

Artículo de divulgación

La IA y el límite de la comprensión científica

*AI and the limits of scientific understanding*Camila Adriana Portanda Polo ¹, Carlos Acuña Ocampo ^{2,*}, Selenne Romero-Servin ³¹ Tecnológico de Monterrey, Escuela de Ingeniería y Ciencias, Monterrey, Nuevo León, México.² Tecnológico de Monterrey, Escuela de Negocios EGADE, Monterrey, Nuevo León, México.³ Universidad de Buffalo, Departamento de Ciencias del Diagnóstico Oral, Nueva York, Estados Unidos.*Autor de correspondencia: carlos.acunaocampo@tec.mx**Recibido:**

1/10/2026

Aceptado:

12/10/2026

Publicado:

13/10/2026

RESUMEN

Podemos describir cada operación de una red neuronal, calcular sus errores y modificar sus parámetros. Cuando preguntamos qué aprendió para reconocer un rostro, una enfermedad o un objeto, la respuesta deja de ser evidente. El objetivo de este artículo es explicar esa diferencia. A partir del análisis de imágenes, muestra cómo una red construye representaciones a lo largo de sus capas y por qué conocer el valor de sus parámetros no basta para interpretar una decisión. También presenta métodos que permiten estudiar esas representaciones y examina el alcance de las explicaciones obtenidas mediante mapas de activación, visualización de características y métodos de atribución. Así, plantea que el límite de nuestra comprensión no es el mismo para todas las preguntas: una interpretación puede resultar útil para entender una operación del modelo y ser insuficiente para justificar una decisión en el contexto donde se aplica.

Palabras clave: inteligencia artificial; redes neuronales; representaciones; interpretabilidad; inteligencia artificial explicable.

ABSTRACT

We can describe every operation of a neural network, calculate its errors, and adjust its parameters. When we ask what it learned to recognize a face, a disease, or an object, the answer is no longer clear. The objective of this article is to explain that difference. Drawing on image analysis, it shows how a network builds representations across its layers and why knowing the values of its parameters is not enough to interpret a decision. It also presents methods for studying these representations and examines the scope of the explanations obtained through activation maps, feature visualization and attribution methods. In doing so, it suggests that the limits of our understanding depend on the question we ask: an interpretation may help us understand an operation within the model yet be insufficient to justify a decision in the context where it is applied.

Keywords: artificial intelligence; neural networks; interpretability; explainable artificial intelligence.

INTRODUCCIÓN

Podemos describir cada operación de una red neuronal, calcular sus errores y modificar sus parámetros. Cuando preguntamos qué concepto construyó para reconocer un rostro, una enfermedad o un objeto, la respuesta deja de ser evidente. La dificultad de comprender el aprendizaje, aun cuando podemos controlar el proceso de entrenamiento, es uno de los problemas más fascinantes de la inteligencia artificial (Lipton, 2018).

Una imagen entra en un algoritmo y, unos instantes después, aparece una respuesta: gato, tumor, peatón o defecto de fabricación. Sabemos que la arquitectura procesó los datos y podemos medir cuántos casos clasificó correctamente. Incluso conocemos el valor de cada uno de sus parámetros. Aun así, todavía no podemos reconstruir y explicar de manera completa, paso a paso, cómo el modelo aprendió a reconocer aquello que buscábamos.

Esta situación suele resumirse diciendo que la inteligencia artificial es una caja negra. La expresión es útil, aunque puede conducir a una idea equivocada. El algoritmo no es un objeto mágico cuyo funcionamiento ignoremos por completo. Conocemos sus componentes, sus ecuaciones y el procedimiento que modifica sus conexiones. El problema aparece en otro nivel: el conocimiento adquirido se distribuye entre millones o miles de millones de parámetros y no queda almacenado como una lista de reglas comprensibles.

El propósito de este artículo es abordar esa diferencia sin presentar los límites de nuestra comprensión como un misterio absoluto. El desarrollo comprende el funcionamiento técnico de los modelos y la visión por computadora, los riesgos de interpretación en decisiones médicas o de alto impacto y la formación de quienes desarrollan estos sistemas. Sabemos por qué el entrenamiento puede mejorar el desempeño; lo que todavía investigamos es cómo se organizan las representaciones que permiten a un modelo reconocer patrones y tomar una decisión.

