

El cerebro y sus interrogantes:

preguntas abiertas en
neurociencia computacional



Felipe Bravo-Márquez

PhD por la Universidad de Waikato, Nueva Zelanda. Profesor Asociado del Departamento de Ciencias de la Computación de la Universidad de Chile y Director de la Iniciativa de Datos e Inteligencia Artificial. Investigador Principal del Centro Nacional de Inteligencia Artificial e Investigador del Instituto Milenio Fundamentos de los Datos. Sus intereses incluyen NLP, aprendizaje automático, inteligencia artificial y recuperación de información.

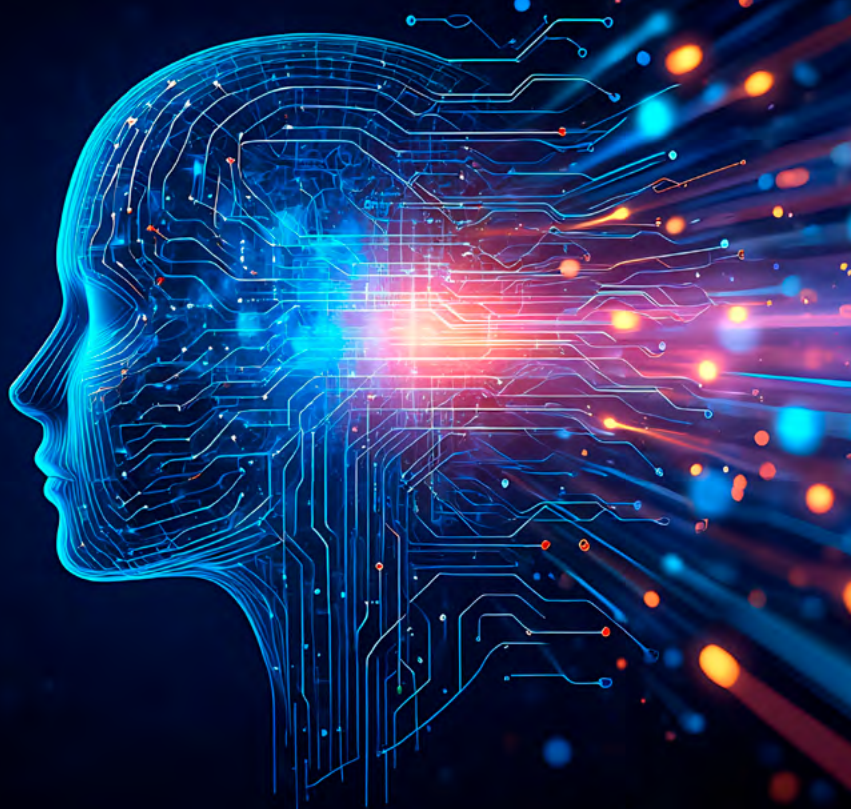
✉ fbravo@dcc.uchile.cl



Rodrigo Carrasco-Davis

PhD por la Gatsby Computational Neuroscience Unit (UCL). Actualmente Investigador Postdoctoral en el Princeton Neuroscience Institute, en el laboratorio de neurociencia de control cognitivo. Su investigación se sitúa en la intersección entre neurociencia computacional y aprendizaje automático, con foco en meta-aprendizaje y sistemas de memoria.

✉ rodrigo.cd@princeton.edu





Resumen / En los últimos años, la inteligencia artificial ha tenido un desarrollo vertiginoso y ha permeado muchos ámbitos de nuestra vida. A pesar del éxito que ha tenido, quedan múltiples interrogantes científicas: ¿cómo podrían los sistemas de IA interactuar, aprender y decidir de forma continua y autónoma según el ambiente que los rodea? ¿Cómo podrían coexistir con un cuerpo físico? ¿O cómo podrían generar motivaciones propias? Algunas de estas cualidades están presentes en la forma en que nosotros, los humanos y otros animales, nos desarrollamos durante nuestra vida, por lo que resulta natural volcar nuestra atención hacia cómo funciona el cerebro para intentar contestar estas interrogantes.

En este artículo presentamos nociones básicas de neurociencia computacional, un área que históricamente ha buscado formalizar matemáticamente los mecanismos que rigen la dinámica del cerebro. Introduciremos conceptos básicos del funcionamiento del cerebro bajo el lente de modelos computacionales, además de destacar preguntas actuales en neurociencia que podrían servir para inspirar avances en inteligencia artificial, o bien ser inspiradas por ellos.

Neurociencia computacional

La neurociencia como tal es un área de investigación dedicada a estudiar el cerebro, aunque este tipo de investigación puede tomar muchas formas. Por ejemplo, uno podría argumentar que la psicología es una forma de investigar el cerebro a través del comportamiento humano. Otro ejemplo podría ser la neurología, que tiene un enfoque médico, con el objetivo de identificar y tratar patologías asociadas al cerebro.

En contraste con estas áreas, la neurociencia computacional busca describir lo observado y generado por el cerebro de forma matemática, es decir, a través de ecuaciones que puedan usarse para simular, predecir y comprender la dinámica de lo observado. Esta descripción puede tener muchos niveles de abstracción, como veremos más adelante. Por ejemplo, existen ecuaciones matemáticas que describen distintos aspectos del cerebro a distintas escalas y niveles computacionales: cómo varía el voltaje de una neurona en el tiempo, cómo un grupo de neuronas reacciona a partir de la actividad de otras, cuánta información y de qué tipo se propaga a través de distintas zonas del cerebro, cómo el cerebro se reconfigura con el tiempo durante el aprendizaje, qué son las memorias y cómo se consolidan, o cómo el cerebro genera acciones y comportamientos.

Lograr este tipo de descripción matemática requiere, muchas veces, abstraerse de ciertos detalles (por ejemplo, a veces se asume que todas las neuronas son “iguales”, aunque realmente no lo son). Esta abstracción es la que permite emplear distintas herramientas matemáticas e ingenieriles, como teoría de circuitos eléctricos, teoría y control de sistemas dinámicos, teoría de la información, teoría de la computación, mecánica estadística y lingüísti-

ca computacional, entre muchas otras. Modelar aspectos del cerebro con estas herramientas nos permite, por ejemplo, simular el sistema en un computador, intervenir en él y analizar las consecuencias de modificar ciertas variables para entender cómo y por qué funciona. Finalmente, el formalismo matemático puede traducirse, y muchas veces coincide, con el utilizado en inteligencia artificial y aprendizaje automático en general. Existe, en otras palabras, un lenguaje común entre la neurociencia computacional y el aprendizaje automático que convierte esa intersección en un espacio fructífero para entender la inteligencia como un fenómeno natural: uno que puede observarse tanto en el cerebro como en agentes de inteligencia artificial, quizás regido por principios o leyes similares. Existe un debate enorme acerca de qué significa *comprender* un sistema complejo como el cerebro o un agente artificial, pero volveremos a esta discusión al final del artículo.

En las siguientes secciones hablaremos con más detalle de cómo funciona el cerebro y de cómo se modelan muchos de estos sistemas en neurociencia computacional. Comenzaremos de menos a más: desde las unidades de cómputo, las neuronas; pasaremos por la neuroplasticidad; luego a entender redes más grandes; y finalmente las partes y funciones más complejas del cerebro, como la toma de decisiones, el aprendizaje y la memoria.

Neuronas como unidad de cómputo del cerebro

El cerebro está hecho de células, como el resto del cuerpo. Sin embargo, la mayoría de las células están diseñadas para construir tejidos, transportar oxígeno o realizar funciones químicas específicas. Las neuronas son diferentes: están especializadas en la comunicación.

Nuestra comprensión moderna de las neuronas comenzó a finales del siglo XIX. Camillo Golgi desarrolló una técnica de tinción que colorea al azar una pequeña fracción de neuronas, permitiendo observar su forma completa por primera vez. Al estudiar estas estructuras ramificadas, Golgi concluyó que el cerebro era una red continua, similar a una malla de fibras fusionadas. Santiago Ramón y Cajal utilizó el mismo método, mostrando que las neuronas son células individuales que se comunican entre sí en puntos de contacto especializados. Esta idea, conocida como la *doctrina de la neurona*, es uno de los fundamentos de la neurociencia moderna.

En términos simples, una neurona es una célula cuyo trabajo es *recibir señales, combinarlas y transmitir nuevas señales*. Estructuralmente, una neurona tiene tres partes principales (ver Figura 1):

- **Dendritas:** estructuras ramificadas que reciben señales de otras neuronas.
- **Cuerpo celular (soma):** integra las señales que llegan desde las dendritas.
- **Axón:** una prolongación que transmite señales hacia otras neuronas.

Una neurona puede conectarse con miles de otras neuronas y recibir conexiones de muchas más. Las neuronas no se tocan directamente. Se comunican a través de pequeñas uniones llamadas *sinapsis*. Está formada por el extremo del axón de la neurona emisora, un pequeño espacio entre las células y una región de la membrana de la neurona receptora.

La transmisión de información dentro de una neurona entre neuronas ocurre de forma química y eléctrica. En una misma neurona, existen distintos iones con carga eléctrica que se desplazan a través de la membrana celular, como sodio, potasio, calcio y cloro. La membrana de la neurona regula qué iones pueden entrar o salir, generando una diferencia de voltaje entre el interior y el exterior de la célula. Este voltaje se conoce como *potencial de membrana*. Cuando la neurona está en reposo, este voltaje se mantiene relativamente estable. Sin embargo, las señales que recibe de otras neuronas pueden modificarlo ligeramente. Algunas lo aumentan y otras lo disminuyen.

Si suficientes señales excitatorias llegan en un corto intervalo de tiempo, el voltaje puede superar un umbral crítico. Cuando eso ocurre, la neurona produce un pulso eléctrico breve llamado *potencial de acción*, o simplemente *spike*. Un *spike* es un evento muy rápido, dura cerca

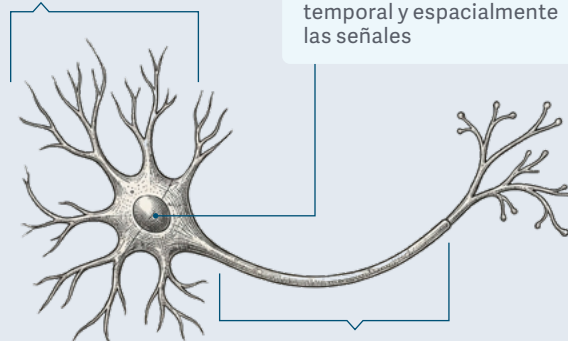
El hardware: La neurona

Dendritas

Input. Estructuras ramificadas que reciben señales de otras neuronas.

Soma

Procesador. El cuerpo celular que integra temporal y espacialmente las señales



Axón

Output. Prolongación que transmite la señal procesada hacia la red

El cerebro humano contiene ~ 86 mil millones de neuronas y cientos de billones de conexiones.

Basado en la **Doctrina de la Neurona** (Santiago Ramón y Cajal): son unidades discretas, no una malla continua.

Figura 1 / La neurona: las dendritas reciben señales de otras neuronas conectadas a ella. Dependiendo de la magnitud de ese estímulo, la neurona puede generar un potencial de acción o spike desde el Soma. Este se propaga a lo largo del axón hasta las vesículas sinápticas en su terminal, donde se libera el neurotransmisor que transmite la señal a las neuronas siguientes, formando la sinapsis.

de un milisegundo y funciona bajo un principio de *todo o nada*: ocurre o no ocurre. Por esta razón, se parece más a un pulso digital que a una señal analógica. La información no está en la amplitud del pulso, sino en *cuándo ocurre* y *con qué frecuencia*.

