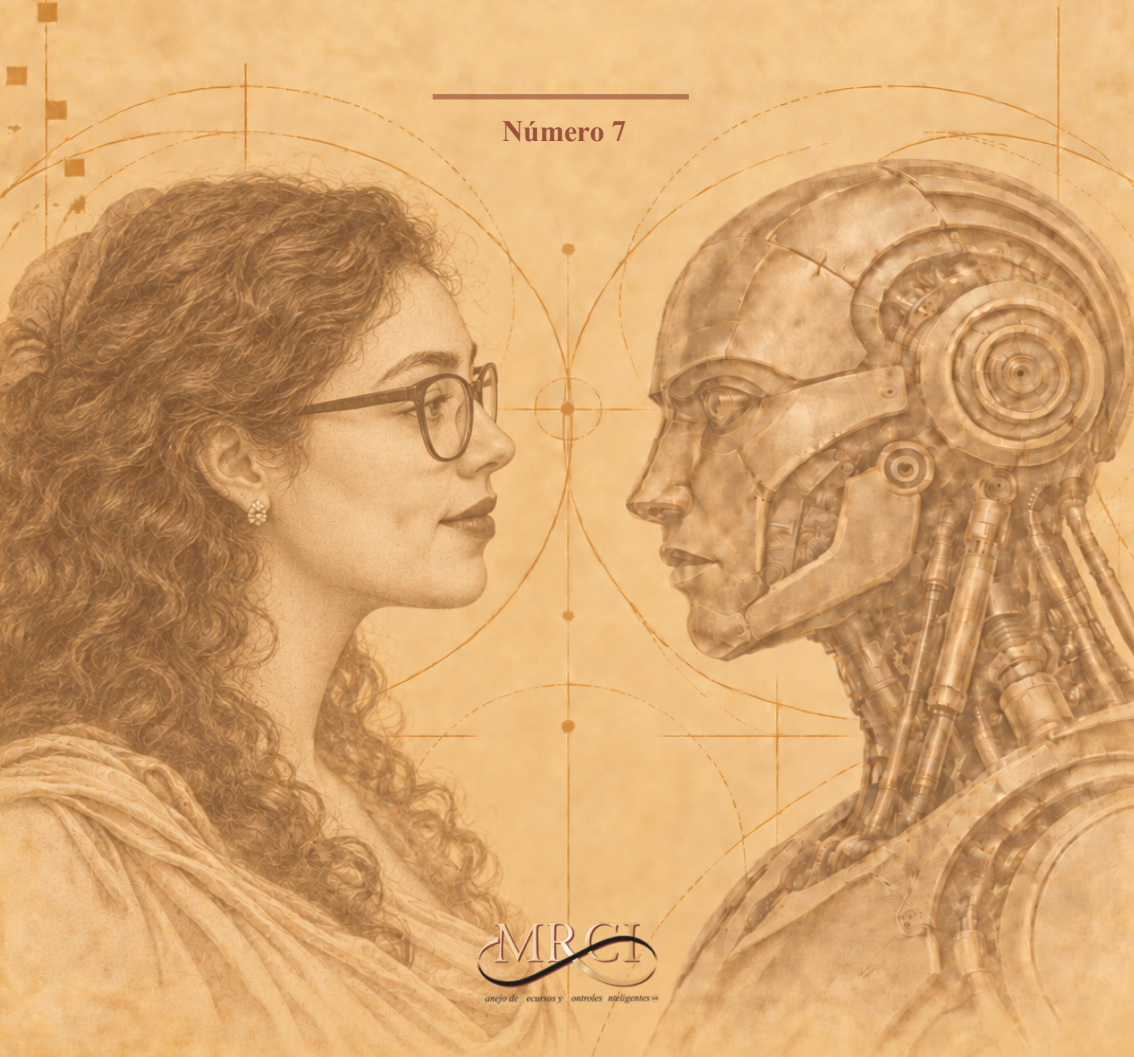
COLECCIÓN

MITOS Y REALIDADES DE LA IA

Entre la promesa y la realidad

REDES NEURONALES Y LOS LÍMITES DE SU COMPARACIÓN CON EL CEREBRO

Número 7



MRCI

anejo de ecursos y ontroles ntéligentes

DR. JOSÉ ANTONIO PÉREZ RAMOS

REDES NEURONALES Y LOS LÍMITES DE SU COMPARACIÓN CON EL CEREBRO

SOBRE EL AUTOR

Doctor en Ciencias de lo Fiscal por el Instituto de Especialización para Ejecutivos. Maestro en Fiscal por la Universidad Autónoma "Benito Juárez" de Oaxaca. Maestro en Estudios Avanzados en Derechos Humanos por la Global Open University. Licenciado en Contaduría Pública por la Universidad Autónoma "Benito Juárez" de Oaxaca y Licenciado en Filosofía por la Universidad Nacionalista de México.

Socio fundador y director general de la Firma Manejo de Recursos y Controles Inteligentes (Firma MRCI). Fiscalista del año 2009 por la revista *Defensa Fiscal*. Doctor honoris causa por la Asamblea General del Claustro Doctoral Iberoamericano (2016). Doctor honoris causa por el Instituto Mexicano de Líderes de Excelencia (2017). Doctor honoris causa por 1Million Startups, Latinomics, Leaderships Forum y la Fundación Humanist World (2018). Doctor honoris causa por el Claustro de Doctores de la AISOL (American International School of Law) (2019).

Autor de diversas obras, entre las que destacan *Ius mutantur o del derecho que cambia* (Tirant Lo Blanch, 2025), *Futurología* (MRCI, 2025) y *Hasta que la muerte nos separe* (MRCI, 2025). Ha participado como coordinador en *Cuestiones tributarias. Problemas y controversias en el México actual* (Tirant Lo Blanch, 2023), *El costo de la justicia* (Apex Iuris, 2019), *Remuneraciones estratégicas inteligentes* (MRCI, 2015) y *Retos de la justicia tributaria* (Tirant Lo Blanch-AMDF, 2025).

Titular de *Pie de Página*, programa para la divulgación de los derechos humanos en México, a través de la plataforma Legal Cloud. Expositor en foros nacionales e internacionales. Colaborador en diversas revistas jurídicas, entre las que destacan *Defensa Fiscal* y *Legal Cloud*.

REDES NEURONALES Y LOS LÍMITES DE SU COMPARACIÓN CON EL CEREBRO

DR. JOSÉ ANTONIO PÉREZ RAMOS

PRIMERA EDICIÓN, ATEMPORAL

Derechos reservados, propiedad de
Dr. José Antonio Pérez Ramos

Comentarios y opiniones: investigacion@mrci.com.mx

Título original: Redes Neuronales y los límites de su comparación con el cerebro.

