# *Opacidad epistémica y justificación científica en las redes neuronales profundas*

Alumno: Beatriz Penín Fernández

Tutor: Jesús Vega Encabo

Grado en Ciencia, Tecnología y Humanidades
Curso: 2025/2026

Universidad Autónoma de Madrid y Universidad Autónoma de Barcelona

Universidad Autónoma de Madrid

Universitat Autònoma de Barcelona

## Abstract

This study examines the epistemic conditions under which deep neural networks (DNNs) can produce legitimate scientific knowledge despite their constitutive opacity. Drawing on the case of AlphaFold 2, a program whose development was awarded the Nobel Prize in Chemistry in 2024, it is contended that this opacity, while irreducible in its most fundamental forms, does not invalidate the epistemic value of these technologies. Although techniques such as explainable AI or mechanistic interpretability allow certain forms of opacity to be partially addressed, the lack of intelligibility and comprehensibility that structurally characterizes DNNs precludes the justification of their results through a transparent connection between the internal functioning of the model and the phenomenon under study. The aim of this work is to show that such irreducibility does not exclude DNNs from scientific practice, insofar as they can function as research instruments and mediating models capable of orienting scientific discovery within a framework of justification and reliability external to the internal functioning of the system. Following the computational reliabilism proposed by Durán and Formanek, it is argued that verification and validation methods, robustness analysis, implementation history, and expert knowledge constitute the epistemic basis upon which the scientific community is capable of justifying the results of opaque systems such as AlphaFold 2, and which must form an indispensable part of their use in science.

**Keywords**: epistemic opacity, deep neural networks, AlphaFold2, computational reliabilism, irreducibility, scientific justification

## Resumen

Este estudio examina las condiciones epistémicas bajo las cuales las redes neuronales profundas (DNN) pueden producir conocimiento científico legítimo pese a su opacidad constitutiva. A partir del caso de AlphaFold2, programa cuyo desarrollo fue galardonado con el Premio Nobel de Química de 2024, se sostiene que esta opacidad, aunque irreductible en sus formas más fundamentales, no invalida el valor epistémico de estas tecnologías. Si bien técnicas como la inteligencia artificial explicable o la interpretabilidad mecánica permiten reducir parcialmente ciertas formas de opacidad, la falta de inteligibilidad y comprensibilidad que caracteriza estructuralmente a las DNN no permite justificar sus resultados mediante una conexión transparente entre el funcionamiento interno del modelo y el fenómeno estudiado. El objetivo de este trabajo es mostrar que dicha irreductibilidad no excluye a las DNN de la práctica científica en tanto que pueden funcionar como instrumentos de investigación y modelos capaces de orientar el descubrimiento científico dentro de un marco de justificación y fiabilidad externo al funcionamiento interno del sistema. Siguiendo el fiabilismo computacional propuesto por Durán y Formanek, se argumenta que los métodos de verificación y validación, el análisis de robustez, la historia de implementaciones y el conocimiento experto constituyen la base epistémica sobre la cual la comunidad científica es capaz de justificar los resultados de sistemas opacos como AlphaFold2 y que han de formar una parte indispensable de su uso en la ciencia.

**Palabras clave**: opacidad epistémica, redes neuronales profundas, AlphaFold2, fiabilismo computacional, irreductibilidad, justificación científica

# ÍNDICE

# 1. Introducción

¿Qué condiciones debe satisfacer una herramienta para que el conocimiento que produce pueda considerarse legítimo? Uno de los ideales epistémicos más destacados en el proceso de legitimación del conocimiento ha sido la idea de transparencia. Conocer algo supone, a través de este, poder dar cuenta de los procesos que conducen a los resultados, hacer accesibles los mecanismos que los generan.

La ciencia, en tanto práctica orientada al conocimiento, acoge este ideal en tradiciones como el empirismo, el acceso transparente de cada dato y mecanismo. Sin embargo, las aspiraciones científicas no siempre consideran, o ni pueden considerar, esa condición como suficiente. Los desafíos científicos han obedecido históricamente más a la demostración de un hecho dentro de un desconocimiento parcial de los procesos implicados, de algún tipo de "opacidad", entendiendo esto como la falta de acceso a ciertos elementos epistémicamente relevantes.

En adición a esto, podría argumentarse que esta falta de acceso es cada vez mayor. Esta deriva hacia una ciencia dentro de lo opaco, puede verse en el análisis que Paul Humphreys realiza en su libro *Extending Ourselves* (2004) sobre el uso científico de instrumentos que no permiten una transparencia de los resultados que muestran. La introducción de instrumentos como el telescopio, supuso un problema epistémico en la medida en que aquello que el telescopio mostraba no era accesible a simple vista, y por tanto las imágenes producidas debían *justificarse.* Humphreys defiende que en el caso de lo computacional este desplazamiento es más notorio porque sus capacidades de transformación de datos, de cálculo, modelización y representación exceden las capacidades humanas ordinarias. Este estudio parte, precisamente, de la enmarcación de las redes neuronales profundas como el caso que más se aleja actualmente de la concepción de un instrumento transparente.

Las redes neuronales profundas o DNN (Deep Neural Networks) son modelos capaces de realizar tareas complejas con predicciones de una precisión no alcanzada hasta ahora por otras herramientas computacionales. Como ejemplo de ello, el modelo AlphaFold2, cuyo desarrollo fue reconocido con el Premio Nobel de Química de 2024, permitió generar predicciones estructurales de más de doscientos millones de proteínas, permitiendo así cubrir prácticamente la totalidad de las proteínas conocidas.

Explicar hasta qué punto las redes neuronales profundas se alejan de la posible transparencia de lo computacional es precisamente una de las primeras tareas de este estudio. En otros sistemas computacionales, la justificación de los resultados puede apoyarse en el análisis detallado de los

algoritmos que los producen. Sin embargo, si los procesos internos de las redes neuronales profundas no parecen fácilmente accesibles y su opacidad no pudiera reducirse para alcanzar una explicación detallada de cómo se han producido sus resultados, el ideal de transparencia no puede ser usado como una condición de su justificación.

La hipótesis que guiará este estudio es que las opacidades que presentan las redes neuronales profundas, en distintos niveles y posibilidades de reducción, no implican necesariamente su exclusión del proceso de investigación científica. Su papel como herramientas epistémicas u orientación científica trascienden la necesidad de una transparencia en el sentido que exige el empirismo tradicional y desplazan los criterios de legitimidad hacia formas externas de justificación, que si bien cuentan con precedentes históricos en la ciencia, deben ser estudiadas y adecuadas en el contexto de estas tecnologías.

Para defender esta tesis, en primer lugar, se describirá el funcionamiento general de las redes neuronales profundas y su posible papel como herramientas epistémicas tomando AlphaFold2 como ejemplo representativo. Se analizarán a continuación los distintos niveles de transparencia y opacidad que afectan a estos modelos, distinguiendo entre trazabilidad, inteligibilidad y comprensibilidad, así como entre opacidad interna y externa con el fin de esclarecer hasta qué punto estas formas de opacidad pueden reducirse mediante técnicas computacionales y de interpretabilidad. Se examinarán los distintos papeles epistémicos que estas pueden desempeñar en la práctica científica más allá de lo instrumental con el fin de precisar que tipo de justificaciones podrían exigir para finalmente abordar la cuestión de la justificación científica bajo condiciones persistentes de opacidad. AlphaFold2 funcionará entonces como caso central para mostrar cómo una DNN puede adquirir legitimidad epistémica sin contar con un procedimiento completamente transparente, apoyándose en cambio en una red de fiabilidad, contraste experimental y conocimiento experto.