LO QUE SABEMOS DEL APRENDIZAJE

En términos matemáticos, una red neuronal es una función con una gran cantidad de parámetros. Recibe una entrada, realiza operaciones sucesivas y produce una salida (Figura 1). Durante el entrenamiento comparamos

esa salida con la respuesta esperada mediante una función de pérdida. Si el resultado es incorrecto, el algoritmo calcula cómo contribuyó cada parámetro al error. Después modifica ligeramente esos valores para reducir la pérdida. Este proceso, conocido como propagación hacia atrás y descenso de gradiente, se repite miles o millones de veces (LeCun et al., 2015).

Nada de esto ocurre al azar. Podemos inspeccionar los datos, definir la función que se optimiza, elegir la arquitectura, controlar la tasa de aprendizaje y observar la evolución del error. También podemos reproducir el entrenamiento bajo condiciones determinadas. Así, sabemos cómo aprende el algoritmo: busca una configuración de parámetros que produzca respuestas cada vez más cercanas a los ejemplos que le proporcionamos.

Lo que no especificamos es el conjunto exacto de conceptos intermedios que debe formar. Al programar un clasificador de aves no escribimos una regla que diga dónde comienza un ala, qué textura corresponde a una pluma o cómo distinguir un pico de una rama. Mostramos ejemplos y dejamos que el sistema descubra representaciones útiles para disminuir el error. El aprendizaje ocurre precisamente porque la máquina construye soluciones que no fueron codificadas de manera explícita.

Geoffrey Hinton comparó el diseño del algoritmo de aprendizaje con «diseñar el principio de la evolución» (como se citó en Pelley, 2024). Las características que permiten distinguir una imagen se aprenden a partir de los datos mediante un procedimiento general de entrenamiento (LeCun et al., 2015). Aunque elegimos la arquitectura y el objetivo que buscamos optimizar, las representaciones que se forman durante el aprendizaje requieren un estudio propio. Conocer las condiciones bajo las que se construyen nos ayuda a investigar su origen, pero no basta para comprender qué información contienen y cómo intervienen en una decisión.

Podemos saber por qué el entrenamiento mejora el desempeño y, al mismo tiempo, no saber qué representación interna surgió. Es como observar una pintura formada por miles de pequeños puntos de color: podemos conocer la posición y el tono de cada uno, pero ninguno contiene por sí solo la figura representada. La imagen aparece únicamente cuando todos se combinan. En el modelo, conocemos sus parámetros, pero resulta complejo explicar cómo su interacción da lugar a los patrones que ha aprendido.



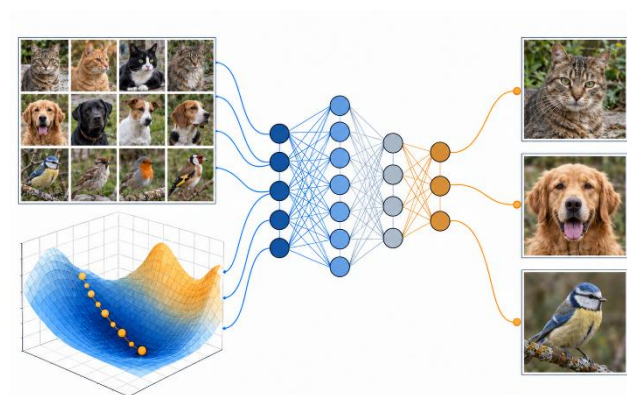


Figura 1. Entrenamiento de una red neuronal con imágenes. Las imágenes del conjunto de entrenamiento se utilizan para ajustar los parámetros de la red y aprender a distinguir entre categorías.

DE LOS PÍXELES A LAS REPRESENTACIONES

Las redes neuronales convolucionales permiten observar esta diferencia con claridad. Para una computadora, una fotografía no comienza como un pájaro, un rostro o un pulmón. Comienza como una matriz de números que representa la intensidad y el color de cada píxel. Las primeras capas, como se ilustra en la Figura 2, suelen responder a cambios simples: bordes, líneas, orientaciones o contrastes. Las capas posteriores combinan esas señales y forman texturas, contornos, partes de objetos y configuraciones cada vez más complejas.