Colección mitos y realidades de la inteligencia artificial.

Entre la promesa y la realidad.

Autor: Dr. José Antonio Pérez Ramos.

Queda prohibida la reproducción total y parcial de esta obra denominada REDES NEURONALES Y LOS LÍMITES DE SU COMPARACION CON EL CEREBRO, por cualquier medio, sin autorización escrita del autor.

PRINTED IN MEXICO
IMPRESO EN MÉXICO

ÍNDICE

LA PREGUNTA QUE EL NOMBRE NO RESUELVE ..... - 7 -

MÉTODO Y LÍMITES DE ESTA REVISIÓN ..... - 8 -

QUÉ HACE UNA UNIDAD ARTIFICIAL ..... - 10 -

POR QUÉ CONECTAR UNIDADES CAMBIA EL PROBLEMA ..... - 13 -

APRENDER EXIGE UN CRITERIO DE MEJORA ..... - 17 -

PARÁMETROS Y REPRESENTACIONES NO SON LO MISMO ..... - 20 -

LO QUE UNA COMPARACIÓN BIOLÓGICA DEBE ESPECIFICAR ..... - 23 -

EL PROBLEMA DEL APRENDIZAJE BIOLÓGICAMENTE PLAUSIBLE . - 25 -

CUATRO PREGUNTAS QUE NECESITAN PRUEBAS DISTINTAS ..... - 28 -

UNA RED PUEDE CONTRIBUIR A ESTUDIAR LA VISIÓN ..... - 31 -

PREDECIR UNA REGIÓN NO REPRODUCE SUS DEPENDENCIAS ... - 34 -

CONDUCTA PARECIDA Y SEÑALES DIFERENTES ..... - 36 -

CÓMO LEER UNA CIFRA DE SEMEJANZA ..... - 40 -

DISEÑAR UNA COMPARACIÓN QUE PUEDA FALLAR ..... - 43 -

MITOS Y OBJECIONES QUE CONVIENE SEPARAR ..... - 46 -

IMPLICACIONES PARA ENSEÑAR Y UTILIZAR REDES ..... - 48 -

CONCLUSIÓN ..... - 51 -

REFERENCIAS ..... - 52 -

La pregunta que el nombre no resuelve

Un dibujo de círculos conectados puede representar una red artificial, un circuito biológico o una estructura completamente ajena a ambos; la forma gráfica ayuda a seguir relaciones, pero no identifica por sí misma qué hacen las conexiones, cómo cambian o a qué corresponde cada círculo en el mundo, de modo que el parecido visual proporciona un punto de partida pedagógico y no una conclusión científica.

La pregunta central es en qué sentido una red neuronal artificial puede compararse con el cerebro y qué pruebas serían necesarias para sostener una correspondencia concreta; responder exige abandonar tanto la afirmación de que una red es un pequeño cerebro como la idea de que cualquier comparación carece de valor por tratarse de un programa.

Una calculadora puede producir la misma respuesta aritmética que una persona mediante un procedimiento distinto, mientras un modelo de una corriente nerviosa puede representar con detalle una variable biológica sin realizar ninguna tarea cotidiana que reconoceríamos como inteligente; estos dos ejemplos conceptuales muestran

que parecido funcional y fidelidad biológica son dimensiones diferentes.

La tesis de este trabajo es que una analogía neuronal debe formularse como una hipótesis delimitada, con un objeto, un nivel de descripción y una prueba; si se afirma que una red predice ciertas respuestas visuales, corresponde examinar esas predicciones, pero si se afirma que reproduce el circuito que las genera, hacen falta observaciones capaces de discriminar mecanismos alternativos.

La distinción también protege la evaluación práctica, porque un sistema de clasificación puede resultar útil sin reproducir la visión humana y puede resultar inadecuado aunque algunos de sus patrones internos se parezcan a mediciones cerebrales; la decisión de usarlo depende de la tarea, los errores, las condiciones de operación y sus consecuencias.

Método y límites de esta revisión

Se realizó una revisión documental crítica con fecha de corte del 3 de octubre de 2026, a las 00:12 UTC (2 de

octubre en Oaxaca), concentrada en seis investigaciones primarias que permiten contrastar cálculo artificial, modelación biofísica, aprendizaje y comparación visual; se consultaron los textos originales mediante publicaciones, repositorios de autores y archivos científicos, distinguiendo los resultados de cada estudio de la interpretación desarrollada aquí.

La selección es intencional y no constituye una revisión sistemática de toda la neurociencia computacional; permite examinar tipos de inferencia, pero no calcular un balance estadístico de todas las arquitecturas, ni establecer cuál es el modelo cerebral más completo disponible, ni atribuir a una investigación histórica el estado definitivo del campo.

Los cálculos, tablas comparativas y casos numéricos que siguen son construcciones explicativas, identificadas como tales; no se entrenó una red experimental ni se midieron cerebros para este artículo, y los resultados publicados pertenecen exclusivamente a los autores que se citan.

El alcance se concentra en los límites de la comparación neuronal, mientras el detalle de transformers corresponde

a otro artículo de la colección y la conciencia artificial requiere una discusión independiente; tampoco se presupone que toda inteligencia artificial utilice redes neuronales, pues aquí se estudia una familia de modelos y no una definición universal de la disciplina.

Qué hace una unidad artificial

Para comenzar, conviene utilizar una unidad definida por operaciones que cualquiera pueda reconstruir: recibe entradas numéricas, multiplica cada una por un peso, agrega un sesgo y aplica una función a la suma; esta descripción identifica una transformación matemática y no afirma que una célula nerviosa ejecute exactamente esa secuencia.

En el ejemplo de dos entradas se escribirá z = w1 × x1 + w2 × x2 + b, donde x1 y x2 son los datos de entrada, w1 y w2 los pesos, b el sesgo y z el resultado previo a la activación; la salida será a = máximo entre 0 y z, una operación conocida como unidad lineal rectificada o ReLU.

El peso determina cómo contribuye una entrada a esa combinación, mientras el sesgo desplaza el punto a partir



del cual la salida puede hacerse positiva; si una entrada representa una cantidad medida en cierta escala, cambiar esa escala sin ajustar lo demás también altera el resultado, de modo que el valor de un peso aislado carece de una interpretación universal.