# 2. Descripción de las DNN

## 2.1 Las DNN como herramientas epistémicas

Con el fin de dilucidar qué tipo de instrumento son las redes neuronales profundas en el contexto científico, usaré la distinción conceptual que Margaret Morrison y Mary Morgan propusieron en su libro *Models as Mediators* (1999) con respecto a lo que separa el instrumento simple del instrumento de investigación. El primero actúa sobre el mundo sin representar ningún aspecto de él, al igual que un martillo transforma su entorno pero no esclarece nada sobre este. El instrumento de investigación, en cambio, incorpora una capacidad representativa que lo distingue; “No aprendemos mucho del

martillo. Pero otras herramientas (quizás más sofisticadas) pueden ayudarnos a ello. El termómetro es un instrumento de investigación: es físicamente independiente de una cacerola con mermelada, pero se puede introducir en la mermelada hirviendo para medir su temperatura." (p.11) El termómetro, en este ejemplo, comprende una dimensión representativa dentro de un marco científico que le permite transmitir conocimiento sobre el fenómeno que registra. Ahora bien, ¿es este el sentido en el que se ha demarcado a las redes neuronales profundas? ¿Su labor científica trasciende la de un instrumento simple?

La práctica científica reciente parece sugerir esta superación. Juan Durán, en "Beyond transparency: Computational reliabilism as an externalist epistemology of algorithms." (2025) documenta varios casos de investigaciones con redes neuronales profundas que han ampliado "la gama de capacidades de experimentación disponibles para los investigadores" (p.56). Entre ellos, el caso de referencia en este estudio, AlphaFold 2, un sistema desarrollado por DeepMind cuyo objetivo es predecir la estructura tridimensional que adoptará una proteína a partir únicamente de su secuencia de aminoácidos (Jumper et al., 2021).

Determinar las totalidad de las millones de secuencias de estructuras proteicas desde lo experimental, mediante técnicas como la cristalografía de rayos X o la espectroscopía de resonancia magnética nuclear, requiere entre meses y años de trabajo por proteína, y predecirla computacionalmente a partir de principios físicos resulta igualmente inabarcable, ello hizo que hasta la creación de AlphaFold2 el llamado "problema del plegamiento proteico" permaneciese inconcluso durante más de cincuenta años (Royal Swedish Academy of Sciences, 2024b).

AlphaFold 2, en lugar de derivar el plegamiento de principios teóricos previos, fue entrenado sobre una base de datos de estructuras proteicas ya determinadas experimentalmente, de modo que sus redes neuronales profundas "aprendían" autónomamente a predecir la estructura tridimensional a partir de regularidades presentes en los datos. Fue así el primer método capaz de predecir regularmente estructuras de proteínas con precisión atómica, "incluso en casos en los que no se conocía ninguna estructura similar" (Jumper et al., 2021, p.1). Este rendimiento fue el que llevó a la Academia Real Sueca de Ciencias a conceder el Premio Nobel de Química de 2024 a John Jumper y Demis Hassabis (Royal Swedish Academy of Sciences, 2024b). Este contexto innegablemente dota a esta tecnología de un reconocimiento mayor que el de un instrumento simple, e implica la aceptación de su capacidad representativa, de las redes neuronales profundas como un instrumento de investigación, o incluso algo que supere esto.

El uso de redes neuronales profundas tan relevantes para la comunidad científica ha propiciado distintos modos de justificación de sus resultados. Por ello, con el fin de hacer un seguimiento de este

caso, han de definirse primero los mecanismos que hacen funcionar a este sistema antes de juzgar el estatuto epistemológico de sus resultados.

## 2.2 Descripción del funcionamiento

Para entender realmente a qué refiere la literatura científica cuando usa el término "aprender" en las redes neuronales profundas, es necesaria una descripción funcional de este proceso. Una red neuronal profunda, como la presente en AlphaFold 2, puede describirse como un sistema que aproxima una función matemática entre un espacio de entrada y un espacio de salida, esto es, una correspondencia del tipo:

$$f\theta: X \rightarrow Y$$

donde X representa el espacio de los datos de entrada, Y el espacio de las salidas posibles, y θ el conjunto de parámetros del modelo, típicamente agrupados como pesos y sesgos. Estos pesos determinan la influencia que la activación de un nodo ejerce sobre los nodos de la capa siguiente. La información fluye desde una capa de entrada a través de una o varias capas ocultas hasta alcanzar una capa de salida. A medida que el procesamiento avanza, se vuelve menos clara la determinación de qué características o combinaciones de características están siendo representadas en cada nivel.

El sistema adquiere su comportamiento a través de un proceso iterativo de optimización: partiendo de unos parámetros iniciales, la red procesa ejemplos de entrenamiento y compara sus predicciones con los resultados esperados. La integración de una función de pérdida mide entonces la discrepancia entre ambos y guía el ajuste de los pesos mediante un algoritmo de *retropropagación del error*, este distribuye la responsabilidad del fallo a lo largo de todas las capas de la red. El proceso se repite miles o millones de veces, con el objetivo de que el sistema "aprenda" a generalizar en lugar de "memorizar" los ejemplos vistos, con el objetivo de ofrecer predicciones correctas sobre datos nuevos (Goodfellow, Bengio y Courville, 2016). Como señala Duede en "Deep learning opacity in scientific discovery" (2023), el buen desempeño del modelo se evalúa inductivamente a partir de esta capacidad para generalizar, de mantener un bajo error en conjuntos de prueba independientes extraídos de la misma distribución, es decir, de la reiteración de aciertos bajo condiciones controladas.

Su formalización matemática, sin embargo, tan sólo denota su naturaleza algorítmica, una secuencia finita de pasos definidos para resolver un problema. El descenso por gradiente estocástico, la retropropagación, la actualización iterativa de pesos, procedimientos característicos del aprendizaje automático son completamente especificados y deterministas dado un estado inicial, sin embargo, esto no equipara las redes neuronales profundas a un sistema de instrucciones explícitas.

Considérese el problema de determinar si un número entero es par o impar. Mediante el uso de una calculadora científica, como con cualquier otro programa clásico, el problema se resolvería aplicando una regla explícita que un programador especificó de antemano, el más habitual sería la división del número entre dos y la comprobación de si el resto es cero. El proceso es completamente trazable, reconstruible paso a paso, y su corrección puede verificarse directamente a partir de las instrucciones que lo definen.

A modo de comparativo, una red neuronal profunda entrenada para resolver el mismo problema no recibiría instrucciones en ese sentido. A pesar de la trivialidad y simpleza de la cuestión, el funcionamiento de la red sería el mismo, tomaría miles de ejemplos de números etiquetados como pares o impares, y a partir de ellos determinaría *autónomamente*, mediante la optimización iterativa de sus parámetros, la función que los clasifica correctamente. El resultado puede ser idéntico al de la calculadora, pero el proceso que lo genera es esencialmente distinto. En cierto sentido, podría decirse que la red genera una función que captura esa regularidad de manera operativa, lo que supone que, aunque esa función pueda ser matemáticamente equivalente a la definición de paridad, la red no posee esta como tal y no puede articular que está aplicando una regla de divisibilidad. Es decir, la red no alcanza una definición de lo que es ser un número par del mismo modo en que lo alcanza un sistema simbólico; lo que posee es una función ajustada que discrimina correctamente, sin que esa función esté vinculada a ningún contenido conceptual articulable.

Como resultado de esto, existe otra distinción relevante entre las redes neuronales profundas y los instrumentos computacionales más simples, que constituye además una preocupación directa para el ámbito científico, y es la que sugiere que los dos resultados que habíamos valorado como idénticos podrían no serlo. Según Duede y Davey en “A Priori Knowledge in an Era of Computational Opacity: The Role of Artificial Intelligence in Mathematical Discovery” (2025), un resultado producido por un sistema computacional, como el realizado en el ejemplo anterior, puede legitimarse de dos formas, o bien que “lo que el sistema está haciendo debe ser matemáticamente transparente para nosotros" (p. 8), o bien que, de ser un programa opaco, cuente con una “prueba de una afirmación matemática”(p.9). Esta propuesta plantea dos cuestiones que estructuran el análisis de este estudio. En primer lugar, observar si el funcionamiento de las redes neuronales profundas constituye efectivamente una forma de opacidad. En segundo lugar, establecer qué tipo de garantías epistémicas se han exigido como condición de legitimidad científica y de qué forma estas condicionan la práctica contemporánea. El término opacidad, necesario para responder ambas cuestiones habrá primero de ser definido en el caso de las redes neuronales profundas.