Modelos de una neurona

Como mencionamos anteriormente, uno de los primeros esfuerzos de la neurociencia computacional fue desarrollar

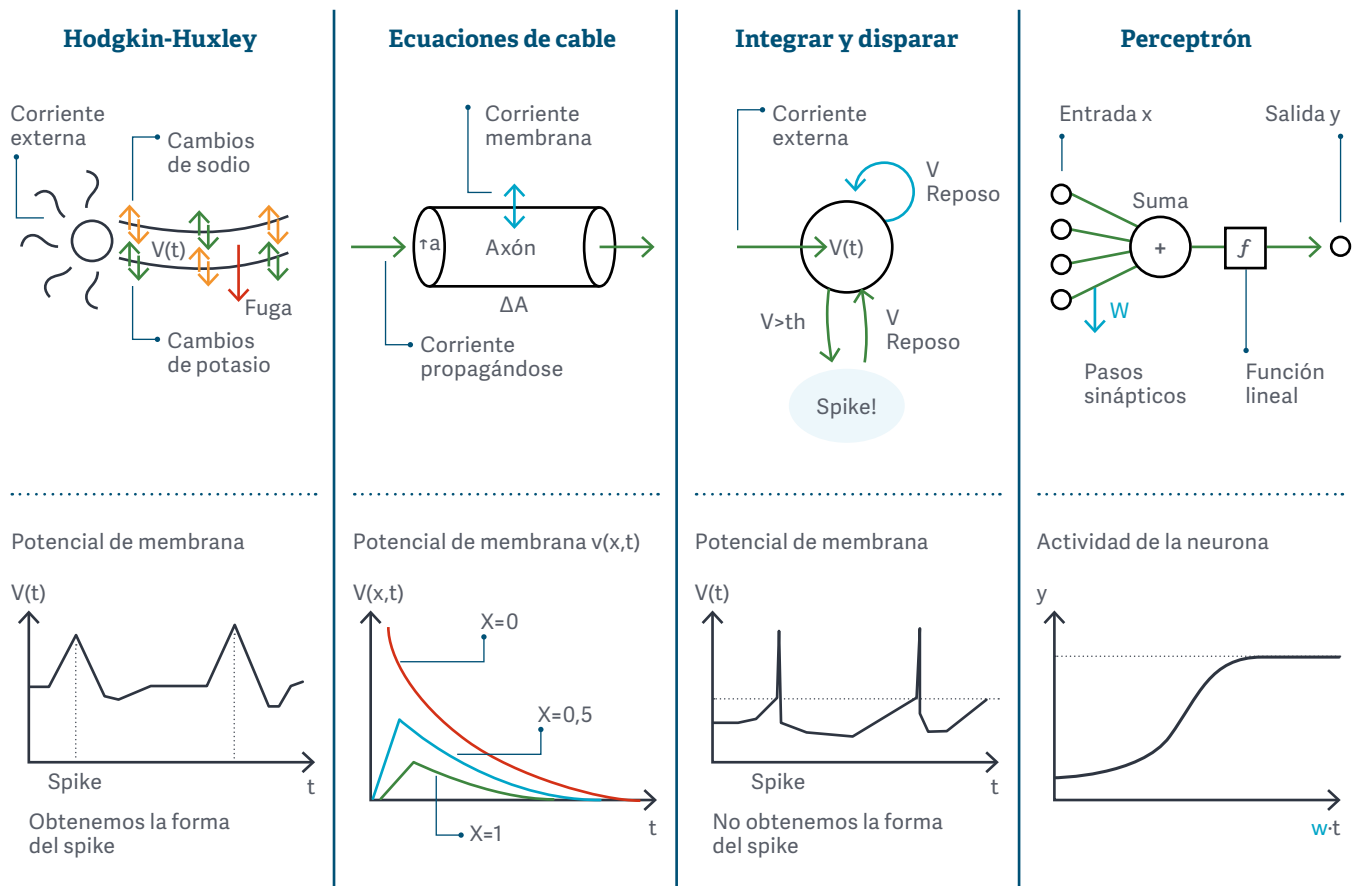


Figura 2 / Modelos matemáticos del comportamiento de una neurona. **Hodgkin-Huxley:** sistema de ecuaciones diferenciales acopladas que describe el potencial de membrana mediante el modelado de compuertas iónicas de sodio y potasio, y una corriente de fuga. Es el modelo más detallado biofísicamente y reproduce la forma precisa de los spikes. **Ecuación de cable:** describe cómo se propaga el potencial a lo largo del axón, expresándolo en función de la posición y el tiempo. Es útil cuando la geometría espacial de la neurona importa. **Integrar y disparar:** el potencial de membrana tiende al reposo, pero aumenta al recibir una corriente externa; si alcanza un umbral, se emite un spike y el potencial se resetea al valor de reposo. A diferencia de Hodgkin-Huxley, el spike no tiene forma biofísica explícita sino que se define como un evento discreto. **Perceptrón:** abstrae toda restricción biológica y modela la salida de otras neuronas como tasas de disparo continuas, las pondera por pesos sinápticos, las suma y aplica una función de activación no lineal. Es el modelo más simplificado y el que da lugar a las redes neuronales artificiales.

modelos matemáticos que describieran el comportamiento eléctrico de una neurona (ver Figura 2).

Ecuaciones de Hodgkin-Huxley: Las ecuaciones de Hodgkin-Huxley (HH) [1] describen cómo fluyen los iones de sodio y potasio a través de los canales en la membrana neuronal. Se construyen usando teoría de circuitos eléctricos: el potencial de membrana se modela como un capacitor, y el flujo de iones como corrientes que entran y salen de él, formando un sistema de ecuaciones diferenciales acopladas. Son computacionalmente costosas y tienen muchos parámetros, pero describen con gran precisión el comportamiento eléctrico de la neurona, incluyendo la forma y dinámica de los spikes.

Ecuaciones de cable: Las neuronas pueden tener geometrías complejas, especialmente en el axón y las dendritas. Las ecuaciones de cable modelan la neurona como un cilindro con resistencia interna y capacitancia de membrana, describiendo cómo el voltaje se propaga y decae a lo largo de la estructura mediante una ecuación diferencial parcial en espacio y tiempo [2]. Son útiles cuando la morfología importa; por ejemplo, para entender cómo una señal en una dendrita distal llega atenuada al soma.

Integrar-y-Disparar: El modelo Integrar-y-Disparar (LIF, por sus siglas en inglés) puede entenderse como una versión simplificada de HH [3]. El potencial de la neurona tiende al reposo, pero aumenta al recibir una corriente externa. Una

vez que alcanza un umbral, se emite un *spike* reseteando el voltaje manualmente, sin modelar explícitamente los canales iónicos. Esto simplifica considerablemente la dinámica y reduce el costo computacional.

Neurona artificial (Perceptrón): El perceptrón es el modelo más abstracto de todos. Simplemente suma las señales entrantes ponderadas y las pasa por una función de activación no lineal, sin modelar canales iónicos, propagación espacial ni *spikes*. Esta simplicidad es precisamente su fortaleza: permite escribir matemáticamente redes de millones de neuronas conectadas, analizarlas y simularlas en un computador convencional. En contraste, simular una red equivalente con neuronas HH requeriría un poder de cómputo enormemente mayor para apenas cientos o miles de neuronas. El costo de esta abstracción es la pérdida de la descripción biofísica de las neuronas: para replicar el comportamiento de una neurona biológica con todo su detalle, se requieren muchas neuronas perceptrón [4, 5]. Muchos modelos modernos de redes neuronales asumen este tipo de neuronas simplificadas, y es un área de investigación activa entender qué efectos no pueden observarse en redes debido a esta simplificación [6, 7].

Para entender cómo las neuronas se influyen entre sí, debemos observar con más detalle los espacios de contacto entre ellas, llamados sinapsis, que juegan un rol fundamental en el aprendizaje.

Sinapsis

Cuando un *spike* llega al final del axón, desencadena un proceso llamado *transmisión sináptica*. Pequeñas vesículas liberan moléculas llamadas neurotransmisores en el espacio entre las neuronas. Estas moléculas atraviesan la brecha entre neuronas en una fracción de milisegundo y se unen a receptores en la neurona siguiente (ver Figura 3). Al activarse, estos receptores modifican las propiedades eléctricas de la membrana de la neurona receptora [8].

El resultado es un pequeño cambio en el voltaje de esa neurona. Algunas sinapsis aumentan el voltaje y hacen más probable que la neurona dispare. Estas se conocen como *sinapsis excitatorias*. Otras lo disminuyen y reducen la probabilidad de disparo. Estas son *sinapsis inhibitorias* [9].

Una sinapsis individual suele ser muy débil. Rara vez provoca por sí sola que una neurona dispare. Sin embargo, cada neurona recibe miles de señales de este tipo. La neurona integra todos estos pequeños cambios de voltaje a lo largo del tiempo y del espacio. Como vimos anteriormente en los modelos HH e integrar-y-disparar, la neurona genera un *spike*. Si no, permanece en potencial de reposo.

La interfaz de red: Sinapsis

Una sinapsis individual es demasiado débil para forzar una acción. La computación real ocurre al integrar miles de micro-señales en el cerebro. Las neuronas modulan continuamente la probabilidad de que otras neuronas se activen.

→ **Excitatorias (+):** Aumentan el voltaje, acercando al umbral de disparo.

→ **Inhibitorias (-):** Disminuyen el voltaje, suprimiendo el disparo.

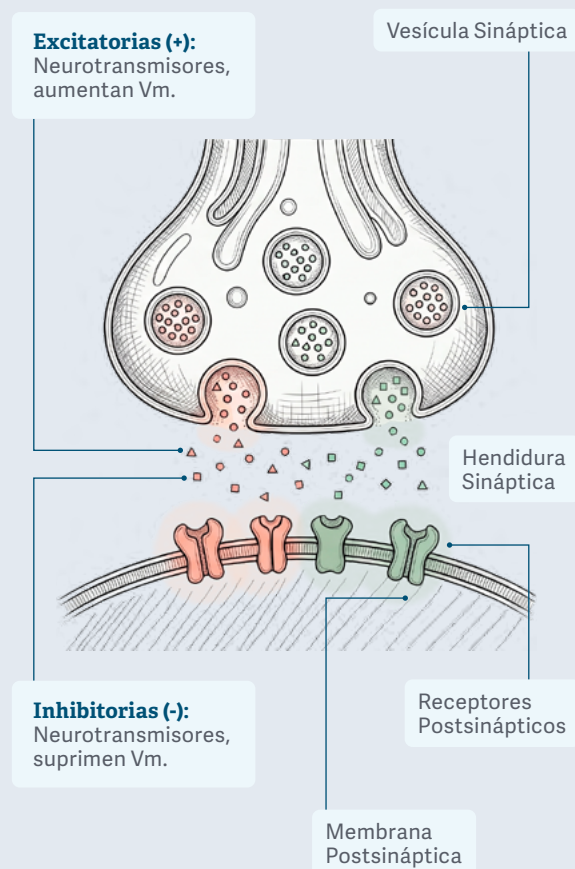


Figura 3 / Sinapsis: espacio de conexión entre neuronas.

Cuando la neurona presináptica genera un *spike*, libera neurotransmisores desde la vesícula sináptica hacia la neurona postsináptica, modificando su potencial de membrana. Los neurotransmisores pueden ser excitatorios, acercando el potencial al umbral de disparo, o inhibitorios, alejándolo. Típicamente una sinapsis es enteramente excitatoria o inhibitoria, aunque existe evidencia de sinapsis que liberan ambos tipos, como la que se muestra en este esquema [10].



La plasticidad sináptica es uno de los principales mecanismos que permiten aprender cosas nuevas, pero no es el único.

Plasticidad

Las conexiones entre neuronas no son fijas. Cambian con la experiencia. Esta propiedad se conoce como *plasticidad*. Cuando grupos de neuronas se activan juntos de manera repetida, las sinapsis entre ellas pueden fortalecerse. Cuando la actividad no está coordinada, esas conexiones pueden debilitarse. Con el tiempo, estos pequeños ajustes modifican cómo fluye la actividad a través de la red.

Una de las ideas más influyentes para describir este proceso se resume en una frase simple:

“Las neuronas que se activan juntas tienden a fortalecer su conexión”.

Este principio se conoce como *aprendizaje hebbiano* [11], propuesto por Donald Hebb en los años cuarenta. Décadas de experimentos han identificado mecanismos biológicos que implementan esta idea. Uno de los más estudiados es la *plasticidad dependiente del tiempo de disparo* (STDP) [12, 13], donde el momento preciso en que dos neuronas disparan determina si su conexión se fortalece o se debilita. A nivel físico, estos cambios implican modificaciones en la sinapsis: variaciones en la cantidad de receptores, en la liberación de neurotransmisores, en el tamaño de la conexión o incluso en el número de sinapsis entre neuronas.

Con el tiempo, estos cambios microscópicos reorganizan los patrones de actividad en la red. Grupos de neuronas comienzan a activarse de manera consistente juntos, formando lo que los neurocientíficos llaman *circuitos o redes neuronales*. Así, el aprendizaje no se almacena en una neurona específica ni en un registro centralizado. Se distribuye en los patrones de conexión entre muchas neuronas.

Un ejemplo simple es el reconocimiento de un rostro. La primera vez que vemos una cara nueva, la actividad en el sistema visual es débil e inestable. Con la repetición, las neuronas que representan sus rasgos tienden a activarse juntas, sus sinapsis se fortalecen y el patrón se estabiliza. Con el tiempo, basta una pequeña parte de la señal para activar el resto del patrón, y reconocemos a la persona casi instantáneamente. A través de miles de millones de estos cambios microscópicos, el cerebro ajusta continuamente su estructura para interpretar mejor el mundo y guiar la acción.