Esta organización jerárquica no se introduce como un diccionario visual, sino que emerge porque resulta útil para resolver la tarea. Zeiler y Fergus (2014) mostraron que era posible proyectar hacia el espacio de la imagen las activaciones de distintas capas y observar esta progresión, desde patrones básicos hasta combinaciones más complejas de formas y texturas. Trabajos posteriores de visualización de características produjeron imágenes capaces de activar neuronas o canales específicos y permitieron explorar algunas de las representaciones aprendidas por la red (Olah et al., 2017).

La red no necesariamente aprende las mismas categorías que empleamos las personas. Una unidad puede responder a una textura presente en distintos objetos, mientras que otra puede activarse ante combinaciones de patrones que no corresponden a una forma reconocible. Aunque en

ocasiones aparecen detectores que parecen corresponder a ruedas, ventanas u hojas, la información también puede estar distribuida entre muchas unidades, sin que ninguna represente por sí sola un concepto completo.

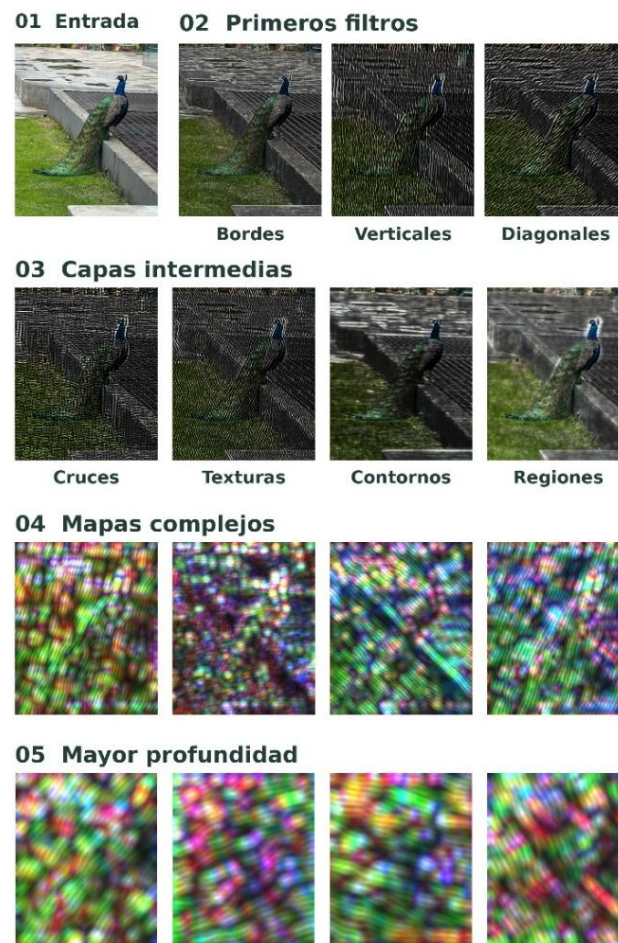


Figura 2. Las representaciones internas y los límites de su comprensión. Una red neuronal artificial para el análisis de imágenes combina información a lo largo de sus capas y construye representaciones de mayor complejidad. En las primeras podemos identificar respuestas asociadas con bordes, orientaciones y texturas, mientras que en las más profundas el significado de los patrones se vuelve incomprensible.

POR QUÉ LOS PARÁMETROS NO SE CONVIERTEN EN UNA EXPLICACIÓN

Un modelo moderno puede contener millones de números. Cada uno tiene un valor conocido, pero leerlos no equivale a comprenderlos. La primera dificultad es la

escala. Una predicción depende de una cadena extensa de operaciones en la que pequeñas contribuciones se combinan. La segunda es la no linealidad: el efecto de una señal cambia según las demás señales presentes. La tercera es que el significado puede estar distribuido. Un mismo concepto puede depender de muchas neuronas y una misma neurona puede participar en varios conceptos.

Esta última propiedad se relaciona con la superposición. Cuando el espacio disponible es limitado, una red puede representar más características que dimensiones utilizando combinaciones de activaciones. El resultado es eficiente para calcular, pero complicado para interpretar. Buscar la neurona del gato o la neurona de la enfermedad puede ser una pregunta mal formulada: la representación puede existir como un patrón compartido por muchas unidades (Elhage et al., 2022).