## 3. Causas y tipos de opacidad en DNN

Uno podría llegar a la conclusión de que el hecho de poder describir el funcionamiento de estas redes, como se ha hecho en la sección anterior, llevaría a deducir cierta transparencia en ellas. Sin embargo, esta transparencia no es uniforme o absoluta. Esta sección buscará explicar en primera instancia cómo el acceso epistémico a los mecanismos internos de un sistema admite distintos niveles, y cómo el tipo de resultados justificados a través de la transparencia que cada uno permite varía en consecuencia.

Esta valoración implica también que la aceptación de un sistema como productor de resultados reproducibles, la valoración de una serie de algoritmos "trazables" o rastreables como correctos, no significa la comprensión de qué los produce, ni de cómo sus componentes interactúan, ni de su adecuación describiendo un fenómeno complejo. Esta idea cimentará el segundo punto de esta sección a través de un análisis de los tipos de opacidad y su posible reductibilidad mediante técnicas de explicación e interpretabilidad.

La idea de reductibilidad e irreductibilidad en ciertos tipos de opacidad se basa en la crítica que Paul Humphreys desarrolla en *Extending Ourselves* a lo no observado y lo no observable en ciencia. Humphreys sostiene que en la ciencia contemporánea lo relevante para la producción de conocimiento científico ya no sigue al empirismo tradicional de que un fenómeno sea directamente observable por un sujeto humano, menciona que "el límite de lo observable no es fijo, sino que se expande continuamente a medida que mejora nuestro acceso instrumental al mundo." (Humphreys, 2004, p. 15). Entiende así que la actual aceptación científica de los resultados de un instrumento depende de que existan procesos de detección con los que justificar inferencias sobre él.

Este desplazamiento que describe hace que ciertas opacidades de las redes neuronales profundas tengan la perspectiva de ser reducidas, de expandir lo "observable", sin embargo, la hipótesis que guiará esta sección es que no todas las opacidades poseen el mismo peso epistemológico. Según su naturaleza y su grado de reducibilidad, algunas pueden proporcionar información suficiente para verificar determinados datos y procesos, constituyendo así un avance relevante para la computación moderna. Otras, en cambio, son *esencialmente* irreductibles. En estos casos, que cubrirán el objetivo global de este estudio, la justificación de resultados en descubrimientos científicos exige recurrir a otros métodos de detección, validación y fiabilidad que además implican un cambio en la producción de conocimiento científico.

## 3.1 Trazabilidad, inteligibilidad y comprensibilidad: ubicando la opacidad

Con el fin de diferenciar hasta qué punto se alcanza transparencia en el funcionamiento de las redes neuronales profundas, John Zerilli en “Explaining Machine Learning Decisions”, hace uso de tres aspectos del acceso epistémico a la lógica de las redes neuronales pertinentes en este estudio: la “trazabilidad”, la inteligibilidad y la comprensibilidad. La trazabilidad refiere a la capacidad de ejecutar el modelo computacionalmente y reproducir sus operaciones; en este sentido no difiere de la de una calculadora u otro sistema computacional, por muy sofisticados o más complejos que sean los algoritmos de las redes neuronales, ambos producen outputs verificables a partir de inputs dados (Zerilli, 2022). En el programa de AlphaFold2, esto correspondería a afirmar que el output obtenido corresponde al procedimiento computacional declarado, lo cual sería correcto, el código abierto de AlphaFold permite afirmar que el resultado ha sido generado mediante una ejecución concreta del procedimiento algorítmico implementado (Royal Swedish Academy of Sciences, 2024b).

En el caso de las redes neuronales profundas, se suceden problemas de transparencia en las siguientes dos concepciones de Zerilli. La primera, la inteligibilidad, refiere a la posibilidad de descifrar semánticamente las relaciones internas del modelo, es decir, de asignar un significado interpretable a sus componentes y a las transformaciones que realiza sobre ellos. En las redes neuronales profundas existe precisamente una ausencia de significado de los valores internos de la red neuronal profunda (pesos, sesgos y activaciones).

Podría objetarse que esta situación no es específica de las redes neuronales profundas, dado que cualquier estructura matemática, considerada de manera puramente formal, admite múltiples interpretaciones posibles. Las variables de una ecuación, por ejemplo la de un oscilador armónico,

$$m\frac{d^2x}{dt^2} + \gamma\frac{dx}{dt} + kx = 0$$

pueden interpretarse siendo X la posición de una masa acoplada a un muelle y γ la fricción que actúa sobre el sistema, de la misma manera que X podría medir "la distancia" en moles por litro (mol/L) a la que se encuentra una mezcla química de alcanzar su estabilidad, hasta que la "fricción" γ estabilice la muestra. Boge utiliza este ejemplo en su artículo "Two dimensions of opacity and the deep learning predicament" y describe que "cualquier modelo matemático, concebido puramente como una estructura formal, admite múltiples interpretaciones. [...] la interpretación entonces se logra asignando significados a sus símbolos matemáticos" (Boge, 2022, p. 50), mostrando así que la ambigüedad formal de un modelo matemático no es problemática en sí misma, siempre que exista una significación simbólica explícita que fije qué valores están siendo modelados y cómo. Los pesos,

sesgos y funciones de activación en una DNN no están diseñados para representar entidades, procesos o magnitudes del mundo en este sentido, los valores internos cumplen más bien una función exclusivamente instrumental dentro del proceso de optimización de la red. Carecen, por ello, de inteligibilidad, puesto que no es posible asignarles un significado interpretable.

La segunda problemática se da en lo tocante a la comprensibilidad. Aun suponiendo que sus estructuras funcionales pudieran recibir alguna interpretación semántica, persistiría en el estudio de las redes neuronales profundas la imposibilidad de *comprender* cómo esas estructuras interactúan para producir el resultado global, una incomprensibilidad. Zerilli explica este término a través del ejemplo de los bosques aleatorios *(random forests)*, también modelos de aprendizaje automático pero construidos a partir de la combinación de múltiples árboles de decisión. En ellos, cada árbol individual puede ser inteligible y su lógica seguirse paso a paso, cada nodo representa una regla de decisión cuyo significado puede articularse; sin embargo, el modelo en su conjunto, la interacción de los cientos o miles de árboles realizando tareas simultáneamente resulta incomprensible. No es posible seguir de manera directa cómo la combinación de todas esas reglas individuales; cómo ese *bosque* produce una decisión global.

Las redes neuronales profundas comparten esta dificultad con los bosques aleatorios. Jenna Burrell, en su artículo "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms" (2016) señala cómo el aprendizaje automático, característico de ambas tecnologías, suele describirse como afectado por la llamada "maldición de la dimensionalidad" (p.2), esto es, cómo la naturaleza del sistema funciona sobre miles o decenas de miles de características simultáneamente, y consecuentemente la lógica interna de sus decisiones se vuelve inabarcable para el razonamiento humano. Las redes neuronales profundas en este sentido comparten con los bosques aleatorios y otras tecnologías basadas en el aprendizaje automático una incomprensibilidad en primera instancia derivada de la escala. Burrell, no obstante, señala que esta incomprensibilidad no se sustenta únicamente en esta característica. Es correcto que, razonar sobre el algoritmo se vuelve más difícil cuantas más cualidades o características se proporcionan como inputs, pero no por su alta dimensionalidad en sí, sino porque cada uno desplaza de manera imperceptible la clasificación resultante, “los conjuntos de datos pueden ser extremadamente grandes, y el código puede estar escrito con claridad, sin embargo, es la interacción entre ambos dentro del mecanismo del algoritmo lo que genera la opacidad” (Burrell, 2016, p. 2). Esto significa que la incomprensibilidad de las DNN no es simplemente una cuestión que podría resolverse reduciendo el tamaño del modelo o simplificando su arquitectura, deriva del comportamiento en conjunto del sistema.