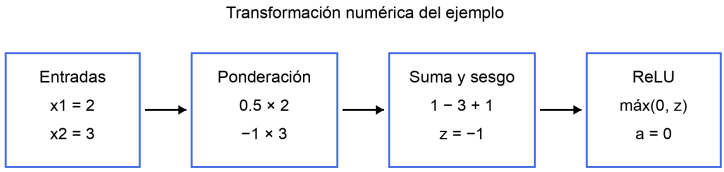
Tomemos x1 = 2, x2 = 3, w1 = 0.5, w2 = −1 y b = 1; la combinación es 0.5 × 2 − 1 × 3 + 1 = −1, por lo que la salida rectificada es 0, un resultado que puede comprobarse sin mencionar imágenes, lenguaje, intenciones o estados mentales.

Si el segundo peso cambia a −0.2 y el resto permanece igual, la combinación pasa a 0.5 × 2 − 0.2 × 3 + 1 = 1.4 y la salida es 1.4; este cambio permite observar sensibilidad a un parámetro, pero todavía no demuestra aprendizaje, porque no se especificó qué salida era deseable ni cómo se eligió la modificación.

La palabra activación puede generar otra confusión: en este ejemplo designa un número calculado, no un disparo eléctrico registrado en una célula; sería posible construir otro modelo con eventos discretos o variables que evolucionan en el tiempo, aunque esa posibilidad no

convierte automáticamente la operación actual en un modelo de todos esos fenómenos.

Figura 1. *Una unidad artificial definida por sus operaciones*



Nota. Elaboración propia; las flechas indican dependencia de cálculo y los valores pertenecen al ejemplo hipotético, no a una neurona biológica medida.

Lo que el diagrama deja fuera es tan importante como lo que muestra: no especifica moléculas, anatomía, tiempos de conducción ni historia de desarrollo; estas omisiones son legítimas para enseñar una función, pero se volverían problemáticas si alguien utilizara el mismo dibujo para afirmar que ya describió el funcionamiento de una neurona humana.

El nombre de la unidad tampoco obliga a interpretarla como un concepto único, porque una salida de 1.4 sólo



adquiere significado respecto de su posición en un sistema; podría alimentar otra operación, contribuir a una clasificación o formar parte de una representación, y ninguna de esas posibilidades se deduce únicamente del número.

Por qué conectar unidades cambia el problema

Una red se forma al organizar dependencias entre unidades, de modo que los resultados de algunas se convierten en entradas de otras; una capa es una agrupación de transformaciones dentro de esa organización y no debe identificarse automáticamente con una región anatómica, pues el término describe aquí una estructura computacional.

La composición permite construir funciones que una unidad aislada no expresa bajo las mismas restricciones; para observarlo sin depender de un sistema comercial, consideremos dos entradas binarias y una tarea que debe responder 1 cuando son distintas y 0 cuando son iguales, conocida como disyunción exclusiva o XOR.

Puede construirse una red pequeña con h1 = máximo entre 0 y x1 − x2, h2 = máximo entre 0 y x2 − x1 y salida y = h1 + h2; los pesos se fijan deliberadamente para explicar la función, por lo que este ejemplo demuestra una capacidad de representación y no que cierto procedimiento de entrenamiento encontrará esos valores.

Tabla 1. *Comprobación exhaustiva de una red construida para XOR*

Entradas x1 y x2	Primera unidad h1	Segunda unidad h2	Salida y
0 y 0	0	0	0
0 y 1	0	1	1
1 y 0	1	0	1
1 y 1	0	0	0

Nota. Elaboración propia; el universo de prueba se limita a las cuatro combinaciones binarias definidas y los valores se obtienen por sustitución directa.

La tabla permite una verificación completa porque el dominio contiene sólo cuatro casos; si las entradas fueran



cualquier número real, la misma red calcularía la distancia absoluta entre ambos números, pero ya no sería correcto llamar XOR a toda su conducta sin explicar el cambio de dominio y de interpretación.

El ejemplo también permite separar arquitectura, parámetros y función: la arquitectura establece dos rutas rectificadas que se suman, los parámetros fijan las diferencias y la función resultante depende de ambas cosas; otra elección de pesos dentro de una estructura semejante podría generar un comportamiento distinto.

¿Por qué importa la activación no lineal? Si una primera transformación es h = A × x + b y una segunda es y = C × h + d, entonces su composición es y = C × A × x + C × b + d; al eliminar h aparece otra transformación afín, por lo que encadenar esas dos operaciones sin no linealidad no crea una familia funcional más amplia que la de una sola transformación afín compatible.

La conclusión anterior es algebraica y tiene un alcance preciso: se refiere a esas transformaciones, no a todas las redes ni a todas las razones para factorizar matrices; introducir restricciones, objetivos de optimización o

mecanismos diferentes puede modificar el problema práctico, aunque no invalida la igualdad demostrada.

Una explicación que reduzca el valor de una red a tener muchas capas pierde esta distinción, porque importa qué operaciones contienen, qué restricciones comparten y qué información conservan; contar niveles en un dibujo no permite calcular directamente su utilidad ni asignarle un nivel de inteligencia humana.

También puede contarse el tamaño de una red sin recurrir a comparaciones biológicas: una capa densa de dos entradas y tres unidades tiene seis pesos y tres sesgos, y una salida que recibe esas tres unidades agrega tres pesos y un sesgo; el conjunto posee trece parámetros ajustables, aunque sólo se hayan definido tres unidades ocultas y una unidad de salida.

Confundir parámetros con neuronas produce incluso un error dentro de este sistema diminuto, antes de intentar cualquier comparación con un cerebro; las cifras describen objetos diferentes y sólo pueden relacionarse después de especificar qué se está contando y con qué finalidad.

Aprender exige un criterio de mejora

Una red puede cambiar sus parámetros sin mejorar en la tarea que interesa, por lo que aprendizaje no debe utilizarse como sinónimo de cualquier modificación; hace falta identificar un objetivo, una regla de ajuste y una evaluación que permita juzgar qué consiguió el cambio.

Rumelhart et al. (1986, pp. 533–535) describieron un procedimiento que calcula cómo contribuyen los pesos al error de salida y utiliza esa información para ajustarlos, además mostraron representaciones internas obtenidas en tareas concretas; la contribución pertinente aquí es la relación explícita entre función, error y actualización, no una demostración de identidad con el aprendizaje cerebral.

Para reconstruir un paso sin ocultarlo detrás de vocabulario técnico, se empleará una unidad de salida lineal con $x1 = 2$, $x2 = 1$, $w1 = 0.2$, $w2 = 0.1$ y $b = 0$; su predicción es $y = 0.5$, mientras la salida deseada se fija en $t = 1$ y la pérdida se define como $L = (y - t)^2 / 2$.