La plasticidad sináptica es uno de los principales mecanismos que permiten aprender cosas nuevas, pero no es el único. Otros ejemplos incluyen la plasticidad intrínseca (cambios en la excitabilidad interna de la neurona) [14], la neurogénesis (creación de neuronas nuevas, especialmente en el hipocampo) [15], y la modulación por neurotransmisores como la dopamina, la serotonina o la acetilcolina [16].

Si bien se conocen varios mecanismos locales de plasticidad sináptica, como el aprendizaje hebbiano, sigue siendo una incógnita cómo el cerebro aprende a nivel global, coordinando el cambio de muchas neuronas a la vez. Para entender esto mejor, veamos qué ocurre en las redes neuronales.

Redes neuronales artificiales y biológicas

El cerebro humano contiene alrededor de 86 mil millones de neuronas de distintos tipos y cientos de billones de conexiones entre ellas. En esta sección aplicaremos una abstracción adicional para discutir modelos de redes neuronales, que sacrifican detalle biológico a cambio de simplicidad y mayor claridad conceptual. Adoptaremos los siguientes supuestos:

- **Un solo tipo de neurona simplificada.** Supondremos que todas las neuronas siguen el mismo modelo, en particular el perceptrón, donde la salida de una neurona se conecta como entrada a otra. Existen modelos de redes neuronales donde este no es el caso [17].
- **Salida continua.** Supondremos que la salida de cada neurona es un número continuo, interpretable como análogo a la frecuencia de disparo de spikes [18].
- **Las neuronas como único sustrato de cómputo.** Supondremos que todo el procesamiento ocurre en las neuronas, dejando de lado otras células como los astrocitos y distintos tipos de glía. Aunque durante mucho tiempo se consideraron principalmente células de soporte, hoy se investiga activamente su posible rol en la modulación de la actividad neuronal y en procesos de plasticidad [19, 20].

Como consecuencia de estas simplificaciones, cada neurona individual no es capaz de resolver problemas complejos por sí sola: es el sistema en su conjunto el que tiene la capacidad de procesar y responder a información compleja. En otras palabras, las redes neuronales son sistemas distribuidos donde la capacidad computacional emerge de la cooperación de muchas neuronas.

Quizás la forma más sencilla de conectar neuronas para formar una red es una red neuronal de dos capas (ver Figura 4): una capa de entrada, que recibe las señales, y una capa de

salida, que produce la respuesta. Este tipo de arquitectura ha sido utilizada tanto en neurociencia como en aprendizaje automático debido a su simplicidad y a su capacidad para resolver problemas complejos, ya que es un aproximador universal de funciones. Luego veremos que existen redes con muchas capas, y con conectividades más complejas.

Si bien el supuesto de que cada neurona es estructuralmente similar es razonable, su conectividad resulta ser más específica dependiendo del área del cerebro que estemos estudiando. Esto es importante porque la conectividad pre-define qué neuronas están conectadas entre sí, y son esas conexiones las que están sujetas a plasticidad durante el aprendizaje. A continuación describiremos diferentes regiones del cerebro que cumplen distintos roles y que, por tanto, presentan arquitecturas de conectividad distintas según las demandas de cada función. Finalmente, discutiremos el aprendizaje en redes neuronales biológicas: cómo las reglas locales como la hebbiana o la STDP, si bien capturan aspectos importantes de la plasticidad, resultan insuficientes para explicar el aprendizaje en redes grandes.

Organización del cerebro

Aunque el cerebro parece un solo órgano, está compuesto por múltiples regiones con distintos tipos de neuronas, patrones de conectividad y funciones. Una forma útil de organizarlas es distinguir tres grandes sistemas presentes en todos los mamíferos [21].

El *tronco encefálico*, ubicado en la base del cerebro, es una estructura evolutivamente antigua que controla funciones esenciales como la respiración y el ritmo cardíaco. Computacionalmente, su rol principal es mantener el sistema activo y regular su estado global mediante estructuras como la formación reticular, que modula el nivel de alerta.

Rodeando el tronco encefálico, el *sistema límbico* incluye la amígdala, el hipocampo y el hipotálamo. Estas regiones están involucradas en emoción, motivación y memoria, y contribuyen señales de valor y refuerzo que influyen en qué experiencias se aprenden y qué conductas se repiten.

Cubriendo la superficie del cerebro, el *neocórtex* es el principal sistema para la percepción, representación del mundo, planificación y control voluntario de acciones. Todos los mamíferos la poseen, pero en humanos está especialmente expandida en áreas asociadas al lenguaje, la abstracción y la toma de decisiones.

Además de estos tres sistemas, algunas estructuras cumplen roles especialmente relevantes para el aprendizaje: el

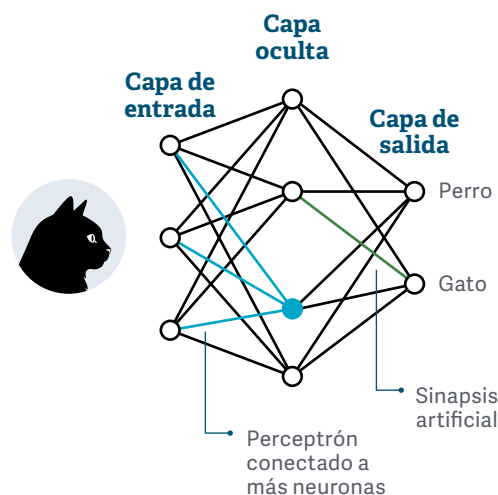


Figura 4 / Red neuronal artificial: compuesta por capas de neuronas, en este caso tres: entrada, oculta y salida, donde cada neurona (perceptrón) recibe como entrada las salidas de la capa anterior, escaladas por pesos sinápticos, que son las conexiones entre neuronas.

tálamo regula qué información sensorial llega a la corteza y cuándo; el *hipocampo* forma nuevas memorias episódicas y participa en su consolidación de conocimiento en el neocórtex; el *cerebelo* está involucrado en predicción y aprendizaje basado en error; y los *ganglios basales* participan en la selección de acciones y en aprendizaje por refuerzo. A continuación exploraremos algunas de estas estructuras, junto con modelos de redes neuronales que se han propuesto para describir su actividad.

El neocórtex: la principal máquina de aprendizaje del cerebro

El neocórtex es probablemente la estructura más notable del cerebro. A pesar de participar en funciones tan diversas como visión, audición, lenguaje, planificación motora y razonamiento abstracto, su organización interna es sorprendentemente uniforme. En casi todas sus áreas se repite el mismo microcircuito de seis capas [22].

Bajo el microscopio, la corteza visual, auditiva, motora y prefrontal se ven muy similares. Lo que cambia no es la estructura básica del circuito, sino los tipos de señales que reciben y las proyecciones que envían a otras regiones del cerebro.

Esto sugiere una idea importante desde el punto de vista computacional: el neocórtex podría implementar un mecanismo general de aprendizaje que luego se especializa según sus conexiones y los datos que recibe. En lugar de tener

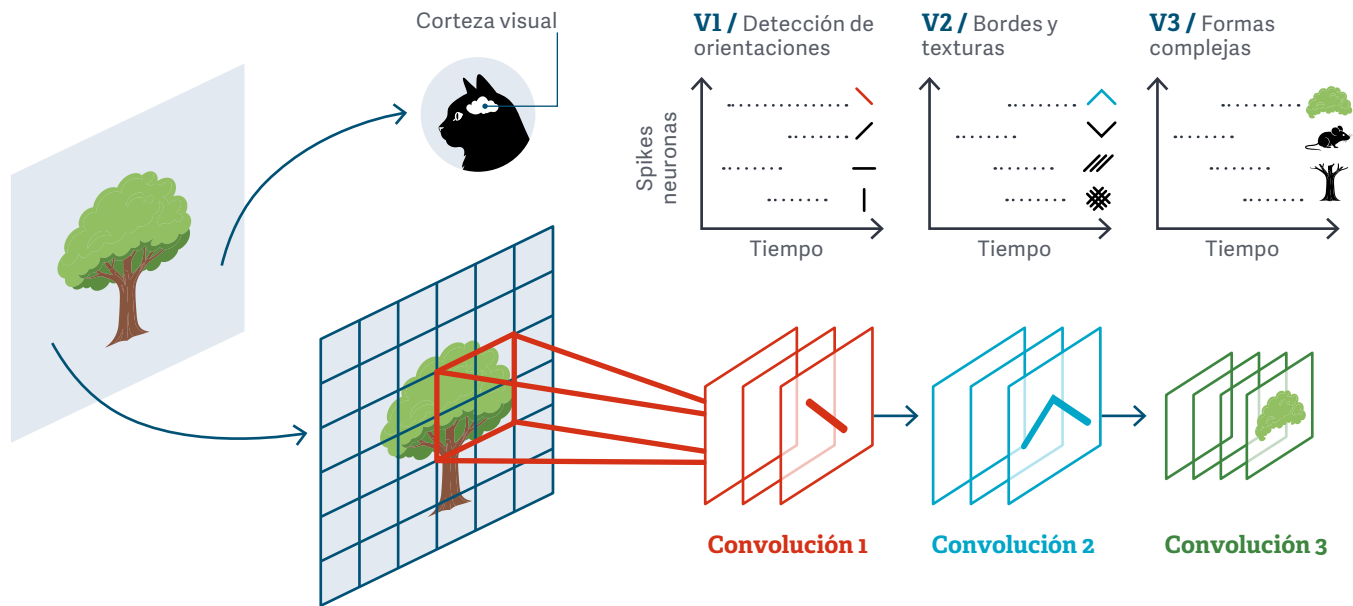


Figura 5 / Red convolucional como modelo del sistema visual: las redes convolucionales son uno de los modelos más utilizados para describir el sistema visual en mamíferos. En estas redes, cada neurona actúa como un filtro local que recorre la imagen de entrada y genera representaciones de lo que detecta. Estas representaciones tienen mucho en común con las observadas en la corteza visual: a medida que se avanza por las capas de la red, las representaciones progresan de características simples, como bordes y esquinas, hacia formas y texturas más abstractas, reflejando la jerarquía visual del cerebro. Esta correspondencia entre redes convolucionales y corteza visual puede explorarse directamente en plataformas como BrainScore, que permiten comparar modelos computacionales con registros neuronales reales.

módulos completamente distintos para visión, lenguaje o razonamiento, el cerebro reutiliza una arquitectura común que se entrena con diferentes flujos de información.

Pequeñas regiones del neocórtex, a veces llamadas columnas corticales, pueden entenderse como unidades locales de aprendizaje [23]. Cada una recibe señales de entrada, detecta regularidades y genera predicciones sobre lo que es probable observar a continuación. Estas predicciones se propagan tanto hacia áreas superiores como hacia áreas inferiores de la jerarquía cortical, generando un sistema fuertemente recursivo.

Esta organización nos lleva naturalmente a una forma moderna de entender la percepción.

Corteza visual

El sistema visual es quizás el ejemplo más estudiado de la conexión entre neurociencia computacional y aprendizaje automático. Hubel y Wiesel [24] descubrieron en la década de 1960 que las neuronas de la corteza visual

primaria V1 (de un gato) responden selectivamente a bordes orientados en posiciones específicas del campo visual, comportándose esencialmente como filtros locales. Esta observación inspiró a Fukushima [25] a proponer el Neocognitron, una red jerárquica que detectaba patrones visuales de forma progresivamente más abstracta. Más tarde, LeCun et al. [26] formalizaron esta idea en las redes convolucionales (CNNs), demostrando su utilidad práctica en tareas de clasificación.

En una red convolucional, las neuronas analizan localmente distintas partes de la imagen de entrada, formando filtros que recuerdan a los detectores de bordes de Hubel y Wiesel. En el cerebro, el proceso análogo comienza en la retina, cuya actividad se transmite a través del tálamo hacia V1, y luego a áreas visuales de mayor nivel como V2 y la corteza inferotemporal (IT), que representan información visual cada vez más abstracta. Esta jerarquía biológica se refleja naturalmente en la arquitectura de capas de una CNN (ver Figura 5).

La similitud entre ambos sistemas ha dado lugar a comparaciones directas. Plataformas como BrainScore¹ permiten

1 <https://www.brain-score.org/>

evaluar qué tan bien los modelos computacionales reproducen grabaciones neuronales reales de la corteza visual. Hoy existen modelos con distintos grados de realismo biológico, contruidos con propósitos diversos: clasificación de imágenes, compresión, o maximización de la transmisión de información, este último con una conexión más directa a teorías de codificación eficiente en el cerebro.