La presencia de según qué características de Zerilli afecta directamente a las caracterizaciones filosóficas que se han hecho sobre la opacidad en redes neuronales profundas, y a cómo no todas

presentan el mismo peso epistemológico ni plantean el mismo tipo de problema para la producción de conocimiento científico. Aquellas que coinciden con una falta de trazabilidad, como sería la ocultación deliberada de los algoritmos por parte de empresas y organizaciones, son, en principio, superables. Lo mismo ocurre con aquellas formas de opacidad que dependen de las limitaciones particulares del agente que examina el sistema, de manera que lo que puede resultar opaco por su sofisticación técnica para un investigador sin formación especializada, puede no serlo para otro con mayores conocimientos.

Siguiendo lo mencionado previamente, lo que el análisis siguiente propone examinar son precisamente aquellas formas de opacidad que presentan problemas de reductibilidad. Según la distinción realizada por Alvarado en “Explaining Epistemic Opacity” (2021) estás serán las que dependen de circunstancias no contingentes del agente, y que podrían ser verdaderamente irreductibles en la medida en que dependen de rasgos constitutivos del propio proceso computacional o del fenómeno estudiado en sí. Alvarado denota esto como una *opacidad epistémica esencial*, que en este estudio encuentra su análogo en las opacidades irreductibles.

## 3.2 Tipos de opacidad y su irreductibilidad

La falta de inteligibilidad y de comprensibilidad están profundamente interrelacionadas en las redes neuronales profundas, por lo que ninguna forma de opacidad puede atribuirse exclusivamente a una u otra dimensión. Sin embargo, dado que la primera responde principalmente a problemas relativos a la estructura interna del sistema, mientras que la segunda apunta en mayor medida a la relación del sistema con el mundo externo que estudia, resulta útil adoptar la distinción entre opacidad interna y opacidad externa que Bjerring, Mainz y Munch (2025) emplean en su análisis de los límites de la inteligencia artificial explicable. Esta distinción permitirá examinar en qué medida cada forma de opacidad constituye un obstáculo en el ámbito científico.

### 3.2.1 Opacidad interna

La opacidad interna concierne al funcionamiento interno del modelo. Florian Boge, en "Two Dimensions of Opacity and the Deep Learning Predicament" (2021), denomina esta dimensión “h-opacity” (how-opacity), u opacidad del cómo. Esta opacidad se ve asociada al modo en que una red neuronal profunda modifica internamente la función que implementa a partir de los datos de entrenamiento, puesto que se desconocen las transformaciones específicas que se suceden a lo largo de las capas, y cómo el proceso de entrenamiento condujo al comportamiento final del sistema.

Søgaard (2024), en su artículo "On the Opacity of Deep Neural Networks", desglosa la “h-opacity” en dos formas más precisas. La primera es la “inference-opacity” u opacidad de inferencia, que refiere a

cuando expertos humanos no pueden, tras su inspección, explicar por qué una red neuronal profunda predice un output y a partir de un input x concreto. La segunda es la “training-opacity” u opacidad de entrenamiento, la cual se da cuando expertos humanos no pueden, tras su inspección, explicar cómo los parámetros de la red fueron inducidos a partir de los datos de entrenamiento. Søgaard identifica, además, cinco propiedades estructurales de las DNN como posibles fuentes de estas formas de opacidad: el tamaño del modelo, la continuidad de sus funciones, la no linealidad de sus activaciones, su carácter instrumental (la ausencia de significado semántico en sus parámetros) y la incrementalidad de su entrenamiento (el hecho de que los parámetros se actualicen iterativamente).

Una manifestación que se podría reconocer en la combinación de varias de estas propiedades estructurales, en particular del carácter instrumental de los parámetros y de la incrementalidad del entrenamiento, es lo que Raleigh y Knoks (2025) denominan “polysemanticity" o polisemanticidad. La intuición detrás de los intentos de interpretabilidad de las redes neuronales era que cada neurona pudiera interpretarse como una unidad funcional con un significado relativamente claro, es decir, que cada unidad de la red fuera sensible a una categoría determinada. Sin embargo, en la práctica se observa que no funcionan así, sino que "las neuronas individuales en las capas ocultas de una DNN son frecuentemente activadas por múltiples características diferentes y no relacionadas entre sí" (p.9). La polisemanticidad refiere a que las representaciones no están localizadas de forma limpia en unidades individuales sino distribuidas y superpuestas a lo largo de la red, lo que hace imposible asignar un significado estable a ninguna unidad individual y constituye, por tanto, un obstáculo fundamental para la inteligibilidad del sistema.

### 3.2.2 Perspectivas de reducción de la opacidad interna

Ahora bien, las perspectivas de reducir este tipo de opacidades se basan en un número limitado de investigaciones. Con respecto a la “h-opacity”, la Inteligencia Artificial Explicable (Explainable AI, XAI) tiene como objetivo desarrollar técnicas capaces de hacer interpretables las decisiones de este tipo de redes. Bjerring, Mainz y Munch, en "Deep Learning Models and the Limits of Explainable Artificial Intelligence" (2025), mencionan entre las técnicas más extendidas el SHAP (Shapley Additive Explanations), que interpreta el comportamiento de una red neuronal profunda especificando la contribución que cada característica del input realiza a un output concreto. Para ello, se basa en el valor de Shapley, procedente de la teoría de juegos cooperativos, que permite cuantificar la importancia relativa de cada variable en una predicción determinada. Como resultado, SHAP proporciona una lista de características ordenadas según su contribución al output, lo que permite al investigador identificar qué aspectos del input han sido más relevantes para la predicción sin necesidad de examinar directamente los miles de millones de parámetros internos del modelo.

Bjerring, Mainz y Munch argumentan que la efectividad de la técnica depende críticamente del tipo de opacidad que se pretende reducir. En lo que respecta a la opacidad interna, o específicamente a la opacidad de inferencia, reconocen las limitaciones asociadas, como la gran capacidad computacional requerida para el cálculo de los valores de Shapley, sin embargo, declaran que "existen razones para ser optimistas respecto a que las técnicas de XAI puedan contribuir a reducir la opacidad interna de los modelos de aprendizaje profundo" (p. 11). Ante estas perspectivas, cabe resaltar que no se aspira a una transparencia total, precisamente porque las explicaciones que el XAI produce son simplificaciones post-hoc, de manera que pueden ser satisfactorias para el investigador y ser, sin embargo, inexactas respecto a lo que la red hace internamente.

A modo de contraste, ante la imposibilidad de reducir por completo la opacidad interna se puede recurrir al "test de φ" de Søgaard (2024). Este parte de modelos generalmente considerados transparentes, como los bosques aleatorios mencionados previamente, o los clasificadores Naive Bayes, modelos que calculan la probabilidad de que un dato pertenezca a una categoría asumiendo que sus variables son independientes entre sí. Si al introducir en ellos una de las cinco propiedades características de las DNN (tamaño, continuidad, no linealidad, instrumentalidad o incrementalidad), el modelo pierde transparencia, esa propiedad puede considerarse una condición suficiente de opacidad. El resultado del test demuestra que los bosques aleatorios pueden ser completamente transparentes a pesar de su no linealidad, es decir, que la relación entre sus entradas y salidas no siga una proporción simple y predecible; y que los clasificadores Naive Bayes lo pueden ser a pesar de su que sus valores internos varían de forma gradual a lo largo de un espectro numérico (un formalismo matemático continuo en lugar de discreto). La conclusión de Søgaard es que la opacidad de las redes neuronales profundas no deriva de ninguna propiedad estructural aislada y que por ello no puede reducirse mediante la superación parcial de ninguna de ellas.