La pérdida inicial es $(0.5 - 1)^2 / 2 = 0.125$; el cuadrado hace que una desviación positiva y una negativa de igual magnitud reciban el mismo costo, aunque elegir ese

criterio es una decisión del ejemplo y no una propiedad inevitable de todo problema de aprendizaje.

El gradiente indica cómo cambia localmente la pérdida cuando cambia cada parámetro; para esta función, las derivadas son (y − t) × x1, (y − t) × x2 y y − t, que toman los valores −1, −0.5 y −0.5, respectivamente.

Con una tasa de aprendizaje de 0.1, se resta a cada parámetro esa tasa multiplicada por su derivada, obteniendo w1 nuevo = 0.3, w2 nuevo = 0.15 y b nuevo = 0.05; la nueva predicción es 0.3 × 2 + 0.15 × 1 + 0.05 = 0.8, y la nueva pérdida es 0.02.

Tabla 2. *Un paso de ajuste y una alternativa demasiado grande*

Condición del cálculo	Predicción	Pérdida definida	Lectura válida
Parámetros iniciales	0.5	0.125	Punto de partida
Paso con tasa 0.1	0.8	0.02	Mejora en este ejemplo
Paso con tasa 1	3.5	3.125	Empeora en este ejemplo



Nota. Elaboración propia; ambos pasos parten de los mismos parámetros iniciales y no representan actualizaciones consecutivas ni resultados experimentales.

La tercera fila se obtiene al usar una tasa de 1 desde el punto inicial: los parámetros resultan 1.2, 0.6 y 0.5, por lo que la predicción es 3.5; la dirección local de descenso no garantiza que cualquier desplazamiento de cualquier tamaño reduzca la pérdida, una precaución visible incluso en un problema sencillo.

En este caso particular puede comprobarse algo más: al actualizar los tres parámetros, el error e = y − t se transforma en e nuevo = e × (1 − 6 × tasa), porque $2^2 + 1^2 + 1^2 = 6$; para reducir su magnitud con una tasa positiva basta que ésta sea menor que un tercio, mientras ese límite no debe trasladarse a redes diferentes.

La retropropagación extiende el cálculo de derivadas a composiciones de operaciones mediante la regla de la cadena; el hecho de que una operación anterior afecte una pérdida posterior permite calcular su contribución sin que



cada unidad deba conocer un concepto humano como equivocación, culpa o intención.

Conviene separar calcular el gradiente de elegir cómo actualizar: el primero informa sobre una variación local y el segundo puede incorporar tasas, promedios u otras decisiones del procedimiento; tampoco debe confundirse reducir la pérdida en un caso con mejorar la conducta en entradas no utilizadas durante el ajuste.

Este último límite resulta decisivo para la analogía cerebral, pues decir que el modelo aprendió algo puede ser una abreviatura válida si se identifica una mejora comprobada; afirmar que vivió una experiencia equivalente a la de un estudiante agrega propiedades que el cálculo anterior nunca examinó.

Parámetros y representaciones no son lo mismo

En los ejemplos anteriores, los parámetros permanecen fijos mientras se calcula una salida y las activaciones cambian cuando llegan otras entradas; esta distinción básica evita confundir una modificación del modelo con un



estado transitorio de su procesamiento, aunque existan diseños donde ambos procesos se intercalen.

Una representación es una forma de organizar información que otra operación puede utilizar, por ejemplo el par de números producido por una capa; para interpretarla hay que precisar qué información permite recuperar, en qué condiciones y con qué procedimiento, en lugar de asignarle significado a partir de una imagen sugerente de sus unidades.

Consideremos una representación h = (x1, x2) y otra h alternativa = (x1 + x2, x1 − x2); de la segunda puede recuperarse x1 sumando sus dos componentes y dividiendo entre dos, y puede recuperarse x2 restándolos y dividiendo entre dos, de modo que ambas conservan la información original bajo aritmética exacta.

Sin embargo, las coordenadas no significan lo mismo: en la primera, cada componente coincide con una entrada, mientras en la segunda cada componente mezcla ambas; una búsqueda de correspondencias entre unidades aisladas podría juzgarlas muy diferentes y perder la



equivalencia informativa demostrada por la transformación reversible.

De esa observación no se sigue que todas las representaciones de una red sean equivalentes, porque una transformación que elimina información puede impedir la recuperación; si sólo se conserva la suma, las entradas 1 y 3 resultan indistinguibles de 2 y 2, un contraejemplo que muestra por qué deben examinarse las operaciones concretas.

Decodificar una propiedad a partir de una activación tampoco prueba que el sistema la utilice para decidir; en una construcción hipotética, una red podría conservar dos datos en su estado interno y calcular la salida exclusivamente con el primero, de modo que un observador recuperaría el segundo sin que éste desempeñara un papel causal en la respuesta.

Para investigar ese papel habría que intervenir sobre la información o sobre la ruta que la transmite y examinar las consecuencias con controles adecuados; observar una correlación sirve para formular la hipótesis, pero no



sustituye la prueba de dependencia y puede ser compatible con varias explicaciones.

La misma cautela se aplica a expresiones como neurona de un objeto o recuerdo almacenado en un peso: pueden designar resultados interpretativos de un estudio específico, pero una etiqueta aislada no establece exclusividad, estabilidad ni necesidad; antes de utilizarla conviene preguntar qué observaciones la sostienen y qué pasaría al modificar el componente.

Lo que una comparación biológica debe especificar

La diferencia entre una unidad artificial y un modelo biofísico puede verse en un trabajo clásico de Hodgkin y Huxley (1952, pp. 500–507): su análisis de una fibra nerviosa gigante relaciona corriente de membrana, voltaje, tiempo y componentes iónicos, utilizando ecuaciones con variables físicas y contrastes experimentales; ese objeto es distinto de una suma ponderada destinada a minimizar un error de clasificación.

Los propios autores explicaron los límites de ciertas interpretaciones físicas de sus ecuaciones y no

presentaron toda simplificación como una descripción molecular definitiva; este antecedente recuerda que incluso un modelo profundamente vinculado con medidas biológicas conserva supuestos y niveles de aproximación (Hodgkin & Huxley, 1952, p. 506).

La comparación útil comienza por identificar la variable que pretende representar: un número sin unidades que ayuda a clasificar imágenes y una magnitud de voltaje no son intercambiables por llevar ambos el nombre de activación; para relacionarlos haría falta una correspondencia explícita y una prueba de que esa correspondencia conserva la propiedad investigada.

Tampoco basta con añadir más detalles a un modelo para volverlo más adecuado para cualquier propósito; si se busca explicar un fenómeno dependiente del tiempo, omitirlo puede ser crucial, mientras una pregunta sobre clasificación estática puede admitir una abstracción más sencilla, siempre que se evalúe su alcance.