Cabe destacar que no todo el modelamiento del sistema visual se hace con redes neuronales. Un ejemplo muy influyente es el trabajo de Olshausen y Field [27], quienes mostraron que los filtros observados en V1 pueden obtenerse a partir de un principio teórico: si se entrena un modelo para representar imágenes naturales de forma eficiente y dispersa (*sparse coding*), los filtros que emergen del entrenamiento se parecen notablemente a los campos receptivos de las células simples de V1. Esta idea se enmarca en la teoría de *efficient coding*, cuya formulación original se atribuye a Horace Barlow [28, 29], quien propuso que el sistema nervioso sensorial está organizado para transmitir la mayor cantidad de información posible minimizando la redundancia de información propagada por los *spikes*. La teoría ha sido aplicada con éxito más allá del sistema visual, incluyendo el sistema auditivo y otras áreas sensoriales, y sigue siendo un marco activo de investigación en neurociencia computacional.

Corteza prefrontal

La corteza prefrontal (PFC, del inglés *prefrontal cortex*) es una de las áreas más complejas del cerebro de los mamíferos. Su complejidad se debe en parte a que sus neuronas participan en una gran variedad de tareas. Muchas regiones del cerebro proyectan su actividad hacia la PFC, que funciona como un orquestador capaz de integrar información de distintas fuentes para guiar la interacción con el ambiente. Una de sus capacidades más estudiadas es el mantenimiento de información contextual a lo largo del tiempo, conocida como memoria de trabajo.

Debido a su acceso a información de múltiples fuentes, la PFC es frecuentemente asociada a capacidades como el pensamiento abstracto, el lenguaje y la planificación. Dada la diversidad de funciones que se le atribuyen, existen distintos modelos computacionales del PFC, cada uno enfocado en explicar aspectos diferentes de estas capacidades.

Uno de ellos es el trabajo de Wang et al. [30, 31], que propone una red recurrente como modelo del PFC. En una red recurrente, parte de la actividad de las neuronas en un momento dado se retroalimenta como entrada a sí mismas y a otras neuronas en el momento siguiente, lo que les permite mantener una memoria de corto plazo en su actividad (ver Figura 6). El modelo de Wang busca describir la capa-

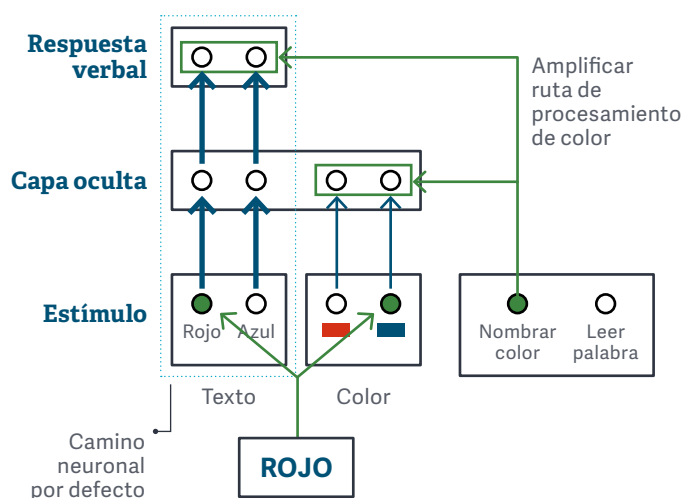
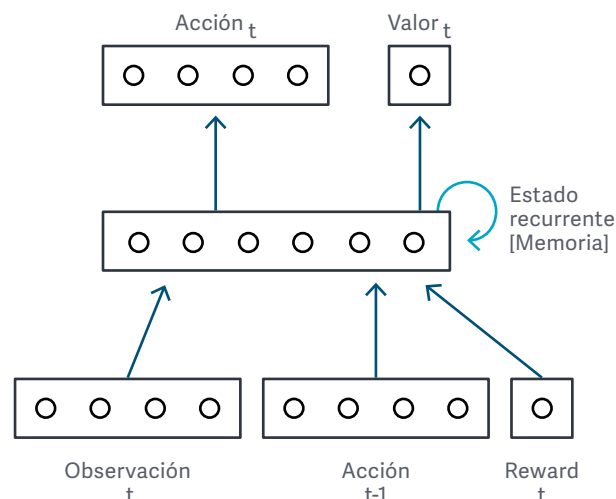


Figura 6 / Arriba: corteza prefrontal como meta-aprendizaje. El modelo propuesto por Wang et al. (2018) es una red recurrente entrenada con aprendizaje por refuerzo en múltiples tareas. La recurrencia permite que la actividad de las neuronas en un instante sirva como entrada en el siguiente, formando una memoria a corto plazo. Esta memoria le permite a la red acumular evidencia del ambiente, inferir en qué tarea se encuentra y aprender por prueba y error sin modificar sus pesos sinápticos. **Abajo: modelo de Stroop** (Cohen et al., 1990). La red modela la tarea Stroop, en la que se debe leer una palabra o mencionar el color de las letras en que está escrita (por ejemplo **ROJO**, donde la respuesta a leer el texto es “rojo”, y la respuesta a decir el color es “azul”). Las personas responden más rápido leyendo que nombrando el color, porque la práctica acumulada de lectura hace que esa ruta neuronal sea más fuerte (ruta por defecto). Sin embargo, una señal de contexto puede amplificar rutas alternativas (utilizando las conexiones en verde) dependiendo de la instrucción (nombrar el color o leer la palabra), permitiendo a la red resolver ambas variantes de la tarea.



Adaptado de Kaufman et al. 2014

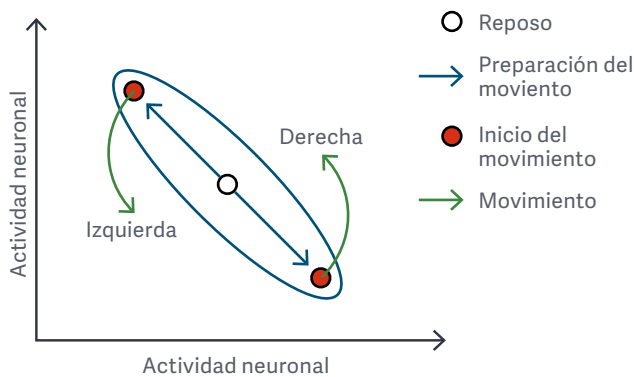


Figura 7 / Espacio de preparación motora en la corteza motora: antes de ejecutar un movimiento, la actividad neuronal evoluciona hacia una condición inicial específica dentro del subespacio nulo (zona azul), donde los cambios de actividad no producen movimiento muscular. Una vez configurada esta condición inicial (puntos rojos), el movimiento se desencadena al proyectarse la actividad hacia el subespacio de salida (fuera de la zona azul), donde sí genera movimiento. La trayectoria resultante (flechas verdes) está determinada por las condiciones iniciales establecidas durante la preparación. Este mecanismo se formaliza matemáticamente con redes recurrentes y teoría de control óptimo [39].

ciudad del PFC de *aprender a aprender*, también llamada meta-aprendizaje [32, 33]. La idea central es que los pesos sinápticos de la red son modificados lentamente a través del entrenamiento con aprendizaje por refuerzo, un proceso análogo a la plasticidad sináptica que ocurre a lo largo de la vida. Sin embargo, una vez entrenada, la red no necesita modificar sus pesos para adaptarse a una tarea nueva: su dinámica recurrente, es decir, la forma en que la actividad de las neuronas evoluciona en el tiempo, implementa por sí sola un algoritmo de aprendizaje rápido. En otras palabras, la red es capaz de aprender tareas nuevas sin modificar sus pesos. La recurrencia cumple aquí el rol de memoria: le permite acumular evidencia del ambiente para inferir el contexto en el que se encuentra y resolver el problema apropiado.

Junto con mantener una representación del contexto, el PFC puede modular la actividad de otras regiones del cerebro [34]. Un modelo clásico de esto es el modelo de Stroop propuesto por Cohen et al. [35], donde señales de control amplifican o atenúan distintas rutas de procesamiento en una red. Este modelo fue diseñado para explicar la tarea Stroop, en la que una persona ve una palabra escrita en un color determinado y debe, dependiendo de la instrucción, decir la palabra escrita o el color de las letras. Por ejemplo, si aparece la palabra **ROJO** (escrita en azul), la respuesta co-

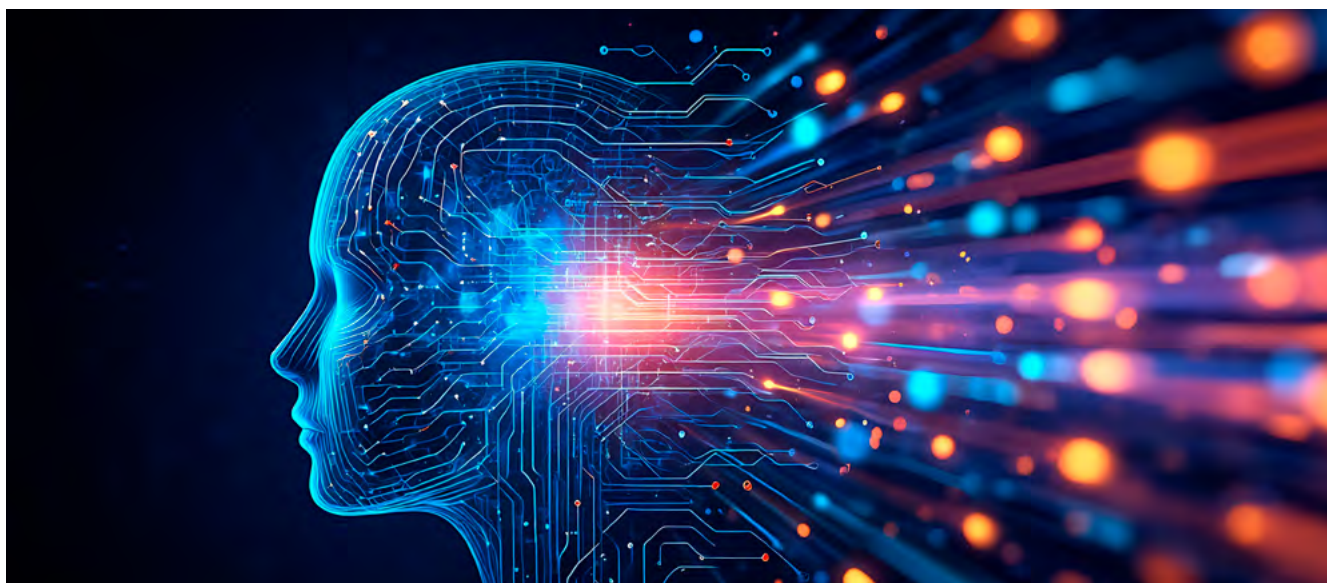
rrecta es 'azul' si la tarea es nombrar el color (ver Figura 6). Las personas pueden resolver esta tarea, pero tardan más cuando hay conflicto entre la palabra y el color. La hipótesis es que la ruta de lectura es más fuerte por práctica acumulada, y la señal de control debe amplificar activamente la ruta del color para superar esa ventaja [36].

Corteza motora

La corteza motora es la región principal responsable de la ejecución de movimientos voluntarios. No trabaja sola: está conectada con la corteza premotora y el área motora suplementaria, que planifican el movimiento antes de ejecutarlo, y con el cerebelo y los ganglios basales, que lo refinan y seleccionan [37]. Una observación clave es que la actividad neuronal en esta región comienza a evolucionar antes de que el movimiento ocurra, durante un período llamado preparación motora.

Una forma de entender este período de preparación es a través del modelo de subespacios de Churchland et al. [38, 39]. La idea central es que la corteza motora funciona como un sistema dinámico: preparar un movimiento consiste en llevar la actividad de la población neuronal a una condición inicial correcta en un espacio de alta dimensión, y el movimiento emerge como la trayectoria que el sistema recorre desde ahí. Un hallazgo elegante de este marco es la existencia de dos subespacios en la actividad neuronal: uno nulo, donde los cambios en actividad no producen movimiento muscular, y uno de salida, donde sí lo producen. Durante la preparación, la actividad evoluciona en el subespacio nulo, lo que permite al cerebro preparar el sistema dinámico antes del movimiento sin activar los músculos prematuramente. Cuando llega la señal de ejecutar, la actividad se proyecta al subespacio de salida y el movimiento ocurre. Este modelo se formaliza matemáticamente como una red recurrente cuya salida muscular es una proyección lineal de la actividad neuronal (ver Figura 7).

Un enfoque complementario es el del control óptimo [40], que trata la corteza motora no como un sistema dinámico sino como un controlador que resuelve un problema de optimización: ejecutar el movimiento deseado minimizando el esfuerzo muscular. Este marco predice bien propiedades observadas en movimientos reales, como la suavidad de las trayectorias y la forma en que la variabilidad se distribuye a lo largo del movimiento. A diferencia del modelo de subespacios, que describe la dinámica neuronal interna, el control óptimo describe el movimiento desde una perspectiva normativa: no cómo lo implementa el cerebro, sino qué problema está resolviendo. Ambos enfoques son complementarios y juntos ofrecen una descripción más completa de cómo la corteza motora planifica y ejecuta movimientos.



