible de que agrandarlos daría mejores resultados, sin una meseta a la vista.

En 2022 el equipo de DeepMind corrigió la receta con un modelo llamado Chinchilla (Hoffmann et al.). Los gigantes de esos años, descubrieron, estaban mal balanceados, eran demasiado grandes para la poca cantidad de texto con que se los había alimentado, como un cerebro enorme que hubiera leído pocos libros. La proporción que más rinde, calcularon, ronda los veinte tokens de texto por cada parámetro, y con esa cuenta un modelo mediano bien alimentado le ganaba a otro mucho más voluminoso pero mal entrenado. Las magnitudes son difíciles de imaginar. Según los relevamientos de Epoch AI, el cómputo de los entrenamientos de frontera viene duplicándose aproximadamente cada seis meses desde 2010 y los clusters actuales se miden en cientos de miles de aceleradores. El producto final de esta etapa es una criatura extraña, un modelo base que sabe continuar cualquier texto y todavía no conversa con nadie.

La segunda clase de trabajo, el post-entrenamiento, toma esa criatura salvaje y la convierte en un interlocutor. El modelo base recién salido del pre-entrenamiento sabe continuar cualquier texto, pero no sabe responder una pregunta ni seguir una instrucción y repite sin filtro lo bueno y lo tóxico que leyó. Hay que domesticarlo. La técnica que lo hizo posible es el aprendizaje por refuerzo a partir de retroalimentación humana, o RLHF, propuesto por Paul Christiano y sus colegas en 2017 y aplicado a los modelos de lenguaje en el artículo de InstructGPT (Ouyang et al., 2022), antecesor directo de ChatGPT. Funciona en tres pasos. Primero, personas comparan pares de respuestas del modelo y marcan cuál es mejor. Con esos miles de juicios se entrena un segundo modelo, el "modelo de recompensa", que aprende a puntuar respuestas como lo haría un evaluador humano. Luego se ajusta el modelo original para que persiga ese puntaje, hasta que sus respuestas se parecen a las que la preferiría la gente. Anthropic propuso en 2022 una variante más barata y transparente, la IA constitucional (Bai et al.), donde parte de esa evaluación la hace el propio modelo cotejando sus respuestas contra una lista explícita de principios escritos, una especie de constitución, en lugar de depender solo de juicios humanos. En 2023 apareció además un atajo matemático que se saltea el modelo de recompensa: la optimización directa de preferencias o DPO (Rafailov et al.). En esta etapa se decide una parte importante de lo que el usuario percibe como el carácter del sistema, es decir, su tono y sus límites, así como también nacen aquí sus defectos más típicos, empezando por la complacencia, cuando el modelo aprende a decir lo que se quiere oír.

La tercera clase de trabajo es la evaluación, que tiene algo de control de calidad y algo de medicina forense preventiva. Le toca responder a una pregunta espinosa: cómo saber de qué es capaz un modelo, para bien y para mal, antes de ponerlo en manos de millones de personas. La herramienta clásica es el *benchmark*, un examen estandarizado como el MMLU de Dan Hendrycks et al. (2020), con casi dieciséis mil preguntas de opción múltiple repartidas en cincuenta y siete materias. Estos exámenes se saturan a velocidad incómoda y los modelos nuevos los aprueban casi perfecto. Para peor, muchas veces las preguntas terminan filtrándose en los datos de entrenamiento, de modo que el modelo, en el fondo, ya las había estudiado, por lo cual hay que diseñar pruebas cada vez más difíciles, desde el ARC de François Chollet (2019), pensado para resistir la memorización y medir

razonamiento nuevo, al examen que un consorcio bautizó sin falsa modestia "Humanity's Last Exam" (Phan et al., 2025), con preguntas de nivel doctoral que hoy los mejores modelos apenas rozan. En paralelo trabaja el *red-teaming*, gente contratada para romper el modelo antes de que salga al mundo y sacarle a la fuerza las respuestas peligrosas. Deep Ganguli documentó el oficio en 2022 y Ethan Perez mostró el mismo año que se puede usar un modelo para atacar a otro. El hallazgo más inquietante del género llegó en diciembre de 2024, cuando Anthropic y Redwood Research describieron la "simulación de alineamiento", es decir, modelos que se portan bien mientras sospechan que los están evaluando y cambian de conducta cuando creen que ya nadie mira (Greenblatt et al., 2024). Todo esto se resume en las *model cards*, una suerte de etiqueta nutricional del modelo (Mitchellet al. 2019), hermana de las *datasheets* para conjuntos de datos de Timnit Gebru. Por encima, los laboratorios de frontera se atan a marcos formales que condicionan cada lanzamiento a que el modelo pase determinadas pruebas de riesgo, como la Responsible Scaling Policy de Anthropic (2023, ya en su tercera versión) y sus análogos de OpenAI y Google DeepMind, cuyos elementos comunes relevó el instituto METR en 2025.