La selección de detalles es una decisión científica que debe defenderse respecto de la pregunta, no respecto del parecido del dibujo; un mapa de una ciudad puede omitir



la altura de los edificios para orientar un recorrido peatonal, pero esa omisión impediría utilizarlo directamente para estudiar sombras, una analogía metodológica sobre modelos que no atribuye propiedades urbanas al cerebro.

En términos prácticos, un informe que diga inspirado en el cerebro debería agregar qué aspecto inspiró el diseño, qué aspecto se modeló cuantitativamente y qué aspecto fue comprobado; esas tres preguntas pueden tener respuestas diferentes, y la inspiración histórica no proporciona por sí misma evidencia de fidelidad actual.

La materialidad tampoco debe utilizarse como un atajo negativo: que dos sistemas estén construidos con materiales distintos impide su identidad material, pero no resuelve si implementan alguna relación funcional comparable; esa cuestión debe examinarse mediante la relación propuesta, en lugar de decidirla sólo por el vocabulario de biología o computación.

El problema del aprendizaje biológicamente plausible

El artículo de Rumelhart et al. (1986, p. 536) reconoció expresamente que el procedimiento presentado, en su



forma de entonces, no era un modelo plausible del aprendizaje cerebral; esta precisión impide convertir una publicación fundamental para el aprendizaje artificial en una certificación de que el cerebro implementa literalmente su algoritmo.

La reserva tampoco obliga a descartar toda investigación sobre mecanismos relacionados: Lillicrap et al. (2016) estudiaron redes donde señales de error se transmitían mediante pesos de retroalimentación fijos y aleatorios, mostrando que la simetría exacta de pesos no era indispensable para aprender en los problemas simulados; su resultado abrió posibilidades algorítmicas y no identificó un mecanismo cerebral ya demostrado.

El contraste permite distinguir plausibilidad de confirmación: mostrar que una regla puede funcionar bajo restricciones menos exigentes responde una objeción a esa posibilidad, pero todavía falta establecer si las estructuras y procesos biológicos pertinentes existen y desempeñan la función atribuida en condiciones observables.

Los autores delimitaron las simplificaciones de sus modelos y las preguntas abiertas sobre señales de error, conexiones y plasticidad, por lo que su investigación resulta más informativa cuando se conserva esa condición; presentar la simulación como descubrimiento de la regla universal del cerebro borraría precisamente la distancia que el estudio permite investigar (Lillicrap et al., 2016, sección Discusión).

Un ejemplo lógico ayuda a entender la diferencia: si se demuestra que una máquina puede resolver una tarea sin un componente antes considerado indispensable, queda refutada la necesidad universal de ese componente para esa clase de solución; no queda demostrado que otra máquina particular use el nuevo procedimiento, pues podría resolver la tarea de una tercera manera.

Esta forma de razonar evita dos excesos frecuentes, el de atribuir al cerebro cualquier método eficaz y el de declarar imposible toda correspondencia porque un algoritmo concreto no satisface alguna condición; las hipótesis deben formularse con suficiente precisión para que una observación pueda apoyarlas, limitarlas o descartarlas.

Una agenda de investigación responsable puede avanzar entre esos extremos: identificar restricciones biológicas relevantes, proponer procedimientos compatibles, comparar sus predicciones y diseñar mediciones que distingan alternativas; reconocer lo que falta no disminuye el valor de un resultado, sino que impide confundir el siguiente problema con uno ya resuelto.

Cuatro preguntas que necesitan pruebas distintas

La expresión se parece al cerebro es demasiado amplia para evaluar un resultado, porque puede referirse a conducta observable, organización de representaciones, mecanismos causales o realización biológica; esas dimensiones pueden estudiarse de manera relacionada, pero una respuesta favorable en una no cancela las preguntas pendientes en las demás.

Una comparación conductual pregunta si dos sistemas responden de forma semejante ante condiciones definidas, incluyendo sus errores; una comparación representacional pregunta si ciertas relaciones internas permiten explicar o predecir mediciones, mientras una comparación

mecanística investiga qué componentes y dependencias producen el resultado.

La fidelidad biológica agrega exigencias sobre variables, procesos y restricciones físicas pertinentes; no consiste en exigir que todo modelo copie cada detalle de un organismo, sino en especificar cuáles de esos detalles forman parte de la afirmación y cuáles quedan deliberadamente fuera.

Tabla 3. *Tipos de semejanza y alcance de una comprobación*

Pregunta	**Evidencia pertinente**	**Conclusión que no se obtiene automáticamente**
Responden de modo parecido	Tareas y errores comparables	Usan el mismo procedimiento
Sus representaciones se relacionan	Predicción y comparación en datos reservados	Cada unidad equivale a una célula

Pregunta	Evidencia pertinente	Conclusión que no se obtiene automáticamente
Comparten una dependencia causal	Intervenciones y predicciones discriminantes	Toda su organización es idéntica
Modelan variables biológicas	Magnitudes, restricciones y contrastes físicos	Poseen todas las capacidades de una persona

Nota. Elaboración propia; las filas organizan exigencias de evaluación y no representan etapas inevitables ni una escala única de inteligencia.

Estas preguntas también impiden utilizar una ausencia de evidencia como prueba absoluta de ausencia: si un estudio sólo midió clasificación visual, no está diseñado para resolver todas las propiedades del sistema; la conclusión debe permanecer dentro de lo observado y de las inferencias que ese diseño permite justificar.

El error inverso aparece cuando cualquier dato favorable se presenta como confirmación global, por ejemplo al convertir una correlación entre respuestas en una equivalencia completa entre arquitecturas; para evitarlo, la afirmación debe poder reescribirse mencionando la tarea, la población, la medición y las condiciones de la comparación.

Una buena pregunta de control es qué resultado haría cambiar la conclusión: si ninguna diferencia imaginable puede contar contra una analogía, probablemente ésta está funcionando como recurso retórico y no como hipótesis; en cambio, una predicción precisa puede conservar valor aunque después se refute parcialmente.

Una red puede contribuir a estudiar la visión

Yamins et al. (2014) investigaron modelos jerárquicos optimizados para reconocimiento visual y su capacidad de predecir respuestas en corteza inferotemporal y V4 de macacos; dentro de la familia estudiada encontraron una relación entre desempeño de categorización y predictividad neuronal, un resultado que muestra la utilidad

científica de comparar modelos artificiales con datos biológicos.