Acción y refuerzo

El cerebro puede entenderse, en gran medida, como un sistema de control: no sólo reacciona a estímulos, sino que genera predicciones sobre las consecuencias de sus acciones, selecciona comportamientos según metas y valores, y compara los resultados con lo que esperaba.

¿Cómo selecciona el cerebro qué acción tomar? Una forma de entender esto es a través del aprendizaje por refuerzo (RL del inglés Reinforcement Learning), un marco teórico en el que un agente aprende qué hacer a partir de señales de recompensa [41]. Imagina que estás aprendiendo a jugar ajedrez: algunos movimientos son mejores que otros a corto y largo plazo, y aprender eso requiere práctica. En principio, uno podría probar movimientos al azar y observar qué ocurre, asignando gradualmente un valor a cada jugada según sus consecuencias.

Uno de los hallazgos más importantes de la neurociencia es la conexión entre la dopamina y el error de predicción de recompensa. Si hago una jugada que creo será excelente y resulta ser un desastre, esperaba un valor alto y obtuve uno bajo: ese error de predicción está codificado por la liberación de dopamina. Niveles altos de dopamina señalan que el resultado fue mejor de lo esperado; niveles bajos, que fue peor [42, 43]. Esta señal es lo que permite al cerebro ajustar el valor de las acciones con la experiencia, y puede ser modelada matemáticamente utilizando la ecuación de Bellman de aprendizaje reforzado [44, 41].

La dopamina no sólo codifica errores de predicción: también responde a estímulos novedosos y sorprendentes, y está

Niveles altos de dopamina señalan que el resultado fue mejor de lo esperado [...] Esto explica su papel en la adicción: sustancias [...] que elevan artificialmente la dopamina generan una señal [...] que el cerebro interpreta como algo valioso de repetir.

involucrada en estados motivacionales y de atención, junto con otros neuromoduladores como la noradrenalina, la acetilcolina y la serotonina [45, 46, 47, 48]. Esto explica su papel en la adicción: sustancias o comportamientos que elevan artificialmente la dopamina generan una señal de “resultado mejor de lo esperado” que el cerebro interpreta como algo valioso de repetir, alterando gradualmente los mecanismos de regulación y el aprendizaje mismo [49].

Aprendizaje en el cerebro

Como hemos visto, las redes neuronales del cerebro son diversas y trabajan en conjunto para producir comportamiento. Esto plantea una pregunta natural: ¿todas las neuronas aprenden de la misma forma? ¿cómo debería ajustarse una conexión en la corteza visual para mejorar en una tarea como el tenis, que involucra también la corteza motora y la PFC?

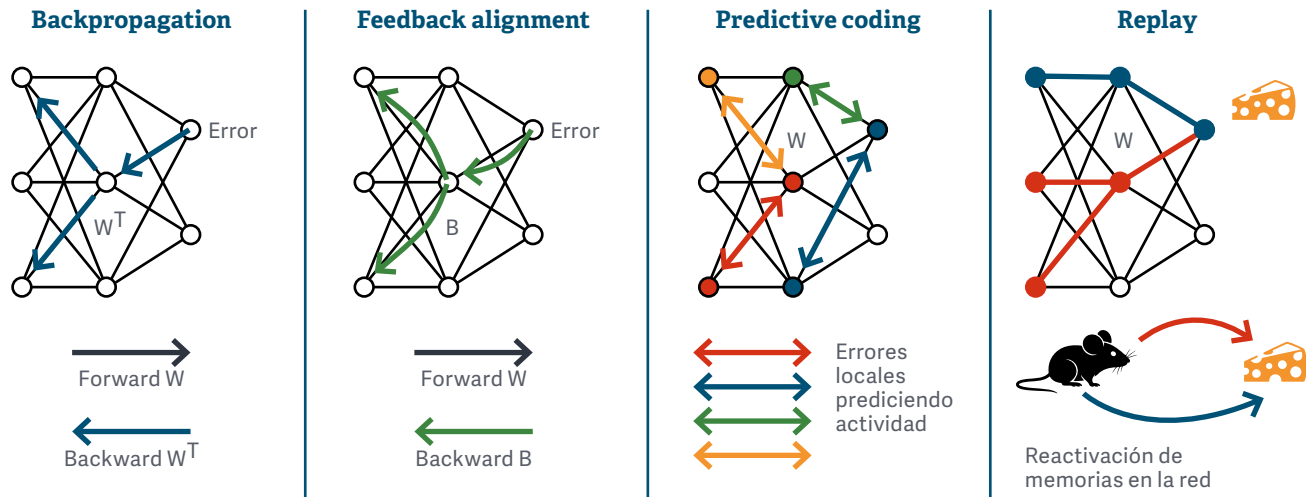


Figura 8 / Algoritmos de aprendizaje en redes neuronales artificiales y biológicas. **Backpropagation:** se calcula el error comparando la salida deseada con la entregada por la red (forward), y se retropropaga a través de los pesos sinápticos transpuestos para ajustar cada conexión (backward). Se piensa que el cerebro no puede implementar este cómputo debido a restricciones biológicas. **Feedback alignment:** similar al anterior, pero el error se retropropaga a través de pesos aleatorios fijos, distintos de los usados en el paso hacia adelante. El ajuste no es óptimo, pero el mecanismo es biológicamente plausible. **Predictive coding:** cada sinapsis se ajusta para predecir la actividad de la neurona siguiente en la red, sin necesidad de retropropagar el error explícitamente (proceso llamado configuración prospectiva). Puede alcanzar un rendimiento comparable al de backpropagation y es biológicamente plausible, aunque computacionalmente más costoso. **Replay:** no es una regla de ajuste sináptico sino un mecanismo de selección de contenido. Durante el descanso y el sueño, patrones de actividad asociados a experiencias previas se reactivan espontáneamente, consolidando el aprendizaje en las sinapsis.

El consenso actual es que el cerebro no puede implementar backpropagation, ya que requiere propagar información precisa a través de conexiones lejanas, algo biológicamente poco plausible.

Esta es una pregunta abierta, y el problema central se conoce como asignación de crédito (credit assignment) [50, 51]. Las reglas locales de plasticidad, como la hebbiana, sólo requieren información de neuronas vecinas, pero no explican bien el aprendizaje a gran escala ni funcionan bien en simulaciones.

Como mencionamos anteriormente, la *dopamina* provee una señal de retroalimentación que indica si el resultado de una acción fue mejor o peor de lo esperado. En términos más formales, esta señal corresponde al error de predicción de recompensa, y es precisamente este tipo de señal el que aparece como ingrediente central en muchas de las propuestas de reglas de plasticidad biológicamente plausibles. Sin embargo, como veremos a continuación, disponer de una señal de error global no resuelve por sí solo el pro-

blema de cómo ajustar cada una de las millones de conexiones involucradas en el aprendizaje (ver Figura 8).

Backpropagation

En inteligencia artificial, el problema de credit assignment está relativamente resuelto mediante la retropropagación del error (*backpropagation*): se calcula el error entre la respuesta de la red y la deseada (forward) y el error se propaga hacia atrás (backward) para ajustar todas las conexiones. Sin embargo, el consenso actual es que el cerebro no puede implementar *backpropagation*, ya que requiere propagar información precisa a través de conexiones lejanas, algo biológicamente poco plausible [52, 53].

Feedback alignment

Existen alternativas intermedias que intentan mantener cierta eficacia de aprendizaje global siendo al mismo tiempo biológicamente plausibles. Una de las más conocidas es el alineamiento de retroalimentación (*feedback alignment*). En *backpropagation* estándar, el error se propaga hacia atrás usando exactamente los mismos pesos que se usaron en el paso hacia adelante, lo que es biológicamente problemático porque requeriría que cada neurona “conoz-

ca” con precisión los pesos de conexiones lejanas. En *feedback alignment*, el error se propaga hacia atrás a través de un conjunto de sinapsis distinto, inicializado aleatoriamente y que no cambia durante el entrenamiento [54]. Sorprendentemente, la red aprende de todas formas, porque los pesos del paso hacia adelante se adaptan gradualmente para alinearse con estas sinapsis fijas. La ventaja biológica es clara: no se requiere transportar información precisa de la red hacia atrás, sino simplemente propagar alguna señal de error a través de conexiones independientes, algo más compatible con la anatomía neuronal conocida. El costo es que aprende más lento y con menor precisión que *backpropagation*, especialmente en tareas complejas.

Predictive coding

El *código predictivo* (*predictive coding*) parte de una idea diferente. En lugar de ajustar los pesos para minimizar directamente el error en la salida de la red, cada neurona ajusta sus conexiones para predecir la actividad de la neurona siguiente en la red. El error que se propaga no es el error global de la tarea, sino la diferencia entre lo que una neurona predijo y lo que la siguiente neurona realmente hizo [55]. A este proceso de ajuste simpático basado en código predictivo se le llama “Configuración Prospectiva”. Esto tiene dos ventajas importantes: primero, la regla de ajuste es local, ya que cada neurona sólo necesita comparar su predicción con la actividad de su vecina inmediata; segundo, a pesar de esta localidad, el ajuste global de la red emerge como consecuencia, sin necesidad de retropropagar el error explícitamente. Este tipo de algoritmo puede alcanzar un rendimiento comparable al de *backpropagation*, aunque requiere más cómputo y converge más lentamente. Su plausibilidad biológica ha generado bastante interés, ya que encaja con observaciones sobre cómo las neuronas corticales parecen codificar errores de predicción en lugar de representaciones directas del mundo. La plasticidad basada en código predictivo es parte de una idea más general [56], según la cual el cerebro está constantemente intentando predecir su próximo input sensorial y ajustando sus representaciones en función del error entre esa predicción y lo que realmente ocurre. Algunos autores van más lejos y postulan que este principio es suficiente para explicar toda la cognición y el comportamiento [57], una postura que genera debate considerable en el campo.

Replay

Más allá de la regla de plasticidad, el cerebro debe también seleccionar qué experiencias consolidar. Una hipótesis prominente es que esto ocurre mediante *replay*: durante

La memoria no se almacena en un único lugar del cerebro. Está distribuida en múltiples sistemas, implementados por distintos circuitos que operan en diferentes escalas de tiempo y cumplen funciones computacionales distintas.

el descanso, especialmente el sueño, el cerebro reactiva patrones de actividad asociados a experiencias previas, y esa reactivación, en conjunto con las reglas de plasticidad, consolida el conocimiento en las sinapsis. Por ejemplo, si en tu práctica de tenis ensayaste el saque muchas veces, durante el sueño se reactivarán neuronas asociadas a su ejecución en la corteza motora, consolidando ese aprendizaje en las sinapsis. Se piensa que el hipocampo es el principal coordinador de este proceso [58], aunque se han observado mecanismos similares en otras áreas de forma independiente [59]. Existen además teorías matemáticas basadas en aprendizaje por refuerzo que proponen criterios para explicar qué experiencias son priorizadas para replay y cuándo ocurre este proceso [60, 61].

Sistemas de memoria

Hasta ahora hemos visto cómo la experiencia modifica las sinapsis y cómo esos cambios alteran los circuitos neuronales. La memoria no se almacena en un único lugar del cerebro. Está distribuida en múltiples sistemas, implementados por distintos circuitos que operan en diferentes escalas de tiempo y cumplen funciones computacionales distintas.

El *hipocampo* permite formar asociaciones nuevas con rapidez, incluso después de una sola experiencia. Funciona como un índice temporal que vincula patrones distribuidos en la corteza. En cambio, el *neocórtex* aprende de manera gradual: sus representaciones se ajustan lentamente a partir de exposiciones repetidas.