En respuesta a este problema se encuentran las investigaciones en la línea de la *interpretabilidad mecánica*, las cuales tratan de comprender el funcionamiento interno de las redes neuronales profundas a partir de componentes individuales, y cuyo instrumento más desarrollado son los "sparse autoencoders". La idea central de estos sistemas es forzar al modelo a expresar sus representaciones internas de manera más "simple". Entrenan una segunda red que toma como input las activaciones internas de la red original e intenta reconstruirlas, pero penalizando las soluciones en las que muchas unidades se activan simultáneamente (propiedad de la polisemanticidad) (Raleigh y Knoks, 2025, p. 10). Si el sistema puede representar la misma información usando pocas unidades muy especializadas en lugar de muchas unidades que responden a características dispares, aprenderá a asignar funciones más concretas y estables a cada componente. En la medida en que esta jerarquía produce componentes más especializados, los sparse autoencoders parecen mitigar la polisemanticidad acercando el sistema

a una relativa inteligibilidad. Sin embargo, desde la perspectiva de Søgaard, esta intervención continúa sin poder resolver el problema de la opacidad de fondo. Los sparse autoencoders reducen la polisemanticidad sin modificar el tamaño del modelo, su no linealidad o su proceso de entrenamiento iterativo. Como señala el test de Søgaard, “Si tanto φ como ψ son propiedades de una DNN y fuentes de opacidad, eliminar φ es insuficiente para hacer al sistema transparente” (p.3) . Otros investigadores son críticos también con el alcance de estas técnicas, Juan M. Durán en "Against Epistemic Transparency of Algorithms" (2024) señala que cualquier intento de explicar una red neuronal profunda mediante otro modelo, como hacen los sparse autoencoders, introduce una nueva capa de opacidad. El modelo explicativo se vuelve, él mismo, un sistema cuyos mecanismos internos requieren justificación, produciendo un *regressus* que no puede resolverse apelando a más explicaciones del mismo tipo.

Lo que puede afirmarse, sin contradecir estas críticas y la propia opinión de Bjerring, Mainz y Munch, es que la reducción parcial de este tipo de opacidad puede cumplir funciones epistémicas relevantes en contextos concretos como comprobar la relevancia de ciertas variables o detectar posibles errores. No obstante, en tanto que esto no equivale a una transparencia global del sistema, la opacidad interna sigue presentándose como irreductible. Por ello, en el caso de sistemas complejos que buscan la explicación de fenómenos, como es el caso de AlphaFold2, su justificación dependerá de procedimientos externos.

### 3.2.3 Opacidad externa

La opacidad externa refiere a una relación distinta y más global de la red neuronal profunda, la que esta mantiene con el ámbito o mundo que pretende estudiar. Bjerring, Mainz y Munch, distinguen dos subcategorías dentro de la opacidad externa: la opacidad de enlace y la opacidad estructural.

La opacidad de enlace surge cuando existe una desconexión entre el modelo y el sistema real que pretende representar: “Cuando existe incertidumbre respecto a si el modelo de aprendizaje profundo está empíricamente respaldado y adecuadamente vinculado al fenómeno que pretende representar, puede sostenerse que el modelo es opaco en términos de enlace” (Bjerring, Mainz y Munch, 2025, p.12). Boge (2021) aborda esta misma dimensión bajo el nombre de w-opacity, world-opacity u "opacidad del qué" en adición a la “opacidad del cómo” anterior. Si esta última refería al funcionamiento del sistema, la “w-opacity” refiere a qué “ha aprendido” realmente la red neuronal profunda sobre el mundo. Esta opacidad abarca desde la ausencia de correspondencia semántica entre los valores internos de la red y las categorías con las que describimos el mundo, una falta de inteligibilidad, hasta la incomprensibilidad de la red. La incomprensibilidad es dada principalmente por los rasgos complejos que la red neuronal profunda es capaz de extraer autónomamente de los

datos y que le permiten discriminar con éxito, pero que ningún investigador especificó de antemano ni puede reconstruir a posteriori. Según Boge, es esta capacidad para "descubrir" lo relevante para el conocimiento científico: "es exactamente la capacidad de las DNN para *"descubrir automáticamente"* características importantes y complejas lo que impulsa la predicción y, al mismo tiempo, proporciona una base para la comprensión de los mecanismos subyacentes" (p.28). Esto se vuelve problemático porque puede conocerse el fenómeno a estudiar, pero desconocer si la red está correctamente conectada con él, de manera que, de usarse en una investigación, la red pierde la forma directa de justificación de sus resultados.

La opacidad externa estructural, según Bjerring, Mainz y Munch (2025), surge cuando el propio fenómeno externo que el modelo estudia es, en sí mismo, incomprensible, cuando la red detecta regularidades pero tal vez no existan conceptos científicos para entenderlos. En este caso, la opacidad reside tanto en la red como en el fenómeno "descubierto". Faries y Raja describen este fenómeno en "Black Boxes and Theory Deserts: Deep Networks and Epistemic Opacity in the Cognitive Sciences" (2022) como "desierto teórico".

### 3.2.4 Perspectivas de reducción de la opacidad externa

En lo que respecta a la opacidad de enlace, Bjerring et al. (2025) sostienen que los métodos contrafactuales, otra de las principales técnicas de XAI, pueden contribuir parcialmente a reducirla. Mientras que SHAP permite identificar cuánto contribuye cada variable a una predicción determinada, las *explicaciones contrafactuales* indican qué valores de las variables de entrada deberían modificarse para obtener un resultado diferente. En este sentido, Wachter et al. (2018) describen la forma general de una explicación contrafactual del siguiente modo:

"El output O fue devuelto por el modelo de aprendizaje profundo porque las características 1, 2, …, n tenían asociados los valores v1, v2, …, vn. Si las características 1, 2, …, n hubieran tenido asociados los valores v1, v2, …, vn, y los valores de todas las demás características del modelo hubieran permanecido constantes, entonces el output O habría sido devuelto por el modelo de aprendizaje profundo" (Wachter et al., 2018, p. 9).

Para identificar estos "contradatos", las técnicas minimizan una función de pérdida que compara distintas perturbaciones de las variables de entrada: el "contradato" óptimo es el que representa la desviación mínima del input original capaz de cambiar la predicción.

Otra vía, también comparativa, para reducir la opacidad de enlace son los benchmarks o “pruebas comparativas de rendimiento”. Los benchmarks utilizan un patrón externo de referencia para evaluar si los outputs se ajustan de manera fiable a este. Durante el proceso de validación del AlphaFold2, Jakob Ortmann en “Of opaque oracles: epistemic dependence on AI in science poses no novel problems for social epistemology” señala cómo la aceptación científica del sistema se dio, en parte, mediante la realización de pruebas comparativas independientes como el uso de benchmarks.

Los benchmarks utilizados fueron estructuras tridimensionales de proteínas previamente no publicadas, lo cual impedía que el sistema se limitara a reproducir ejemplos conocidos. Este proceso se llevó a cabo a través del CASP, el Critical Assessment of protein Structure Prediction, una competición científica internacional que se celebra bienalmente desde 1994 y que supone un estándar de referencia en la evaluación de los métodos de predicción de estructuras proteicas. El CASP realiza esta evaluación en condiciones ciegas, es decir, los participantes reciben secuencias de proteínas cuya estructura ha sido determinada experimentalmente pero no ha sido publicada, y sus predicciones son comparadas con esas estructuras reales por evaluadores independientes. En la edición de 2020, CASP14, AlphaFold2 alcanzó un nivel de precisión que los evaluadores determinaron comparable al de la determinación experimental, con una mediana de puntuación GDT (Global Distance Test), la cual evalúa precisamente la similitud entre una estructura proteica predicha y una estructura determinada experimentalmente, de 92,4 sobre 100 en el conjunto de dominios evaluados.