Además, dos redes con desempeño similar pueden construir soluciones internas diferentes. El entrenamiento no recorre un único camino obligatorio; encuentra una de muchas configuraciones capaces de reducir el error. Conocer el algoritmo de optimización no determina de antemano qué rasgos serán decisivos. La función aprendida depende de la arquitectura, la inicialización, los ejemplos disponibles y las relaciones que el modelo aprende a partir de ellos.

LA MÁQUINA PUEDE APRENDER OTRA LECCIÓN

La libertad para descubrir representaciones es la fuente de la capacidad de la IA y también de sus fallas más desconcertantes. Si todas las fotografías de perros contienen césped y las de gatos fueron tomadas dentro de una casa, una red puede utilizar el fondo como una señal. Obtendrá buenos resultados mientras esa coincidencia se mantenga. Cuando aparezca un gato en el parque, el supuesto conocimiento se desmoronará.

En imágenes médicas, un modelo puede asociar el diagnóstico con la marca del equipo, el hospital de origen, la posición del paciente o un elemento impreso en la radiografía. No recibió la orden de hacerlo. Encontró una regularidad que ayudaba a disminuir el error. En términos matemáticos, el entrenamiento funcionó; desde la perspectiva clínica, el modelo aprendió algo distinto de lo que necesitábamos.

Por ejemplo, en el área de la patología, una especialidad médica dedicada al estudio y diagnóstico de las enfermedades mediante el análisis de sus características microscópicas, un modelo con imágenes histológicas puede aprender a asociar un diagnóstico con características relevantes, pero también con variaciones relacionadas con la tinción, el procesamiento del tejido, el escáner o las condiciones de adquisición de la imagen (Stacke et al., 2021). Por ello, una clasificación correcta no necesariamente implica que el modelo haya aprendido los criterios histopatológicos que utiliza el patólogo para establecer un diagnóstico.

Esta posibilidad explica por qué la precisión no revela por sí sola qué aprendió un algoritmo. Una prueba externa, realizada con pacientes, equipos o instituciones diferentes, puede mostrar si la representación se relaciona con la enfermedad o con las condiciones particulares de la base original. Interpretar el modelo y evaluar su generalización son tareas complementarias. En decisiones de alto impacto, un sistema ligeramente menos preciso, pero más interpretable puede ser preferible si permite detectar errores, justificar resultados y establecer responsabilidades (Rudin, 2019).

CÓMO NOS ASOMAMOS AL INTERIOR

La inteligencia artificial explicable reúne métodos que intentan relacionar una predicción con sus posibles causas. Los mapas de activación resaltan regiones relevantes de una imagen. La visualización de características busca patrones que estimulen componentes internos. Los métodos de atribución estiman cómo cambia la salida cuando se modifica una entrada. Las sondas examinan si cierta información puede recuperarse de una capa. El análisis de circuitos intenta identificar grupos de componentes que realizan una operación reconocible.

Estas herramientas producen evidencia, no una traducción perfecta de la red. Un mapa de calor puede mostrar dónde se concentró la respuesta sin explicar qué rasgo se detectó. Una imagen sintetizada para activar una neurona puede revelar preferencias, pero también depende del método con el que fue creada. Una explicación local describe una predicción concreta y no necesariamente representa el comportamiento general del sistema.

Interpretar un modelo requiere combinar perspectivas. Conviene observar activaciones, alterar entradas, comparar ejemplos, estudiar errores y probar hipótesis. Si pensamos que una red reconoce la forma de un objeto, podemos conservar la forma y cambiar la textura; si sospechamos que utiliza el fondo, podemos reemplazarlo. La explicación gana fuerza cuando permite formular una predicción y esta se confirma mediante una intervención. Estudios como Network Dissection muestran que el análisis sistemático de unidades y conceptos puede aportar evidencia útil, aunque parcial, sobre las representaciones visuales (Bau et al., 2017).

¿CÓMO SE APRENDE A CONSTRUIR LO QUE TODAVÍA ESTAMOS APRENDIENDO A COMPRENDER?

Si bien existen herramientas que permiten observar el comportamiento de un modelo y formular hipótesis sobre lo que pudo haber aprendido, cabe preguntarse algo más reflexivo: ¿qué formación educativa necesita alguien que va a construir estos sistemas? Durante la formación en Ingeniería en Tecnologías Computacionales, esta pregunta surge al entrenar un modelo y obtener resultados sin poder explicar realmente por qué funciona. En la formación tradicional del programador, normalmente se espera que un sistema se comporte exactamente como lo diseña el programador. Estos modelos pueden encontrar estrategias que no anticipamos y aun así cumplir correctamente con la tarea.