Con el tiempo, la información inicialmente dependiente del hipocampo se integra directamente en redes corticales mediante un proceso llamado *consolidación sistémica* [62, 63]. Como vimos anteriormente, el mecanismo de replay reactiva patrones de actividad neuronal en la corteza durante el sueño y los períodos de reposo, consolidando representaciones de largo plazo. Por eso, el daño al hipocampo afec-

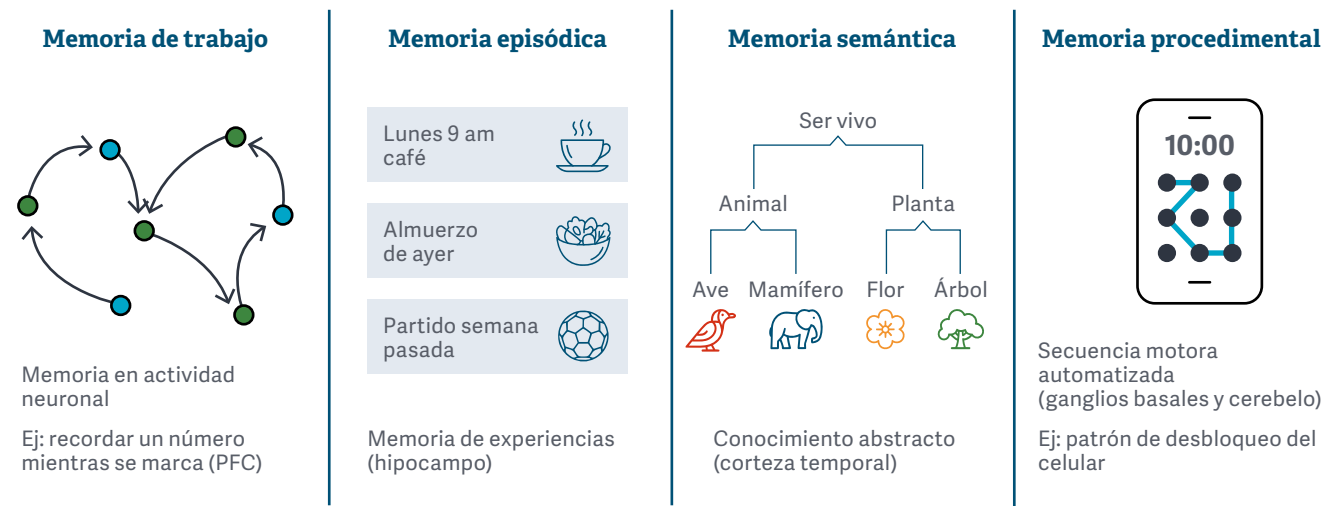


Figura 9 / Tipos de memoria y sus mecanismos. **Memoria de trabajo:** mantenida por actividad neuronal recurrente en la corteza prefrontal (PFC), sin cambios sinápticos duraderos. Ejemplo: recordar un número mientras se marca. **Memoria episódica:** registro de experiencias específicas que incluye qué ocurrió, cuándo y dónde, como el café del lunes o el partido de la semana pasada. Depende críticamente del hipocampo. **Memoria semántica:** conocimiento abstracto sobre el mundo organizado en categorías y relaciones jerárquicas, como la clasificación de seres vivos en animales y plantas. Se almacena en redes distribuidas de la corteza temporal. **Memoria procedimental:** habilidades motoras automatizadas, como el patrón de desbloqueo del teléfono o técnicas en algún deporte, que se ejecutan sin atención consciente a los pasos individuales. Involucra los ganglios basales y el cerebelo.

ta sobre todo la formación de recuerdos nuevos, mientras que muchos recuerdos antiguos permanecen intactos [64].

Sobre esta base se distinguen varios sistemas de memoria (ver Figura 9):

Memoria de trabajo

Permite mantener temporalmente información activa, por ejemplo al recordar un número telefónico mientras lo marcas. Se implementa principalmente mediante patrones sostenidos de actividad en circuitos prefrontales y parietales, más que por cambios sinápticos duraderos [65, 66]. Esta es la memoria mantenida por los modelos propuestos por Wang and Cohen sobre el PFC.

Memoria episódica

Corresponde a recuerdos de experiencias específicas, incluyendo qué ocurrió, dónde y cuándo. Depende críticamente del hipocampo, que vincula representaciones distribuidas en distintas áreas corticales para formar una memoria coherente [67, 68, 69].

Memoria semántica

Contiene conocimiento general sobre el mundo, como hechos, conceptos y relaciones. Se almacena en redes cor-

tales distribuidas, especialmente en áreas temporales y de asociación [70, 71, 72]. Muchas memorias semánticas se originan en experiencias episódicas, pero con el tiempo pierden su contexto y se transforman en conocimiento más abstracto. Se piensa que el mecanismo de replay consolida la memoria semántica a través de múltiples updates de patrones de activación [62, 63].

Memoria procedimental

Corresponde a habilidades y hábitos, como andar en bicicleta, escribir en un teclado o tocar un instrumento. Involucra principalmente a los *ganglios basales* y al *cerebelo*, que aprenden a seleccionar acciones y ajustar su ejecución mediante señales de refuerzo y corrección de error [73, 59].

Redes de Hopfield para almacenar memorias

Las redes de Hopfield [74], trabajo por el cual John Hopfield recibió el Premio Nobel de Física en 2024, son un modelo de memoria asociativa en el que las memorias se almacenan en los pesos sinápticos de una red recurrente. Su funcionamiento tiene dos etapas. Primero, el almacenamiento: cuando se quiere memorizar un patrón, los pesos entre neuronas que se activan juntas se fortalecen siguiendo una regla hebbiana. Este ajuste ocurre de una sola vez, sin entrenamiento iterativo, lo que lo hace atractivo como modelo de aprendizaje rápido a partir de una sola exposición. Segundo, la

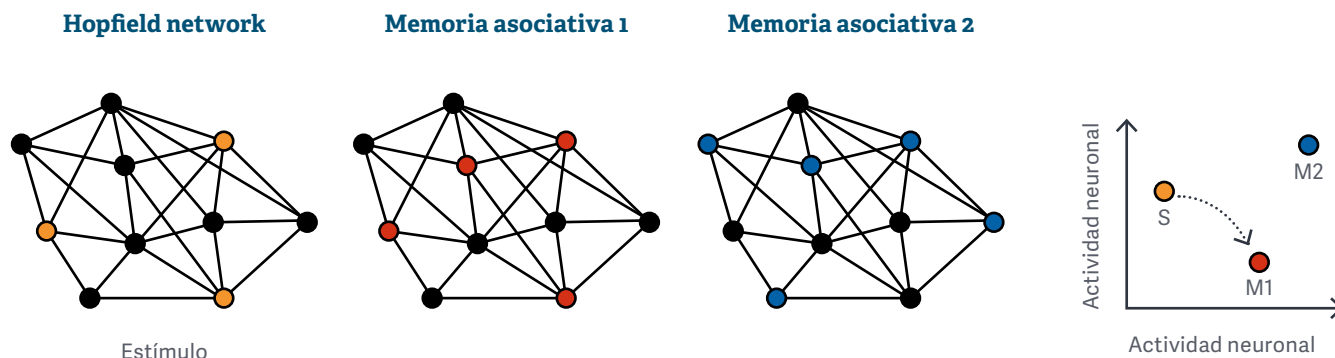


Figura 10 / Red de Hopfield y recuperación de memoria asociativa. Izquierda: una red de Hopfield recibe un estímulo parcial (neuronas naranjas), que activa un subconjunto de la red. La dinámica recurrente de la red evoluciona este patrón de actividad hacia la memoria almacenada más cercana: en este caso, la memoria asociativa 1 (neuronas rojas) o la memoria asociativa 2 (neuronas azules), dependiendo de cuál esté más próxima al estímulo inicial. Derecha: el mismo proceso representado en el espacio de actividad neuronal. El estímulo (S, naranja) corresponde a un punto en ese espacio, y la dinámica de la red lo desplaza hacia el atractor más cercano, en este caso la memoria M1 (rojo). Cada memoria almacenada corresponde a un mínimo de energía en este espacio, y la recuperación equivale a descender hacia ese mínimo.

recuperación: si se presenta a la red un estímulo parcial o ruidoso de un recuerdo almacenado, por ejemplo el rostro a medias de una persona conocida, la dinámica recurrente de la red actúa como un atractor que empuja la actividad neuronal hacia el patrón completo más cercano. En otras palabras, el sistema completa el recuerdo por sí solo (ver Figura 10).

Este modelo ha sido utilizado como descripción del área CA3 del hipocampo, que se cree funciona precisamente como una memoria auto asociativa capaz de recuperar experiencias completas a partir de pistas parciales [75, 76]. Una limitación clásica es que la capacidad de almacenamiento es modesta: más allá de cierto número de patrones, los recuerdos empiezan a interferir entre sí y la recuperación falla.

Una versión moderna de estas redes, propuesta por Ramsauer et al. [77], extiende el modelo a estados continuos y logra una capacidad de almacenamiento exponencialmente mayor. Lo notable es que esta versión moderna es matemáticamente equivalente al mecanismo de atención de los transformers, el componente central de modelos de lenguaje como GPT. Este resultado establece un puente inesperado entre un modelo de memoria biológica de los años ochenta y la arquitectura dominante en IA moderna.

Oportunidades en neurociencia computacional

Existen muchas preguntas abiertas en neurociencia que podrían beneficiarse del modelamiento matemático. Algunas las hemos discutido a lo largo del artículo: ¿cuál es la re-

gla de plasticidad del cerebro? ¿Qué nivel de abstracción es el adecuado para modelar neuronas y redes? ¿Cómo sabe el cerebro qué debe aprender y qué no? ¿Cómo la corteza prefrontal logra implementar algoritmos tan diversos? Aquí presentamos algunas preguntas adicionales que consideramos especialmente relevantes.

Planificación e imaginación

¿Cómo planifica el cerebro hacia el futuro o imagina situaciones que no han ocurrido? Los marcos teóricos más influyentes para este problema son los modelos de aprendizaje por refuerzo basados en modelos del mundo (*model-based RL* [78]) y los algoritmos de búsqueda en árboles de posibilidades, como los utilizados por AlphaGo [79, 80] para resolver el juego de Go, que venció al campeón mundial Lee Sedol por 4 a 1 en marzo de 2016, una década antes de lo que los expertos consideraban posible. Existen modelos de IA con esta capacidad, aunque los modelos de lenguaje más utilizados no la implementan de forma nativa; sin embargo existen intentos de incorporarse [81, 82].

Modularidad del cerebro

¿Por qué el cerebro está organizado en módulos especializados y no es simplemente una red grande de neuronas? Una respuesta posible proviene de la teoría de grafos: distintas funciones cognitivas podrían interferir entre sí si compartieran los mismos recursos, y la modularidad permitiría procesamiento en paralelo de distintos estímulos y tareas. Ver, escuchar, caminar y conversar simultáneamente requiere que distintos sistemas operen de forma relativa-



La inteligencia biológica y la artificial comparten principios fundamentales [...]. Quizás la inteligencia no es sólo un producto de la evolución de los seres vivos o un logro de la ingeniería, sino algo más fundamental en el universo.

mente independiente [36, 83]. Sin embargo, el origen y los límites de esta modularidad siguen siendo debatidos.

Conciencia

¿Qué es la conciencia y por qué tenemos experiencia subjetiva? Este es el llamado problema difícil de la conciencia, y sigue sin resolverse. Gran parte del debate todavía gira en torno a la definición de conciencia. Algunos proponen que la conciencia es una propiedad fundamental de cualquier sistema complejo, no exclusiva de los seres inteligentes [84]. Quizás la teoría más influyente es la del espacio de trabajo global: la conciencia emerge cuando cierta información se vuelve globalmente accesible dentro del cerebro, integrándose y quedando disponible simultáneamente para múltiples sistemas como la memoria, el razonamiento y la planificación. Bajo esta perspectiva, la mayor parte del procesamiento ocurre de forma inconsciente en subsistemas especializados, pero cuando la información entra en este espacio de trabajo compartido, se difunde ampliamente y permite una coordinación flexible entre sistemas. Esa disponibilidad global sería lo que asociamos con la experiencia consciente, y funcionalmente parece facilitar operaciones como el razonamiento abstracto, la planificación a largo plazo y el control deliberado de la conducta [85, 86].

¿Qué queremos decir con entender el cerebro?

Quizás una de las preguntas más fundamentales en neurociencia computacional es que significa “entender” cómo funciona el cerebro. Algunos incluso argumentan que el cerebro no puede entenderse a sí mismo en su totalidad, debido a que se intenta entender un instrumento (el cerebro) con el instrumento mismo, algo que suena contradictorio² [87]. Otros proponen que entender no requiere describir el cerebro funcionando con todas sus partes simultáneamen-

te, sino construir un atlas de las distintas funciones que implementa en distintas regiones, comprendiendo cada una por separado. Pero como vimos, incluso eso es difícil: cada región requiere modelos distintos, que son a su vez difíciles de comprender [88].

Para entender modelos de neurociencia computacional existen, a grandes rasgos, dos aproximaciones complementarias. La primera, quizás la más usada, consiste en proponer un modelo matemático simulable, compararlo con datos experimentales como actividad neuronal o comportamiento animal, e intervenir en su arquitectura, reglas de plasticidad u objetivo para estudiar cómo esas variaciones afectan los resultados [89, 90]. Este método es análogo a los experimentos de lesiones en el cerebro, donde se inactiva una región y se infiere su función observando los efectos en el comportamiento. Es un enfoque de arriba hacia abajo (*top-down*) que parte del sistema complejo y estudia las consecuencias de perturbarlo.