Gracias a este proceso comparativo puede argumentarse que el modelo mantiene cierta conexión con aquello que pretende representar, pero la opacidad interna continúa siendo irreductible de forma completa, al igual que con el uso de las explicaciones contrafactuales. Este límite se debe a que la opacidad de enlace se encuentra delimitada por la propia falta de inteligibilidad y comprensibilidad de la red, las cuales dificultan su relación con el fenómeno estudiado de forma transparente. Se vuelve esencialmente irreductible por la naturaleza misma de estas redes.

Reconocer esta irreductibilidad no equivale, sin embargo, a concluir que los resultados alcanzados mediante redes neuronales profundas carezcan de valor epistémico. Ello encuentra un precedente en la reflexión de Humphreys (2004) sobre la ciencia contemporánea, así como los resultados de un instrumento cuyo mecanismo era observable en términos empiristas tradicionales no eran incorporados a la práctica científica sin la observación directa del investigador como parte del método, las redes neuronales profundas tampoco serán incorporadas a ella sin un método de respaldo equivalente. Lo que ese método ha de proporcionar no es necesariamente un conocimiento exhaustivo del proceso interno, sino una razón de fiabilidad que, de no poder ser *directa*, puede perfectamente ser externa. En esta línea, Durán, en "Beyond transparency: Computational reliabilism as an externalist epistemology of algorithms" (2025), concluye que la justificación epistémica de los outputs de un

sistema algorítmico no proviene de hacer el sistema más inteligible o comprensible, sino de identificar indicadores externos de fiabilidad independientes de su funcionamiento interno. La tesis que se defenderá en lo que sigue es que esta justificación externa no solo es posible sino suficiente: que la comunidad científica dispone ya de prácticas, normas e instituciones que permiten producir conocimiento bajo condiciones de opacidad, y que el caso de AlphaFold2 ejemplifica cómo estos mecanismos se llevan a cabo en la práctica científica contemporánea.

En cuanto a la reducción de la opacidad estructural, se ha visto que el problema reside en la ausencia de un *marco conceptual*, más que en una justificación entre resultado y fenómeno, por ello, la este tipo de opacidad aparece como la forma más irreductible de opacidad externa. En lo que respecta a AlphaFold2, este no constituye un caso de desierto teórico en el sentido de Faries y Raja, puesto que las estructura tridimensionales que estudia se encuentran teóricamente bien delimitadas, el objeto de estudio dispone de conocimiento externo suficiente como para que sus resultados sean reconocidos y evaluados posteriormente. Si el programa generase resultados correctos pero de los que la ciencia actual no dispone de conocimiento, simplemente no serían reconocidos ni evaluados, y tal vez el programa, en lugar de un instrumento de investigacion podria ser utilizado con otro papel epistémico.

# 4. Opacidad en la ciencia

## 4.1 Modelos mediadores

Antes de examinar las prácticas de justificación científica dentro de la opacidad de las redes neuronales profundas, conviene detenerse en una demarcación previa de qué son estos otros papeles epistémicos que pueden desempeñar estas. Se ha analizado la capacidad de producir resultados aceptables de estos instrumentos, sin embargo, su capacidad epistémica podría trascender esto.

Morrison y Morgan, en *Models as mediators* (1999), describen el concepto de instrumento de investigación, que fue aplicado a las DNN en la sección 2.1 de este estudio, con el fin de explicar lo que en el libro se denota como “modelo mediador”, algo cercano a la función de un instrumento de investigación, pero que produce conocimiento en autonomía parcial respecto a la teoría y el mundo: "Los modelos no se sitúan en el medio de una estructura jerárquica entre la teoría y el mundo [...] concebimos los modelos como fuera del eje teoría-mundo. Es esta característica la que les permite mediar eficazmente entre ambos"(Morrison y Morgan, 1999, p. 18). Las DNN pueden entenderse en primer lugar como instrumentos de investigación por su capacidad de detección de patrones y predicción. No obstante, pueden a su vez desempeñar una función análoga a la de los modelos mediadores.

Este concepto de “modelo”, cabe mencionar, se emplea en un sentido distinto al que se ha utilizado en este estudio previamente, referente este a las redes neuronales profundas. Un “modelo mediador” no apela específicamente a modelos computacionales avanzados, en su lugar, usan este término en relación con modelos científicos. Un ejemplo presentado de esto sería el descrito en Models as Mediators (1999) sobre el modelo de London de la superconductividad (p.182). Este modelo, no fue derivado de la teoría electromagnética existente, se ideó a partir de una conceptualización completamente nueva del fenómeno, convirtiéndose simultáneamente en representación y en nueva comprensión teórica.

Para comprender con mayor precisión qué tipo de relación representativa mantienen las DNN con el fenómeno que estudian, resulta útil recurrir también al ejemplo propuesto en el capítulo 5 del volumen de Morrison y Morgan. En este se realiza una distinción entre el modelo como representación y el modelo como representante. En los modelos mediadores clásicos, hay una correspondencia entre la estructura interna del modelo y la estructura del mundo que estudia, Hughes observa que en simulaciones computacionales como el modelo Ising, un modelo físico apoyado por simulación computacional, esto no sucede. El modelo Ising no funciona como representación, mas sí como un representante de una clase entera de sistemas (Hughes, 1999, p. 127), capturando las propiedades estructurales que todos los miembros de una clase comparten, y siendo así cómo puede proporcionar "la mejor explicación disponible del comportamiento de uno de los sistemas, sin ser una representación fiel de cualquier sistema de este tipo" (Hughes, 1999, p. 128). Hughes define estas formas de representación como “recurso epistémico”, “Aprovechamos este recurso al observar lo que podemos demostrar con el modelo. Esto puede implicar una demostración en el sentido decimonónico de la palabra, es decir, una prueba matemática. [...] O puede implicar una demostración en el sentido actual del término” (Hughes, 1999, p.128) En ambos casos, reconoce una función epistémica en tales modelos, ya que ambos estarían permitiendo, de forma última, acceder a un conocimiento más profundo del objeto estudiado.

Esta caracterización resulta especialmente pertinente en casos de opacidad estructural, en los cuales las redes neuronales profundas se limitarían a capturar regularidades sin un marco concreto. Por ejemplo, de no estar AlphaFold2 circunscrito a un dominio lo suficientemente delimitado, ¿podría legitimarse su uso como “recurso epistémico” o “modelo mediador? La condición para demostrar que cierta regularidad corresponde a un fenómeno y no es un simple artefacto de la red neuronal profunda es, al igual que con la opacidad de enlace, los indicadores externos de fiabilidad. En este caso, se trataría o bien de la convergencia de distintos modelos sobre las mismas regularidades, lo cual proporcionaría una razón para pensar que lo detectado no depende del programa si no que podría tratarse de un *representante* de fenómenos sí circunscritos en un marco conceptual conocido. O bien

de un desarrollo teórico posterior que permitiese interpretar lo detectado y proceder a su evaluación, de otra forma el conocimiento simplemente no sería alcanzable, careceríamos de formas de detección de este.

Independientemente del papel que una red neuronal profunda lleve a cabo, ya sea como instrumento de investigación o como modelo mediador, la aceptación de ellos como contribución epistémica dentro de la comunidad científica depende de una demostración, de una *justificación*. En qué momento del proceso científico intervienen estas tecnologías es así determinante en su valoración, puesto que no será lo mismo emplear una DNN como orientación previa en la formulación de una hipótesis que como respaldo de una conclusión ya establecida.