La comparación no consistió en emparejar automáticamente cada unidad artificial con una neurona: los autores ajustaron combinaciones lineales de salidas del modelo para predecir sitios registrados y evaluaron esas predicciones con imágenes nuevas; distinguir el modelo de esa lectura ajustada es necesario para describir lo que efectivamente se comprobó (Yamins et al., 2014, pp. 8620, 8624).

Sus resultados no autorizan a afirmar que cualquier clasificador preciso reproducirá el cerebro, pues el propio trabajo examinó controles y el alcance de la clase de modelos; además, la predictividad se refería a respuestas y condiciones visuales determinadas, no a la totalidad de la actividad cerebral ni a una identidad entre aprendizaje artificial y desarrollo humano.

La lección metodológica que se deriva aquí consiste en reconocer dos objetos de evaluación: la representación generada por una red y el procedimiento que la relaciona con una medición; si el segundo es muy flexible, parte del

éxito podría depender de él, por lo que conviene describirlo, limitarlo y comprobarlo fuera de los datos usados para ajustarlo.

Un ejemplo hipotético ilustra el riesgo: si se permite construir una regla distinta para cada imagen ya conocida, sería posible reproducir muchas respuestas almacenadas sin aprender una relación transferible; reservar imágenes nuevas evita esa forma simple de memorización, aunque otras decisiones del diseño todavía deban revisarse.

También importa qué se reserva: imágenes nuevas de los mismos objetos prueban algo diferente de objetos nuevos o condiciones de presentación nuevas; por ello, la palabra generalización necesita indicar qué dimensión cambió y cuáles permanecieron constantes, en vez de funcionar como una aprobación indistinta del modelo.

Nada de lo anterior invalida el uso científico de redes artificiales, porque un modelo predictivo puede ayudar a formular hipótesis incluso cuando no reproduce todos los mecanismos; el límite está en presentar esa ayuda como evidencia de un alcance que el estudio no puso a prueba.

Predecir una región no reproduce sus dependencias

St-Yves et al. (2023) compararon redes optimizadas para predecir respuestas de resonancia magnética funcional en áreas visuales humanas V1 a V4, incluyendo una arquitectura de rama única y otra con ramas independientes; ambas alcanzaron precisión predictiva semejante, aunque no conservaron las mismas formas de organización jerárquica.

El estudio distinguió jerarquías de composición, dependencia entre representaciones e integración espacial, y encontró que ciertas jerarquías no eran necesarias para obtener predicciones precisas en sus modelos; no demostró que el sistema visual biológico carezca de organización jerárquica, sino que la predictividad de un modelo no obliga a una interpretación anatómica única (St-Yves et al., 2023, secciones Resultados y Discusión).

Esta diferencia es especialmente importante porque los autores advirtieron que un modelo con ramas independientes puede conservar predicciones para unas áreas aunque se elimine otra rama; esa propiedad de la arquitectura artificial limita una lectura mecanística directa,

incluso cuando la representación resulta útil para anticipar datos.

Puede reconstruirse la lógica mediante un contraejemplo propio mucho más sencillo: el sistema A calcula h = x y después y = 2 × h, mientras el sistema B calcula h = x y, por una ruta independiente, y = 2 × x; en observación ordinaria producen los mismos valores de h y y para cualquier entrada.

Si se fuerza h a cero dejando x intacta, el sistema A produce y = 0 y el B mantiene y = 2 × x; las observaciones iniciales no identificaban cuál de las dos organizaciones existía, pero la intervención distingue sus dependencias porque modifica una variable interna sin cambiar la entrada original.

Este cálculo no representa una intervención cerebral real y tampoco prescribe cómo realizarla; su función es demostrar, por construcción, que equivalencia de respuestas observadas e identidad de estructura causal son proposiciones diferentes, por lo que la primera no basta para deducir la segunda.



La enseñanza se extiende a la lectura de diagramas: si una flecha sólo indica que dos medidas están asociadas, no debe describirse como si acreditara una conexión causal necesaria; si indica una dependencia programada, sí permite examinar el programa, pero trasladarla a un organismo requiere evidencia adicional sobre ese organismo.

Además, una intervención debe interpretarse con sus propios límites: eliminar un componente puede tener consecuencias por varias rutas, y un resultado nulo puede reflejar redundancia, compensación o una intervención insuficiente; la prueba causal es una familia de problemas de diseño, no una palabra que transforme cualquier modificación en explicación concluyente.

Conducta parecida y señales diferentes

Geirhos et al. (2019) compararon observadores humanos y redes convolucionales con imágenes que enfrentaban señales de forma y textura, y encontraron diferencias en las respuestas de los modelos evaluados; también mostraron que cambiar los datos de entrenamiento podía aumentar la preferencia por forma en una arquitectura

estudiada, por lo que el resultado no debe atribuirse como destino inmutable a toda red convolucional.

La inferencia corresponde al diseño de estímulos, participantes y modelos examinados, no a una ley universal sobre toda visión humana y artificial; el trabajo es útil porque convierte una intuición sobre qué señal guía la clasificación en una prueba comparativa, además de distinguir la arquitectura de las condiciones en que se ajusta (Geirhos et al., 2019, secciones 2 y 3).

El principio puede adaptarse a un caso profesional hipotético: una organización quiere clasificar documentos en categorías A y B, pero las imágenes de A tienen siempre borde oscuro y las de B borde claro; una regla que sólo observa el borde y otra que observa información pertinente del documento pueden acertar por igual en ese conjunto.

Supongamos un conjunto inicial de cuarenta imágenes, veinte de cada categoría, donde el borde coincide perfectamente con la etiqueta; ambas reglas construidas alcanzan cuarenta aciertos, equivalentes al cien por ciento, aunque sólo una utiliza el contenido que interesa para la tarea.

Para distinguirlas se prepara otro conjunto sintético de cuarenta imágenes: diez documentos A con borde oscuro, diez A con borde claro, diez B con borde oscuro y diez B con borde claro; se supone, exclusivamente para el ejemplo, que el contenido permanece reconocible y que la regla basada en contenido clasifica sin error.

La regla del borde acierta los diez A oscuros y los diez B claros, pero falla los otros veinte, por lo que obtiene veinte aciertos de cuarenta y una exactitud del cincuenta por ciento; la regla del contenido conserva cuarenta de cuarenta, una diferencia completamente determinada por las reglas construidas y no observada en una aplicación real.