La segunda aproximación busca leyes o principios universales a través del análisis matemático, más parecido a lo que haría una física teórica [91]. Un gas está compuesto de millones de partículas con posiciones y velocidades específicas, pero la ley de los gases describe propiedades globales como la temperatura y la presión de forma exitosa sin entrar en el detalle de cada partícula. De forma análoga, muchos análisis matemáticos de redes neuronales describen propiedades globales de representación y aprendizaje sin modelar cada neurona individualmente, tomando prestadas herramientas de sistemas dinámicos, termodinámica y mecánica estadística [92, 93, 72]. La ventaja es que estas expresiones matemáticas se pueden entender de forma transparente bajo supuestos claros. La desventaja es que requieren tanta abstracción que a veces es difícil reconectar sus conclusiones con lo observado experimentalmente.

Es aquí donde existe una sinergia enorme con la inteligencia artificial: los modelos usados para describir partes del cerebro son muy parecidos a los modelos de aprendizaje de máquinas. Entender neurociencia computacional es, en gran medida, entender aprendizaje de máquinas, y viceversa. La idea que subyace a mucha de esta investigación es que la inteligencia biológica y la artificial comparten principios fundamentales, y que ambas son instancias de algo más profundo. Quizás la inteligencia no es sólo un producto de la evolución de los seres vivos o un logro de la ingeniería, sino algo más fundamental en el universo, donde el cerebro y la inteligencia artificial son dos instancias distintas de estudiarlo y entenderlo. **B**

2 <https://www.sainsburywellcome.org/web/blog/will-we-ever-understand-brain-12-leading-neuroscientists-reflect>
<https://www.thetransmitter.org/systems-neuroscience/the-limits-of-neuroscience/>

Referencias

- [1] Hodgkin, A. L., & Huxley, A. F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of Physiology*, 117(4), 500–544. <https://doi.org/10.1113/jphysiol.1952.sp004764>.
- [2] Rall, W. (1962). Theory of Physiological Properties of Dendrites. *Annals of the New York Academy of Sciences*, 96(4), 1071–1092. <https://doi.org/10.1111/j.1749-6632.1962.tb54120.x>.
- [3] Tuckwell, H. C. (1988). *Introduction to Theoretical Neurobiology: Volume 1: Linear Cable Theory and Dendritic Structure* (Vol. 1). Cambridge University Press. <https://doi.org/10.1017/CBO9780511623271>.
- [4] Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6), 386–408. <https://doi.org/10.1037/h0042519>.
- [5] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.
- [6] Arcas, B. A. y, Fairhall, A. L., & Bialek, W. (2000). What Can a Single Neuron Compute? *Neural Information Processing Systems*. <https://api.semanticscholar.org/CorpusID:10445944>.
- [7] Beniaguev, D., Segev, I., & London, M. (2021). Single cortical neurons as deep artificial neural networks. *Neuron*, 109(17), 2727–2739. e3. <https://doi.org/10.1016/j.neuron.2021.07.002>.
- [8] Dittman, J. S. (2024). Taking a closer look at the synapse. *Proceedings of the National Academy of Sciences*, 121(33), e2412457121. <https://doi.org/10.1073/pnas.2412457121>.
- [9] Carvalho, T. P., & Buonomano, D. V. (2009). Differential Effects of Excitatory and Inhibitory Plasticity on Synaptically Driven Neuronal Input-Output Functions. *Neuron*, 61(5), 774–785. <https://doi.org/10.1016/j.neuron.2009.01.013>.
- [10] Kim, S., Wallace, M. L., El-Rifai, M., Knudsen, A. R., & Sabatini, B. L. (2022). Co-packaging of opposing neurotransmitters in individual synaptic vesicles in the central nervous system. *Neuron*, 110(8), 1371–1384.e7. <https://doi.org/10.1016/j.neuron.2022.01.007>.
- [11] Hebb, D. O. (1949). *The organization of behavior; a neuropsychological theory* (pp. xix, 335). Wiley.
- [12] Markram, H., Lübke, J., Frotscher, M., & Sakmann, B. (1997). Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science*, 275(5297), 213–215. <https://doi.org/10.1126/science.275.5297.213>.
- [13] Bi, G., & Poo, M. (1998). Synaptic Modifications in Cultured Hippocampal Neurons: Dependence on Spike Timing, Synaptic Strength, and Postsynaptic Cell Type. *Journal of Neuroscience*, 18(24), 10464–10472. <https://doi.org/10.1523/JNEUROSCI.18-24-10464.1998>.
- [14] Ross, M. T., Flores, D., Bertram, R., Johnson, F., & Hyson, R. L. (2017). Neuronal Intrinsic Physiology Changes During Development of a Learned Behavior. *eNeuro*, 4(5). <https://doi.org/10.1523/ENEURO.0297-17.2017>.
- [15] Disouky, A., Sanborn, M. A., Sabitha, K. R., Mostafa, M. M., Ayala, I. A., Bennett, D. A., Lu, Y., Zhou, Y., Keene, C. D., Weintraub, S., Gefen, T., Mesulam, M.-M., Geula, C., Maienschein-Cline, M., Rehman, J., & Lazarov, O. (2026). Human hippocampal neurogenesis in adulthood, ageing and Alzheimer's disease. *Nature*, 1–10. <https://doi.org/10.1038/s41586-026-10169-4>.
- [16] Lee, S., Liebana, S., Clopath, C., & Dabney, W. (2024). Lifelong Reinforcement Learning via Neuromodulation (arXiv:2408.08446). *arXiv*. <https://doi.org/10.48550/arXiv.2408.08446>.
- [17] Barton, A. J., Valdés, J. J., & Orchard, R. (2009). Neural networks with multiple general neuron models: A hybrid computational intelligence approach using Genetic Programming. *Neural Networks, Advances in Neural Networks Research: IJCNN2009*, 22(5), 614–622. <https://doi.org/10.1016/j.neunet.2009.06.043>.
- [18] Tavanaei, A., Ghodrati, M., Kheradpisheh, S. R., Masquelier, T., & Maida, A. S. (2019). Deep Learning in Spiking Neural Networks. *Neural Networks*, 111, 47–63. <https://doi.org/10.1016/j.neunet.2018.12.002>.
- [19] Murphy-Royal, C., Ching, S., & Papouin, T. (2023). A conceptual framework for astrocyte function. *Nature Neuroscience*, 26(11), 1848–1856. <https://doi.org/10.1038/s41593-023-01448-8>.
- [20] Cooper, M. L., Selles, M. C., Cammer, M., Redd, C., Gildea, H. K., Sall, J., Chiurri, K. E., Cheung, P., Wheeler, D. G., Saab, A. S., Liddelow, S. A., & Chao, M. V. (2026). Astrocytes connect specific brain regions through plastic networks. *Nature*, 1–9. <https://doi.org/10.1038/s41586-026-10426-6>.
- [21] Maldonado, K. A., & Alsayouri, K. (2026). *Physiology, Brain*. In StatPearls. StatPearls Publishing. <http://www.ncbi.nlm.nih.gov/books/NBK551718/>.
- [22] Taylor, H. P., Huynh, K. M., Thung, K.-H., Lin, G., Lyu, W., Lin, W., Ahmad, S., & Yap, P.-T. (2026). Functional hierarchy of the human neo-cortex across the lifespan. *Nature*, 1–10. <https://doi.org/10.1038/s41586-026-10219-x>.
- [23] Hawkins, J., Ahmad, S., & Cui, Y. (2017). A Theory of How Columns in the Neocortex Enable Learning the Structure of the World. *Frontiers in Neural Circuits*, 11. <https://doi.org/10.3389/fncir.2017.00081>.
- [24] Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology*, 148(3), 574–591. <https://doi.org/10.1113/jphysiol.1959.sp006308>.



- [25] Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4), 193–202. <https://doi.org/10.1007/BF00344251>.
- [26] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, 1(4), 541–551. <https://doi.org/10.1162/neco.1989.1.4.541>.
- [27] Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583), 607–609. <https://doi.org/10.1038/381607a0>.
- [28] Barlow, H. B. (2012). Possible Principles Underlying the Transformations of Sensory Messages. In W. A. Rosenblith (Ed.), *Sensory Communication* (p. 0). The MIT Press. <https://doi.org/10.7551/mitpress/9780262518420.003.0013>.
- [29] Manookin, M. B., & Rieke, F. (2023). Two Sides of the Same Coin: Efficient and Predictive Neural Coding. *Annual Review of Vision Science*, 9(Volume 9, 2023), 293–311. <https://doi.org/10.1146/annurev-vision-112122-020941>.
- [30] Wang, J. X., Kurth-Nelson, Z., Kumaran, D., Tirumala, D., Soyer, H., Leibo, J. Z., Hassabis, D., & Botvinick, M. (2018). Prefrontal cortex as a meta-reinforcement learning system. *Nature Neuroscience*, 21(6), 860–868. <https://doi.org/10.1038/s41593-018-0147-8>.
- [31] Wang, J. X. (2021). Meta-learning in natural and artificial intelligence. *Current Opinion in Behavioral Sciences, Computational Cognitive Neuroscience*, 38, 90–95. <https://doi.org/10.1016/j.cobeha.2021.01.002>.
- [32] Son, J., Lee, S., & Kim, G. (2024). When Meta-Learning Meets Online and Continual Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–20. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. <https://doi.org/10.1109/TPAMI.2024.3463709>.
- [33] Vettoruzzo, A., Bouguelia, M.-R., Vanschoren, J., Rögnvaldsson, T., & Santosh, K. (2024). Advances and Challenges in Meta-Learning: A Technical Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7), 4763–4779. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. <https://doi.org/10.1109/TPAMI.2024.3357847>.
- [34] Miller, E. K., & Cohen, J. D. (2001). An Integrative Theory of Prefrontal Cortex Function. *Annual Review of Neuroscience*, 24 (Volume 24, 2001), 167–202. <https://doi.org/10.1146/annurev.neuro.24.1.167>.
- [35] Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed processing account of the Stroop effect. *Psychological Review*, 97(3), 332–361. <https://doi.org/10.1037/0033-295X.97.3.332>.
- [36] Musslick, S., Saxe, A., Hoskin, A., Sagiv, Y., Reichman, D., Petri, G., & Cohen, J. (2020). On the Rational Boundedness of Cognitive Control: Shared Versus Separated Representations. *OSF*. <https://doi.org/10.31234/osf.io/jkhdf>.
- [37] Rizzolatti, G., & Luppino, G. (2001). The Cortical Motor System. *Neuron*, 31(6), 889–901. [https://doi.org/10.1016/S0896-6273\(01\)00423-8](https://doi.org/10.1016/S0896-6273(01)00423-8).
- [38] Churchland, M. M., Cunningham, J. P., Kaufman, M. T., Foster, J. D., Nuyujukian, P., Ryu, S. I., & Shenoy, K. V. (2012). Neural population dynamics during reaching. *Nature*, 487(7405), 51–56. <https://doi.org/10.1038/nature11129>.
- [39] Kaufman, M. T., Churchland, M. M., Ryu, S. I., & Shenoy, K. V. (2014). Cortical activity in the null space: Permitting preparation without movement. *Nature Neuroscience*, 17(3), 440–448. <https://doi.org/10.1038/nn.3643>.
- [40] Todorov, E., & Jordan, M. I. (2002). Optimal feedback control as a theory of motor coordination. *Nature Neuroscience*, 5(11), 1226–1235. <https://doi.org/10.1038/nn963>.
- [41] Sutton, R., & Barto, A. G. (2018). *Reinforcement learning: An Introduction*. MIT Press.
- [42] Schultz, W., Dayan, P., & Montague, P. R. (1997). A Neural Substrate of Prediction and Reward. *Science*, 275(5306), 1593–1599. <https://doi.org/10.1126/science.275.5306.1593>.
- [43] Niv, Y., Duff, M. O., & Dayan, P. (2005). Dopamine, uncertainty and TD learning. *Behavioral and Brain Functions: BBF*, 1, 6. <https://doi.org/10.1186/1744-9081-1-6>.
- [44] Bellman, R. (2010). *Dynamic Programming*. Princeton University Press.
- [45] Bari, A., Xu, S., Pignatelli, M., Takeuchi, D., Feng, J., Li, Y., & Tonegawa, S. (2020). Differential attentional control mechanisms by two distinct noradrenergic coeruleo-frontal cortical pathways. *Proceedings of the National Academy of Sciences*, 117(46), 29080–29089. <https://doi.org/10.1073/pnas.2015635117>.
- [46] Cardozo Pinto, D. F., Pomrenze, M. B., Guo, M. Y., Touponse, G. C., Chen, A. P. F., Bentzley, B. S., Eshel, N., & Malenka, R. C. (2025). Opponent control of reinforcement by striatal dopamine and serotonin. *Nature*, 639(8053), 143–152. <https://doi.org/10.1038/s41586-024-08412-x>.
- [47] Chantranupong, L., Beron, C. C., Zimmer, J. A., Wen, M. J., Wang, W., & Sabatini, B. L. (2023). Dopamine and glutamate regulate striatal acetylcholine in decision-making. *Nature*, 621(7979), 577–585. <https://doi.org/10.1038/s41586-023-06492-9>.
- [48] Collins, A. G. E., & Frank, M. J. (2016). Surprise! Dopamine signals mix action, value and error. *Nature Neuroscience*, 19(1), 3–5. <https://doi.org/10.1038/nn.4207>.