## 4.2 Descubrimiento y justificación

A fin de precisar en qué consiste exactamente las funciones epistémicas en las redes neuronales profundas, el recorrido que Duede (2023) hace de la filosofía de la ciencia, a través de Reichenbach y Popper, es especialmente esclarecedor. Reichenbach (1938) estableció una distinción fundamental entre dos momentos separables de la práctica científica, el contexto de descubrimiento, como proceso mediante el cual se llega a formular una hipótesis, y el contexto de justificación. como procedimiento por el cual dicha hipótesis es evaluada y validada. Esta distinción permitía admitir que la concepción de una idea científica pudiera ser intuitiva, azarosa, o incluso opaca, siempre que el resultado final pudiera someterse posteriormente a criterios públicos de justificación. Duede utiliza esta idea para defender que, cuando los outputs de una DNN son utilizados como orientaciones en un descubrimiento, su opacidad resulta irrelevante, asimilando su papel heurístico a otras prácticas habituales tales como la abducción, la analogía o los experimentos mentales. Del mismo modo que la intuición de un matemático no justifica por sí sola un teorema pero puede orientar su formulación, los resultados de una DNN no justifican una hipótesis aunque sí pueden señalar conexiones que de otro modo habrían permanecido fuera del alcance de la intuición humana.

En “Understanding from Machine Learning Models” (2022), Emily Sullivan distingue entre dos tipos de explicaciones que una modelo computacional como las DNN puede proporcionar en función de la cuestión científica que un investigador esté tratando de responder. Por un lado, las explicaciones “*how-possibly*” identifican una posibilidad a las causas o dependencias de algún fenómeno, limitándose a mostrar que una determinada relación causal o dependencia es compatible con los datos disponibles. Las explicaciones “*how-actually*”, en cambio, establecen "cómo el fenómeno objetivo es causado efectivamente o cuáles son las dependencias reales concernientes al fenómeno" (p. 15), así, responden a lo que *realmente* se da. Como argumenta Sullivan, los modelos de aprendizaje profundo

son capaces de proporcionar explicaciones *how-possibly* sin poder proporcionar las *how-actually* que un contexto de justificación exigiría. Esta concepción permite precisamente el papel epistémico de las DNN como modelos mediadores u orientativos en el ámbito científico. Cuando las opacidades son irreductibles, el modelo por sí mismo solo puede ofrecer explicaciones “how-possibly”, mostrando que cierta relación, dependencia o patrón podría ser relevante, pero no que lo sea efectivamente en el fenómeno real. Sus resultados funcionan como orientaciones heurísticas para la investigación, de manera que la *justificación* posterior de su uso dependerá siempre de criterios externos.

Esta conclusión converge con la posición de Juan Durán en "Beyond Transparency: Computational Reliabilism as an Externalist Epistemology of Algorithms" (2025), así como con la segunda reclamación de Duede y Davey sobre las “pruebas de una afirmación matemática” de que la necesidad es, lejos de recuperar la transparencia, establecer qué tipo de justificación es posible y suficiente. De ello depende el propio estatuto epistémico de estas tecnologías dentro de la ciencia, una red neuronal profunda sin criterios externos de fiabilidad será irremediablemente relegada a un papel heurístico, orientativo o imaginativo, como tantos otros recursos en el ámbito científico, pero no podrá ser reconocida como instrumento de investigación ni como modelo mediador y así sus resultados formar parte legítima del conocimiento científico.

# 5. Justificación científica en condiciones de opacidad

## 5.1 Criterios externos de fiabilidad

La forma en que las redes neuronales profundas se incorporan a nuevos procesos de justificación y validación científica viene dada por el contexto de sistemas computacionales anteriores que ya habían planteado problemas análogos, aunque de menor alcance.

El problema de la justificación de modelos computacionales complejos en el contexto de las simulaciones científicas, por ejemplo, provenía de la imposibilidad de verificación matemática. Eric Winsberg, en su libro “Science in the Age of Computer Simulation” (2010), muestra cómo la metodología de justificación de los resultados en sistemas complejos como simuladores, el cual se dividía en verificación y validación, era inadecuado para la naturaleza de estos modelos. La verificación consiste en la comprobación de que los outputs del sistema aproximen las soluciones verdaderas a las ecuaciones del modelo original, y la validación, la comprobación del modelo elegido como representación adecuada del sistema real. En la práctica, según Winsberg, estas dos actividades no se realizan separadamente, habitualmente lo que ocurre es que los resultados son “todos ellos sancionados a la vez: los simulacionistas intentan maximizar la fidelidad a la teoría, al rigor

matemático, a la intuición física y a los resultados empíricos conocidos. Pero es la confluencia simultánea de estos esfuerzos, más que el establecimiento de cada uno por separado, lo que en última instancia nos da confianza en los resultados" (Winsberg, 2010, p. 23). La incorporación de redes neuronales profundas a la práctica científica debe entenderse por ello como una continuación y ampliación de los problemas epistemológicos planteados por estas simulaciones computacionales.

Durán y Formanek formalizan este principio en "Grounds for Trust: Essential Epistemic Opacity and Computational Reliabilism" (2018), proponiendo un *"fiabilismo computacional"*. Toman en este artículo inspiración en el fiabilismo de proceso de Goldman, el cual viene a establecer que un agente está justificado en creer una proposición si esa creencia resulta de un proceso fiable. En el caso computacional, la creencia de un investigador estaría justificada si los resultados del sistema siguiesen un proceso fiable de formación de creencias, es decir, si y sólo si siguiesen una tendencia consistente de producir resultados correctos en situaciones similares.

Dado que en condiciones de opacidad, los resultados bien pueden no ser correctos, la solución que proponen es lo que denominan una cadena retrospectiva de fiabilidad, una adaptación de la metodología de justificación que Winsberg trataba también de ajustar. En la construcción de esta cadena de fiabilidad, Durán y Formanek nombran cuatro fuentes contribuyentes, en distinto grado, a la atribución de fiabilidad en un sistema computacional complejo:

- Métodos de verificación y validación, ya discutidos a propósito de Winsberg, que constituyen la fuente más robusta por su dependencia de demostraciones matemáticas. Por un lado, garantizan que la implementación de las teorías establecidas se lleva a cabo correctamente; por otro, proporcionan razones sólidas para confiar en los resultados porque estos coinciden (idealmente) con los datos empíricos.

- Análisis de robustez, consiste en la examinación de un conjunto de modelos para determinar si todos predicen un resultado común, denominado "propiedad robusta", y en identificar las estructuras del modelo que generan esa propiedad con el fin de formular "teoremas robustos", descritos como "una afirmación condicional que vincula la estructura común con la propiedad robusta, precedida de una cláusula ceteris paribus" (Durán y Formanek, 2018, p. 17). De manera similar a los métodos contrafactuales y otras técnicas de XAI, que realizan tareas similares sobre un modelo individual, el análisis de robustez estudia también diferentes resultados, aunque de modelos distintos.

- Historia de implementaciones exitosas y fallidas, una evidencia acumulada sobre las condiciones bajo las cuales el sistema produce resultados fiables. La confianza científica

sustentada por un rendimiento reiterado frente a pruebas independientes de predicción estructural.

- El conocimiento experto. Durán y Formanek recalcan que, en el contexto de los sistemas computacionales, el experto no es necesariamente alguien que domina simultáneamente la teoría subyacente y su implementación computacional, el matemático que conoce en profundidad la teoría pero ignora su implementación es tan experto como el informático que sabe implementarla pero desconoce la teoría, lo que define al experto es la pericia en alguno de los dominios relevantes para evaluar su adecuación. Citando a Claus Beisbart en el mismo artículo, "los científicos creen los resultados de sus simulaciones porque confían en las suposiciones sobre las que se construyen" ((Durán y Formanek, 2018, p. 22), ellos serían los desarrolladores, pero son entonces los expertos quienes deben analizar esas suposiciones. La debilidad que presentan, sin embargo, es el potencial parcial o sesgado del experto, por lo que se exige que sea complementado, siempre, por las otras tres fuentes.

## 5.2 Criterios externos usados en AlphaFold 2

El caso de AlphaFold, permite vislumbrar con mayor concreción qué forma toma en la práctica la justificación externa que el fiabilismo presenta en la sección previa.