Esta reflexión nos demuestra que aprender a construir modelos y aprender a comprenderlos son habilidades relacionadas, pero no iguales. La primera requiere conocer arquitecturas, funciones y manejo de datos; la segunda implica desarrollar una actitud mucho más cercana a la investigación científica y a la ingeniería inversa. Los futuros ingenieros en IA deben ser capaces de observar el comportamiento de un modelo, crear sus propias hipótesis sobre lo que pudo haber aprendido, intervenir sobre sus datos y utilizar los resultados para reconstruir, en la medida de lo posible, las representaciones y estrategias que desarrolló el modelo. Con esta formación, podríamos enfrentarnos a modelos que no comprendemos por completo, pero ante los cuales al menos tenemos la esperanza de poder investigar.

CONCLUSIONES

El límite de nuestra comprensión se hace evidente cuando intentamos establecer qué podemos concluir a partir de los resultados de una red neuronal. Que un modelo clasifique correctamente una imagen no demuestra que haya aprendido aquello que consideramos relevante, por lo que su desempeño debe acompañarse de evidencia sobre qué información utiliza y cómo se relaciona con la tarea que buscamos resolver.

Interpretar una red neuronal también requiere del conocimiento del área en la que se utiliza. Identificar qué información influye en una predicción puede ser insuficiente para justificar una decisión clínica, pues el valor de ese hallazgo depende del contexto en el que se interpreta. Esta exigencia amplía la formación de quienes desarrollan inteligencia artificial, pues construir un modelo implica también aprender a investigar sus resultados y reconocer hasta dónde permiten sostener una conclusión.

La inteligencia artificial nos permite construir modelos cuya capacidad para realizar una tarea puede superar nuestra comprensión de cómo lo consiguen. Esto plantea un desafío para el conocimiento científico, porque obtener un resultado y comprender el proceso que lo produce son logros distintos. Reconocer esa diferencia permite valorar los avances de la IA sin confundir su capacidad de responder con nuestra capacidad de comprenderla. La formación de quienes desarrollan IA debe incorporar, junto con las habilidades técnicas, la capacidad de formular hipótesis y examinar críticamente lo que aprenden los modelos.

Literatura citada

- Bau, D., Zhou, B., Khosla, A., Oliva, A., & Torralba, A. (2017). Network dissection: Quantifying interpretability of deep visual representations. En *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 6541–6549). <https://doi.org/10.1109/CVPR.2017.354>
- Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M., & Olah, C. (2022). Toy models of superposition.



Transformer Circuits Thread. https://transformer-circuits.pub/2022/toy_model/index.html

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521, 436–444. <https://doi.org/10.1038/nature14539>

Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM*, 61(10), 36–43. <https://doi.org/10.1145/3233231>

Olah, C., Mordvintsev, A., & Schubert, L. (2017). Feature visualization. *Distill*, 2(11), e7. <https://doi.org/10.23915/distill.00007>

Pelley, S. (2024, 16 de junio). “Godfather of Artificial Intelligence” Geoffrey Hinton on the promise, risks of advanced AI. CBS News. <https://www.cbsnews.com/news/geoffrey-hinton-ai-dangers-60-minutes-transcript/>

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1, 206–215. <https://doi.org/10.1038/s42256-019-0048-x>

Stacke, K., Eilertsen, G., Unger, J., & Lundström, C. (2021). Measuring domain shift for deep learning in histopathology. *IEEE Journal of Biomedical and Health Informatics*, 25(2), 325–336. <https://doi.org/10.1109/JBHI.2020.3032060>

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. En D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *Computer vision – ECCV 2014* (pp. 818–833). Springer. https://doi.org/10.1007/978-3-319-10590-1_53

Aviso legal/Nota del editor: Las declaraciones, opiniones y datos contenidos en todas las publicaciones son exclusivamente de los autores y colaboradores, y no de Agraria ni de sus editores. Agraria y sus editores no se responsabilizan de ningún daño a personas o bienes que resulte de las ideas, métodos, instrucciones o productos mencionados en el contenido.