- [49] Wise, R. A., & Jordan, C. J. (2021). Dopamine, behavior, and addiction. *Journal of Biomedical Science*, 28(1), 83. <https://doi.org/10.1186/s12929-021-00779-7>.
- [50] Richards, B. A., & Lillicrap, T. P. (2019). Dendritic solutions to the credit assignment problem. *Current Opinion in Neurobiology, Neurobiology of Learning and Plasticity*, 54, 28–36. <https://doi.org/10.1016/j.conb.2018.08.003>.
- [51] Ororbia, A. G. (2023). Brain-Inspired Machine Intelligence: A Survey of Neurobiologically-Plausible Credit Assignment (arXiv:2312.09257). arXiv. <https://doi.org/10.48550/arXiv.2312.09257>.
- [52] Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J., & Hinton, G. (2020). Backpropagation and the brain. *Nature Reviews Neuroscience*, 21(6), 335–346. <https://doi.org/10.1038/s41583-020-0277-3>.
- [53] Etter, G. (2025). The brain as a blueprint: A survey of brain-inspired approaches to learning in artificial intelligence (arXiv:2511.04455). arXiv. <https://doi.org/10.48550/arXiv.2511.04455>.
- [54] Lillicrap, T. P., Cownden, D., Tweed, D. B., & Akerman, C. J. (2016). Random synaptic feedback weights support error backpropagation for deep learning. *Nature Communications*, 7(1), 13276. <https://doi.org/10.1038/ncomms13276>.
- [55] Song, Y., Millidge, B., Salvatori, T., Lukasiewicz, T., Xu, Z., & Bogacz, R. (2024). Inferring neural activity before plasticity as a foundation for learning beyond backpropagation. *Nature Neuroscience*, 27(2), 348–358. <https://doi.org/10.1038/s41593-023-01514-1>.
- [56] Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1), 79–87. <https://doi.org/10.1038/4580>.
- [57] Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1521), 1211–1221. <https://doi.org/10.1098/rstb.2008.0300>.
- [58] Ólafsdóttir, H. F., Bush, D., & Barry, C. (2018). The role of hippocampal replay in memory and planning. *Current Biology*, 28(1), R37–R50.
- [59] Thompson, E. J., Rollik, L., Waked, B., Mills, G., Kaur, J., Geva, B., Carrasco-Davis, R., George, T., Domine, C., Dorrell, W., & Stephenson-Jones, M. (2024). Replay of procedural experience is independent of the hippocampus (p. 2024.06.05.597547). bioRxiv. <https://doi.org/10.1101/2024.06.05.597547>.
- [60] Mattar, M. G., & Daw, N. D. (2018). Prioritized memory access explains planning and hippocampal replay. *Nature Neuroscience*, 21(11), 1609–1617.
- [61] Liu, Y., Mattar, M. G., Behrens, T. E. J., Daw, N. D., & Dolan, R. J. (2021). Experience replay is associated with efficient nonlocal learning. *Science*, 372(6544), eabf1357. <https://doi.org/10.1126/science.abf1357>.
- [62] McClelland, J., McNaughton, B., & O'Reilly, R. (1995). Why There are Complementary Learning Systems in the Hippocampus and Neocortex: Insights from the Successes and Failures of Connectionist Models of Learning and Memory. *Psychological Review*, 102, 419–457. <https://doi.org/10.1037/0033-295X.102.3.419>.
- [63] Sun, W., Advani, M., Spruston, N., Saxe, A., & Fitzgerald, J. E. (2023). Organizing memories for generalization in complementary learning systems. *Nature Neuroscience*, 26(8), 1438–1448. <https://doi.org/10.1038/s41593-023-01382-9>.
- [64] Squire, L. R. (2009). The Legacy of Patient H.M. for Neuroscience. *Neuron*, 61(1), 6–9. <https://doi.org/10.1016/j.neuron.2008.12.023>.
- [65] Reynolds, G. D., & Romano, A. C. (2016). The Development of Attention Systems and Working Memory in Infancy. *Frontiers in Systems Neuroscience*, 10. <https://doi.org/10.3389/fnsys.2016.00015>.
- [66] Oberauer, K., Lewandowsky, S., Awh, E., Brown, G. D. A., Conway, A., Cowan, N., Donkin, C., Farrell, S., Hitch, G. J., Hurlstone, M. J., Ma, W. J., Morey, C. C., Nee, D. E., Schweppe, J., Vergauwe, E., & Ward, G. (2018). Benchmarks for models of short-term and working memory. *Psychological Bulletin*, 144(9), 885–958. <https://doi.org/10.1037/bul0000153>.
- [67] Tulving, E. (2002). Episodic memory: From mind to brain. *Annual Review of Psychology*, 53, 1–25. <https://doi.org/10.1146/annurev.psych.53.100901.135114>.
- [68] Calderón, C. B., Loof, E. D., Ergo, K., Snoeck, A., Boehler, C. N., & Verguts, T. (2021). Signed Reward Prediction Errors in the Ventral Striatum Drive Episodic Memory. *Journal of Neuroscience*, 41(8), 1716–1726. <https://doi.org/10.1523/JNEUROSCI.1785-20.2020>.
- [69] Giallanza, T., Campbell, D., & Cohen, J. D. (2024). Toward the Emergence of Intelligent Control: Episodic Generalization and Optimization. *Open Mind*, 8, 688–722. https://doi.org/10.1162/opmi_a_00143.
- [70] Patterson, K., Nestor, P. J., & Rogers, T. T. (2007). Where do you know what you know? The representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience*, 8(12), 976–987. <https://doi.org/10.1038/nrn2277>.
- [71] Chen, L., Lambon Ralph, M. A., & Rogers, T. T. (2017). A unified model of human semantic knowledge and its disorders. *Nature Human Behaviour*, 1(3), 0039. <https://doi.org/10.1038/s41562-016-0039>.
- [72] Saxe, A. M., McClelland, J. L., & Ganguli, S. (2019). A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, 116(23), 11537–11546. <https://doi.org/10.1073/pnas.1820226116>.



- [73] Foerde, K., & Shohamy, D. (2011). The role of the basal ganglia in learning and memory: Insight from Parkinson's disease. *Neurobiology of Learning and Memory, Memory Impairment and Disease*, 96(4), 624–636. <https://doi.org/10.1016/j.nlm.2011.08.006>.
- [74] Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8), 2554–2558. <https://doi.org/10.1073/pnas.79.8.2554>.
- [75] Alonso, N., & Krichmar, J. L. (2024). A sparse quantized hopfield network for online-continual memory. *Nature Communications*, 15(1), 3722. <https://doi.org/10.1038/s41467-024-46976-4>.
- [76] Kang, L., & Toyozumi, T. (2024). Distinguishing examples while building concepts in hippocampal and artificial networks. *Nature Communications*, 15(1), 647. <https://doi.org/10.1038/s41467-024-44877-0>.
- [77] Ramsauer, H., Schäfl, B., Lehner, J., Seidl, P., Widrich, M., Adler, T., Gruber, L., Holzleitner, M., Pavlović, M., Sandve, G. K., Greiff, V., Kreil, D., Kopp, M., Klambauer, G., Brandstetter, J., & Hochreiter, S. (2021). Hopfield Networks is All You Need (arXiv:2008.02217). arXiv. <https://doi.org/10.48550/arXiv.2008.02217>.
- [78] Luo, F.-M., Xu, T., Lai, H., Chen, X.-H., Zhang, W., & Yu, Y. (2022). A Survey on Model-based Reinforcement Learning (arXiv:2206.09328). arXiv. <https://doi.org/10.48550/arXiv.2206.09328>.
- [79] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587), 484–489. <https://doi.org/10.1038/nature16961>.
- [80] McGrath, T., Kapishnikov, A., Tomašev, N., Pearce, A., Hassabis, D., Kim, B., Paquet, U., & Kramnik, V. (2022). Acquisition of Chess Knowledge in AlphaZero. *Proceedings of the National Academy of Sciences*, 119(47), e2206625119. <https://doi.org/10.1073/pnas.2206625119>.
- [81] Ding, J., Zhang, Y., Shang, Y., Feng, J., Zhang, Y., Zong, Z., Yuan, Y., Su, H., Li, N., Piao, J., Deng, Y., Sukiennik, N., Gao, C., Xu, F., & Li, Y. (2025). Understanding World or Predicting Future? A Comprehensive Survey of World Models (arXiv:2411.14499). arXiv. <https://doi.org/10.48550/arXiv.2411.14499>.
- [82] Hafner, D., Pasukonis, J., Ba, J., & Lillicrap, T. (2025). Mastering diverse control tasks through world models. *Nature*, 640(8059), 647–653. <https://doi.org/10.1038/s41586-025-08744-2>.
- [83] Béna, G., & Goodman, D. F. M. (2025). Dynamics of specialization in neural modules under resource constraints. *Nature Communications*, 16(1), 187. <https://doi.org/10.1038/s41467-024-55188-9>.
- [84] Chalmers, D. J. (2010). *The Character of Consciousness*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780195311105.001.0001>.
- [85] Baars, B. J. (2005). Global workspace theory of consciousness: Toward a cognitive neuroscience of human experience. In S. Laureys (Ed.), *The Boundaries of Consciousness: Neurobiology and Neuropathology* (Vol. 150, pp. 45–53). Elsevier. [https://doi.org/https://doi.org/10.1016/S0079-6123\(05\)50004-9](https://doi.org/https://doi.org/10.1016/S0079-6123(05)50004-9).
- [86] Ferrante, O., Gorska-Klimowska, U., Henin, S., Hirschhorn, R., Khalaf, A., Lepauvre, A., Liu, L., Richter, D., Vidal, Y., Bonacchi, N., Brown, T., Sripad, P., Armendariz, M., Bendtz, K., Ghafari, T., Hetenyi, D., Jeschke, J., Kozma, C., Mazumder, D. R., ... Melloni, L. (2025). Adversarial testing of global neuronal workspace and integrated information theories of consciousness. *Nature*, 642(8066), 133–142. <https://doi.org/10.1038/s41586-025-08888-1>.
- [87] Jonas, E., & Kording, K. P. (2017). Could a Neuroscientist Understand a Microprocessor? *PLOS Computational Biology*, 13(1), e1005268. <https://doi.org/10.1371/journal.pcbi.1005268>.
- [88] Carandini, M. (2012). From circuits to behavior: A bridge too far? *Nature Neuroscience*, 15(4), 507–509. <https://doi.org/10.1038/nn.3043>.
- [89] Lillicrap, T. P., & Kording, K. P. (2019). What does it mean to understand a neural network? arXiv:1907.06374 [Cs, q-Bio, Stat]. <http://arxiv.org/abs/1907.06374>.
- [90] Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., Berker, A. de, Ganguli, S., Gillon, C. J., Hafner, D., Kepecs, A., Kriegeskorte, N., Latham, P., Lindsay, G. W., Miller, K. D., Naud, R., Pack, C. C., ... Kording, K. P. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, 22(11), 1761–1770. <https://doi.org/10.1038/s41593-019-0520-2>.
- [91] Saxe, A., Nelli, S., & Summerfield, C. (2021). If deep learning is the answer, what is the question? *Nature Reviews Neuroscience*, 22(1), 55–67. <https://doi.org/10.1038/s41583-020-00395-8>.
- [92] Goldt, S., Advani, M., Saxe, A. M., Krzakala, F., & Zdeborová, L. (2019). Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup. *Advances in Neural Information Processing Systems*, 32. https://proceedings.neurips.cc/paper_files/paper/2019/hash/cab070d53bd0d200746fb852a922064a-Abstract.html.
- [93] Mei, S., Misiakiewicz, T., & Montanari, A. (2019). Mean-field theory of two-layers neural networks: Dimension-free bounds and kernel limit (arXiv:1902.06015). arXiv. <https://doi.org/10.48550/arXiv.1902.06015>.