Tabla 4. *Prueba sintética de dos reglas que antes coincidían*

Grupo de diez documentos	Etiqueta correcta	Respuesta por borde	Respuesta por contenido
A con borde oscuro	A	A	A

Grupo de diez documentos	Etiqueta correcta	Respuesta por borde	Respuesta por contenido
A con borde claro	A	B	A
B con borde oscuro	B	A	B
B con borde claro	B	B	B

Nota. Elaboración propia; cada fila contiene diez casos sintéticos, las reglas son deterministas y no se atribuyen estos resultados a una red entrenada ni a una organización real.

El ejemplo enseña por qué una calificación agregada puede ocultar una dependencia impropia: mientras las dos señales coinciden, los aciertos no permiten saber cuál se utiliza; separarlas experimentalmente cambia la pregunta y puede revelar diferencias que una muestra ordinaria no mostraba.



En una aplicación real habría que comprobar que modificar el borde no altere también el contenido relevante ni introduzca un artefacto nuevo, además de evaluar suficientes casos independientes; si la edición degrada toda la imagen, una caída de desempeño no aislaría necesariamente el papel del borde.

La metáfora de un cerebro que entendió el documento puede desviar la atención de ese problema, mientras una descripción de entradas, señales y salidas conduce a una prueba concreta; una explicación sobria no niega la utilidad del clasificador, sino que ayuda a decidir qué evidencia falta antes de confiar en él.

Cómo leer una cifra de semejanza

Una cifra alta de semejanza sólo resulta informativa cuando se sabe qué resume: coincidencia de etiquetas, correlación entre respuestas, error de predicción o semejanza de relaciones internas son medidas distintas; su interpretación depende del diseño y no puede reducirse a un porcentaje de humanidad.

Considérese el vector hipotético de respuestas observadas (1, 2, 3, 4) y una predicción (12, 14, 16, 18), obtenida al multiplicar cada valor por dos y sumar diez; su correlación de Pearson es uno, porque mantienen una relación lineal positiva perfecta, aunque sus valores numéricos no coinciden.

Si la afirmación investigada fuera que ambas series ordenan de igual manera los estímulos, el resultado sería compatible con ella; si fuera que predicen correctamente la magnitud en las mismas unidades, haría falta examinar escala, desplazamiento y error, de modo que una métrica adecuada para una pregunta puede ser insuficiente para otra.

La solución no consiste en prohibir correlaciones ni en exigir que cada activación artificial tenga las unidades de una medición biológica, sino en describir la transformación que las relaciona y la interpretación autorizada por esa transformación; el problema aparece cuando se cambia silenciosamente de una pregunta a otra al redactar la conclusión.



También debe distinguirse variación explicable de variación total cuando un estudio utiliza estimaciones de ruido: el denominador modifica el significado de una proporción, y omitirlo puede producir una lectura exagerada; no es correcto convertir una fracción de variación predecible en una fracción del cerebro reproducida.

Una comparación de representaciones debe informar además qué estímulos comparte, cómo se agregan las respuestas y qué libertad tiene el ajuste; dos modelos pueden conservar un orden general y diferir en errores concretos relevantes para el uso, de modo que un promedio necesita acompañarse de análisis de casos y condiciones.

La selección de resultados introduce otra dificultad: si se prueban muchas capas y sólo se publica la mejor coincidencia en los mismos datos, la cifra puede reflejar parte de esa búsqueda; reservar una evaluación final independiente permite poner a prueba la elección en lugar de tratar el máximo encontrado como si hubiera sido una predicción previa.

Estas advertencias son criterios de lectura y propuestas metodológicas del artículo, no acusaciones sobre los estudios citados; su propósito es que el lector pueda preguntar qué se midió, qué se ajustó y qué se reservó antes de aceptar una interpretación amplia de una gráfica.

Diseñar una comparación que pueda fallar

Una evaluación útil puede comenzar con una frase estrecha, por ejemplo que un modelo mantendrá el mismo patrón de respuestas que un grupo definido ante una familia especificada de cambios visuales; esa afirmación es más laboriosa que decir procesa como el cerebro, pero también permite establecer qué observaciones serían favorables y cuáles la debilitarían.

Después se necesita separar semejanza esperada y desempeño deseado, porque una aplicación puede requerir superar ciertos errores humanos en lugar de reproducirlos; si el objetivo es científico, compartir errores puede resultar informativo, mientras si el objetivo es operativo, esos mismos errores pueden constituir una razón para rechazar el sistema.

El siguiente paso consiste en definir controles, incluyendo alternativas que puedan resolver la tarea por procedimientos diferentes; comparar sólo con una red claramente insuficiente puede mostrar una mejora, pero deja abierta la posibilidad de que otro mecanismo igualmente eficaz explique los mismos datos sin la correspondencia propuesta.

También conviene fijar qué partes del sistema pueden ajustarse y con qué información, porque un modelo y una transformación de lectura muy flexible forman una combinación distinta de un modelo evaluado con una transformación restringida; la descripción debe permitir reconstruir ambos, no sólo mostrar el nombre de la arquitectura principal.

La división entre desarrollo y evaluación final debe preservar independencia respecto de las decisiones que se pretenden validar; si una condición de prueba se utiliza repetidamente para escoger variantes, termina formando parte del proceso de selección y deja de desempeñar el mismo papel como comprobación externa.

En estudios de intervención, resulta necesario anticipar explicaciones alternativas: una modificación puede cambiar simultáneamente varias señales, destruir información general o introducir una entrada ajena al ámbito de uso; los controles deben ayudar a distinguir esos efectos de la dependencia específica que interesa investigar.

Por último, se necesita una regla de interpretación anterior al entusiasmo por los resultados: qué se considerará apoyo limitado, qué obligará a revisar la hipótesis y qué quedará sin resolver aunque la prueba salga bien; este cierre evita convertir cualquier patrón encontrado después en una confirmación retrospectiva de la misma historia.

El procedimiento propuesto no exige que toda persona usuaria realice experimentos neurocientíficos, sino que adapte esas preguntas a la afirmación que recibe; para comprar una herramienta de clasificación suele ser más pertinente exigir evidencia sobre su tarea que una comparación ornamental con el cerebro.



Mitos y objeciones que conviene separar

El primer mito afirma que tener neuronas artificiales demuestra que el sistema funciona como un cerebro; el término describe una tradición de modelación y una clase de componentes, pero el nombre no verifica correspondencias entre variables, conexiones o reglas de cambio, por lo que la afirmación necesita reformularse de manera comprobable.

El segundo sostiene que una red sólo realiza operaciones numéricas y por eso ningún resultado puede ser significativo; el argumento confunde una descripción de implementación con una evaluación funcional, pues los cálculos pueden producir una clasificación útil y deben juzgarse en relación con el problema que resuelven.