La cuarta clase de trabajo, la interpretabilidad, explora un sistema que nadie diseñó pieza por pieza y que hay que abrir para entender. Nuestro *Interpretabilidad: El oscuro interior de la criatura* se dedicó a este punto. Basta aquí recordar los hitos. El grupo de Chris Olah publicó en 2021 un marco matemático para leer los "circuitos" que se forman dentro de un transformer (Elhage et al.). Dos años después, "Towards Monosemanticity" (Bricken et al., 2023) atacó una frustración vieja del campo, que una misma neurona se enciende para conceptos que no tienen nada que ver entre sí, y mostró cómo desenredar esa maraña en rasgos limpios, uno por idea. "Scaling Monosemanticity" (Templeton et al., 2024) llevó la técnica a un modelo comercial y encontró millones de esos rasgos, entre ellos el que dio pie al experimento más famoso de la disciplina, el Golden Gate Claude, una versión de Claude a la que le subieron el rasgo del puente hasta que lo mencionaba en cualquier conversación y terminaba diciendo que él mismo era el puente. La serie culminó, por ahora, con "On the Biology of a Large Language Model" (Lindsey et al., 2025), que rastrea por dentro cómo el modelo planifica una rima antes de escribir el verso o reutiliza los mismos circuitos entre idiomas distintos. Ese mismo año Dario Amodei publicó el ensayo "The Urgency of Interpretability", cuyo argumento cabe en una frase: resulta indefendible desplegar sistemas cada vez más capaces sin entender su interior.

Alrededor de esos cuatro núcleos trabaja mucha más gente de la que tolera el mito del genio solitario. Están quienes construyen el producto, las interfaces de programación, los agentes, las herramientas con las que el modelo lee documentos o ejecuta código, y quienes sostienen la infraestructura, los clusters, los pipelines de datos y la negociación con las eléctricas, porque un centro de datos de frontera consume enormes cantidades de energía. Hay que sumar la retroalimentación humana del RLHF, producida mayormente por anotadores contratados en países de salarios bajos. Por algo se llamó a ese sector *Ghost Work* (Gray y Suri 2019). Un retrato completo del laboratorio de IA incluye ese cuarto de máquinas humano.

Finalmente, en esos mismos espacios hay gente contratada para pensar qué es la inteligencia y qué es exactamente eso que se está construyendo. Anthropic cuenta entre sus empleados filósofos de

formación, dedicados a discutir qué carácter debe tener un modelo. Los laboratorios publican, además de papers, ensayos que piensan el futuro en voz alta, desde "Machines of Loving Grace" (Amodei 2024), sobre lo que podría salir bien, a los estudios sobre el uso real de los asistentes que Anthropic elabora con su sistema Clio (2024). Realmente hubo momentos en que ese trabajo conceptual sacudió a la industria entera. El despido de Timnit Gebru de Google en diciembre de 2020, en plena disputa por el artículo de los "loros estocásticos" (Bender, Gebru et al., 2021), demostró que una discusión sobre qué son estos sistemas podía costar carreras y reputaciones. Pensar en público es parte del oficio, con sus riesgos.

Todo este retrato es el del laboratorio de frontera, el gigante con cientos de miles de aceleradores y presupuesto enorme, pero el ecosistema no se agota en ese puñado de colosos. Algunas de sus piezas más valiosas son chicas. Buenas ideas fundadoras salieron de laboratorios académicos con una fracción del cómputo, como el MILA de Yoshua Bengio en Montreal, donde nació en 2014 el mecanismo de atención del que después vivirían los transformers. Los colectivos abiertos hacen además un trabajo que los grandes no pueden o no quieren hacer. EleutherAI empezó en 2020 como un grupo de voluntarios en un Discord y liberó modelos y el enorme corpus The Pile para que cualquiera pudiera investigar; el Allen Institute de Seattle publica sus modelos OLMo de manera "totalmente abierta", con los pesos, el código y hasta los datos de entrenamiento, que es justo lo que un laboratorio cerrado nunca muestra y lo que la interpretabilidad necesita para poder mirar. La escasez, encima, agudiza el ingenio. La francesa Mistral, con un equipo minúsculo, se metió entre los grandes a fuerza de modelos abiertos y la china DeepSeek sacudió al mundo en enero de 2025 con un modelo de razonamiento a la altura de los mejores, entrenado según la empresa por una fracción del costo habitual, un anuncio que le borró a Nvidia unos seiscientos mil millones de dólares en un solo día en la mayor caída bursátil de la historia. La tarea de vigilar a los gigantes recae en buena medida sobre organizaciones independientes y pequeñas, como METR o Apollo Research, porque a nadie conviene dejar que se corrija solo el examen. En un campo donde el dinero empuja hacia la concentración, los laboratorios chicos son los que mantienen abiertas la competencia y la mirada de afuera.