El tercer mito identifica una unidad con un concepto humano completo, aunque nuestros ejemplos ya mostraron que una misma información puede distribuirse entre coordenadas diferentes; una interpretación local puede ser útil, pero necesita pruebas sobre su alcance y no puede deducirse del rótulo elegido por quien presenta una visualización.

El cuarto transforma el éxito en predecir una medición en la reconstrucción del mecanismo que la genera; el contraejemplo de dos rutas que coinciden en observación y divergen bajo intervención muestra por qué ese salto no es válido en general, incluso antes de abordar la complejidad de un organismo.

Una objeción razonable advierte que exigir demasiada precisión puede volver incomprensible la divulgación; la respuesta no es eliminar las analogías, sino acompañarlas de un límite breve y concreto, como aclarar que la unidad combina números y que no representa todos los procesos de una célula.

Otra objeción señala que ningún modelo reproduce un objeto completo, de modo que exigir identidad total sería un criterio imposible; precisamente por eso este artículo propone evaluar correspondencias parciales con preguntas delimitadas, sin exigir una copia exhaustiva ni presentar una semejanza parcial como equivalencia global.

Existe además el riesgo de utilizar la cautela como una forma de negar todo progreso: si un estudio permite



predecir datos nuevos mejor que alternativas pertinentes, ese resultado merece reconocimiento dentro de su alcance; señalar lo que no demuestra no debe borrar lo que sí estableció.

La discusión sobre experiencia subjetiva queda fuera de estas demostraciones, porque ni el cálculo de XOR ni una correlación visual proporcionan una prueba suficiente sobre ella; esa delimitación no afirma una respuesta universal sobre cualquier sistema posible, sino que identifica una pregunta que requiere criterios independientes.

Implicaciones para enseñar y utilizar redes

Una explicación introductoria mejora cuando presenta primero una operación reconstruible y después señala qué aspectos omite; al seguir una entrada, un peso, una activación y una salida, el lector adquiere herramientas para preguntar por el sistema real sin depender de imágenes mentales de un cerebro en miniatura.

En formación técnica conviene distinguir capacidad representacional, procedimiento de ajuste y



generalización: la red XOR construida demuestra la primera, el paso de gradiente ilustra el segundo y ninguno de los dos, por sí solo, acredita desempeño en un problema externo; conservar esas fronteras evita que una demostración didáctica parezca una validación de producto.

En un informe profesional resulta preferible describir que un sistema clasificó cierto conjunto bajo ciertas condiciones antes que afirmar que comprendió como una persona; cuando se utilice una abreviatura antropomórfica, su sentido técnico debe quedar claro y no convertirse después en una premisa más fuerte que los resultados disponibles.

Para evaluar proveedores, una pregunta útil es qué ventaja práctica aporta la semejanza cerebral que anuncian: si no se relaciona con una mejora verificable, una restricción relevante o una predicción contrastable, puede tratarse únicamente de una descripción de inspiración; la decisión necesita evidencia sobre el servicio solicitado.

Al documentar errores, interesa registrar qué entrada los produjo, qué salida apareció, qué condiciones cambiaron



y qué alternativas siguen abiertas; una narración sobre lo que la red quiso hacer puede resultar atractiva, pero añade una interpretación que no sustituye el análisis de dependencias observables.

Finalmente, debe mantenerse la responsabilidad humana sobre el uso y la comunicación de resultados: llamar neuronal a una herramienta no la convierte en una autoridad científica autónoma ni descarga sobre ella la obligación de verificar una conclusión; la comparación con el cerebro sirve mejor cuando aumenta la precisión de nuestras preguntas.

Conclusión

Las redes neuronales artificiales pueden explicarse mediante operaciones, composiciones y procedimientos de ajuste que no requieren atribuirles biología humana; al mismo tiempo, pueden aportar modelos valiosos para investigar relaciones entre tareas, representaciones y mediciones cerebrales, siempre que esas relaciones se formulen y se prueben con suficiente precisión.

Los ejemplos matemáticos permiten separar contar parámetros, representar una función y reducir una pérdida, mientras los estudios revisados muestran por qué aprender en una simulación, predecir actividad y reproducir un mecanismo son logros distintos; confundirlos debilita tanto la divulgación como la interpretación científica.

La comparación con el cerebro conserva su valor cuando especifica qué aspecto aproxima, qué observación la respalda y qué queda fuera; pierde ese valor cuando el nombre neuronal sustituye la evidencia o cuando una diferencia material se utiliza para negar cualquier correspondencia funcional posible.

La pregunta práctica no es cuánto se parece el dibujo a un cerebro, sino qué relación se propone y cómo podría ponerse a prueba; responderla permite reconocer capacidades reales, investigar límites y tomar decisiones mejor fundadas sin convertir una metáfora en una explicación completa.

Referencias

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., y Brendel, W. (2019). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness [Las redes convolucionales entrenadas en ImageNet presentan sesgo de textura; el sesgo de forma mejora la exactitud y robustez] (versión 3, 2022). En *International Conference on Learning Representations*. https://doi.org/10.48550/arXiv.1811.12231

Hodgkin, A. L., y Huxley, A. F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve [Una descripción cuantitativa de la corriente de membrana y su aplicación a la conducción y excitación nerviosa]. *The Journal of Physiology, 117*(4), 500–544. https://doi.org/10.1113/jphysiol.1952.sp004764

Lillicrap, T. P., Cownden, D., Tweed, D. B., y Akerman, C. J. (2016). Random synaptic feedback weights support error backpropagation for deep learning [Los pesos sinápticos aleatorios de retroalimentación permiten la retropropagación de errores para el aprendizaje profundo]. *Nature Communications, 7*, Artículo 13276. https://doi.org/10.1038/ncomms13276

Rumelhart, D. E., Hinton, G. E., y Williams, R. J. (1986). Learning representations by back-propagating errors [Aprendizaje de representaciones mediante retropropagación de errores]. *Nature, 323*(6088), 533–536. https://doi.org/10.1038/323533a0

St-Yves, G., Allen, E. J., Wu, Y., Kay, K., y Naselaris, T. (2023). Brain-optimized deep neural network models of human visual areas learn non-hierarchical representations [Los modelos de redes neuronales profundas optimizados para el cerebro en áreas visuales humanas aprenden representaciones no jerárquicas]. *Nature Communications, 14*(1), Artículo 3329. https://doi.org/10.1038/s41467-023-38674-4

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., y DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex [Los modelos jerárquicos optimizados por desempeño predicen respuestas neuronales en la corteza visual superior]. *Proceedings of the National Academy of Sciences, 111*(23), 8619–8624. https://doi.org/10.1073/pnas.1403112111