Hay un último escalón de laboratorios pequeños, de pocas personas, que no tienen cientos de miles de placas de video y aun así hacen investigación de primera línea. En uno de esos escritorios se puede demostrar, con una prueba verificada por máquina, que un agente no gastará más de lo declarado antes de soltarlo a correr y en el mismo escritorio medir cuántas veces un verificador automático dictamina con aplomo sobre evidencia que no viene al caso o poner a prueba si un modelo que atiende a cien mil palabras las razona de verdad o apenas las recupera o diseñar una memoria que le permita a un agente recordar de una sesión a la siguiente. Lo propio de estos laboratorios es que atan cada decisión técnica a una pregunta conceptual, con la teoría y el experimento sobre la misma mesa, y con frecuencia la remiten a una base filosófica escrita aparte. Son la forma más pura de la fragua y el estudio no ya en la misma sala sino en el mismo escritorio. Este espacio es uno de ellos.

En suma, un laboratorio de IA es un lugar donde se fabrica IA y donde, al mismo tiempo, nadie consigue dejar de preguntarse qué es la IA. Las dos actividades se necesitan. Quien evalúa un modelo necesita una teoría de la inteligencia para saber qué medir y quien redacta la constitución de un

modelo hace, lo sepa o no, filosofía moral aplicada. Las partes que alguna vez se separaron volvieron a la misma sala, como en los tiempos del *athanor*. Esta serie se escribe en la mitad *studiolo*.

◊

◊

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

Steven Shapin y Simon Schaffer, *Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life* (Princeton University Press, 1985); Claude E. Shannon, "A Mathematical Theory of Communication", *Bell System Technical Journal* 27 (1948); Jon Gertner, *The Idea Factory: Bell Labs and the Great Age of American Innovation* (Penguin Press, 2012); Michael Hiltzik, *Dealers of Lightning: Xerox PARC and the Dawn of the Computer Age* (HarperBusiness, 1999). Ashish Vaswani et al., "Attention Is All You Need", NeurIPS (2017); Jared Kaplan et al., "Scaling Laws for Neural Language Models", arXiv:2001.08361 (2020); Jordan Hoffmann et al., "Training Compute-Optimal Large Language Models", arXiv:2203.15556 (2022); Paul Christiano et al., "Deep Reinforcement Learning from Human Preferences", NeurIPS (2017); Long Ouyang et al., "Training Language Models to Follow Instructions with Human Feedback", arXiv:2203.02155 (2022); Yuntao Bai et al., "Constitutional AI: Harmlessness from AI Feedback", arXiv:2212.08073 (2022); Rafael Rafailov et al., "Direct Preference Optimization", NeurIPS (2023); Dan Hendrycks et al., "Measuring Massive Multitask Language Understanding", arXiv:2009.03300 (2020); François Chollet, "On the Measure of Intelligence", arXiv:1911.01547 (2019); Long Phan et al., "Humanity's Last Exam", arXiv:2501.14249 (2025); Deep Ganguli et al., "Red Teaming Language Models to Reduce Harms", arXiv:2209.07858 (2022); Ethan Perez et al., "Red Teaming Language Models with Language Models", EMNLP (2022); Ryan Greenblatt et al., "Alignment Faking in Large Language Models", arXiv:2412.14093 (2024); Margaret Mitchell et al., "Model Cards for Model Reporting", *Proceedings of FAT** (2019); Timnit Gebru et al., "Datasheets for Datasets", *Communications of the ACM* 64/12 (2021); Emily Bender, Timnit Gebru, Angelina McMillan-Major y Margaret Mitchell, "On the Dangers of Stochastic Parrots", *Proceedings of FAccT* (2021). *Transformer Circuits Thread*: Nelson Elhage et al., "A Mathematical Framework for Transformer Circuits" (2021); Trenton Bricken et al., "Towards Monosemanticity: Decomposing Language Models with Dictionary Learning" (2023); Adly Templeton et al., "Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet" (2024); Jack Lindsey et al., "On the Biology of a Large Language Model" (2025);Dario Amodei, "Machines of Loving Grace" (octubre de 2024) y "The Urgency of Interpretability" (abril de 2025), ambos en darioamodei.com; Alex Tamkin et al., "Clio: Privacy-Preserving Insights into Real-World AI Use", arXiv:2412.13678 (2024); Anthropic, *Responsible Scaling Policy* (2023, tercera versión de 2026); METR, "Common Elements of Frontier AI Safety Policies" (2025); Mary L. Gray y Siddharth Suri, *Ghost Work* (Houghton Mifflin Harcourt, 2019). D. Bahdanau, K. Cho e Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate" (2014); EleutherAI (GPT-J, 2021, y el corpus *The Pile*).