Los benchmarks descritos en la sección III como como prueba comparativa de rendimiento, constituirían el equivalente más directo a la verificación y validación en términos de Durán y Formanek. Una evaluación que compara los outputs del sistema con estructuras determinadas experimentalmente, estableciendo así la correspondencia entre las predicciones del modelo y propiedades reales de las proteínas.

Con respecto al análisis de robustez y una historia de implementaciones exitosas y fallidas, Jakob Ortmann, en “Of opaque oracles: epistemic dependence on AI in science poses no novel problems for social epistemology” (2025) describe qué otras formas de validación se han llevado a cabo en el caso de AlphaFold2 y que podrían corresponderse con estas. Ortmann menciona cómo laboratorios de todo el mundo han llevado a cabo comprobaciones de robustez extensivas con el fin de identificar las condiciones bajo las cuales el sistema produce resultados fiables y aquellas bajo las cuales falla, “Si los investigadores confían en AlphaFold2, no confían simplemente en una caja negra de la que no se puede saber nada, sino en un sistema que ha sido sometido a exhaustivas pruebas de robustez.” (Ortmann, 2025, p.17).

Ortmann, considera este proceso de justificación de AlphaFold2 suficiente a través de un paralelismo con las nociones de Sullivan del how-possibly y how-actually. En este caso, la pregunta del "por qué", referida a los mecanismos internos que hacen que el sistema funcione como funciona, y la pregunta del "si", referida a si el sistema produce resultados fiables bajo condiciones determinadas. Su argumento es que, la respuesta a la pregunta del "si", o del "how-possibly" de Sullivan, es suficiente para justificar la dependencia epistémica en el sistema: "El conocimiento relevante que cualquier *depositador de confianza* necesita poseer no es necesariamente el conocimiento del por qué, sino más bien todo aquello que sea necesario para justificar la confianza epistémica en un instrumento [...] el conocimiento del *si* puede ser suficiente." (Ortmann, 2025, p. 11).

De hecho, esta condición fue suficiente para una adopción masiva del sistema por parte de la comunidad científica. Más de dos millones de usuarios de 190 países recurrieron a AlphaFold2 en distintos dominios como el diseño de fármacos y la comprensión de mecanismos de resistencia antibiótica. Ese proceso distribuido de uso y evaluación por parte de una comunidad científica experta en dominios muy distintos sería la expresión más amplia del conocimiento experto, una opinión colectiva y acumulada de una comunidad que ha incorporado el sistema a su práctica científica cotidiana. Además, la concesión del Premio Nobel significó que el rendimiento del sistema en el CASP14 y sus numerosos análisis y resultados posteriores representaban una contribución científica suficientemente significativa para merecer el reconocimiento más alto de la comunidad científica internacional (Royal Swedish Academy of Sciences, 2024). Un juicio realizado en consonancia con las verificaciones, validaciones, y análisis previos, no sustentados en la total comprensión interna del sistema.

## 6. Consideraciones finales

¿Es la opacidad de un instrumento como las redes neuronales profundas incompatible con su uso como herramientas epistémicas? A través de este estudio se ha argumentado que no necesariamente. Su complejidad y naturaleza opaca, en distintos niveles, junto con su aplicación a la práctica científica, han suscitado dudas y debates sobre la validez de sus procesos y resultados.

La primera de esas preguntas era si el funcionamiento de estas redes constituye efectivamente una forma de opacidad reducible. La respuesta es que las formas de opacidad que derivan de la falta de inteligibilidad y de comprensibilidad constitutivas de las DNNs no pueden serlo del todo. Si bien las técnicas desarrolladas con este fin ofrecen aproximaciones útiles, y suponen un avance tecnológico que no debe desestimarse como la Explainable AI, esperar que estas soluciones técnicas resuelvan por completo la opacidad sería una empresa vana e inabarcable. Como señalaron Duede y Davey, un

resultado producido por un sistema opaco requeriría así, para ser epistémicamente legítimo, algo funcionalmente equivalente a una prueba de su afirmación, un procedimiento externo que establezca su validez con independencia de la comprensión del proceso interno que lo generó. La cuestión de la opacidad es desplazada hacia el problema de la fiabilidad.

Establecido ese punto, fue necesario determinar qué tipo de aporte ofrecen las redes neuronales profundas al conocimiento científico. En sí mismas, sin criterios externos, sus resultados solo pueden ofrecer explicaciones how-possibly, orientaciones heurísticas comparables a la abducción o experimentos mentales. Sin embargo, para una justificación suficiente, debe existir un vínculo adecuado entre el modelo y el fenómeno de interés, cuando este se alcanza mediante métodos de verificación, análisis de robustez, historia de implementaciones y conocimiento experto, como en el caso de AlphaFold2, pueden entonces legítimamente ser considerados instrumentos de investigación o modelos mediadores.

Cabe mencionar, que existen dimensiones del problema que este trabajo ha dejado deliberadamente fuera de su alcance, qué riesgos implica su uso en determinados dominios de alta responsabilidad, como la medicina, la justicia o las políticas públicas. La pregunta en estos casos no recae sobre la legitimidad de su validez, más bien un debate de responsabilidades sobre el conocimiento experto y sus criterios o intereses particulares, cuestiones que la epistemología puede ayudar a formular, pero que no puede responder por sí sola.

Lo que sí puede afirmarse, a la luz de lo argumentado, es que las redes neuronales profundas no explican en el sentido en que lo hace una demostración matemática o un mecanismo causal articulado, pero son altamente capaces de orientar, e incluso descubrir; y, adecuadamente vinculadas al fenómeno, permiten avanzar en su comprensión. Si la ciencia ha sabido históricamente incorporar instrumentos cuyo funcionamiento excedía la comprensión inmediata de quienes los usaban, es porque, en la línea de Humphreys, el límite de lo epistémicamente accesible no es fijo. Las redes neuronales profundas son por tanto el ejemplo más actual de la necesidad de criterios de fiabilidad en el proceso científico. Tanto en AlphaFold2 como en otras redes neuronales profundas la opacidad no impide la producción de conocimiento, ella se ha llevado a cabo, y seguirá haciéndolo dentro de unos marcos conceptuales fuertes y una verificación empírica externa a las regularidades que puedan ser detectadas.

## Bibliografía


Alvarado, R. (2021). Explaining epistemic opacity [Preprint]. PhilSci-Archive. https://philsci-archive.pitt.edu/19384/

Bassan, S., Amir, G., Corsi, D., Refaeli, I., & Katz, G. (2023). *Formally explaining neural networks within reactive systems*. arXiv. https://doi.org/10.48550/arXiv.2308.00143

Bjerring, J. C., Mainz, J., & Munch, L. (2025). *Deep learning models and the limits of explainable artificial intelligence.* Asian Journal of Philosophy, 4, Article 22. https://doi.org/10.1007/s44204-024-00238-8

Boge, F. J. (2021). Two dimensions of opacity and the deep learning predicament. *Minds and Machines, 32*(1), 43–75. https://doi.org/10.1007/s11023-021-09569-4

Burrell, J. (2016). How the machine "thinks": Understanding opacity in machine learning algorithms. *Big Data & Society*. https://doi.org/10.1177/2053951715622512

Creel, K. A. (2020). Transparency in complex computational systems. *Philosophy of Science, 87(*4), 568–589. https://doi.org/10.1086/709729

Dominici, G., Barbiero, P., Espinosa Zarlenga, M., Termine, A., Gjoreski, M., Marra, G., & Langheinrich, M. (2025). *Causal concept graph models: Beyond causal opacity in deep learning.* International Conference on Learning Representations. https://openreview.net/forum?id=lmKJ1b6PaL

Duede, E. (2023). Deep learning opacity in scientific discovery. *Philosophy of Science*, 90(5), 1089–1099. https://doi.org/10.1017/psa.2023.8

Duede, E., & Davey, K. (2025). A priori knowledge in an era of computational opacity: The role of artificial intelligence in mathematical discovery. *Philosophy of Science*, 92(5), 1394–1404. https://doi.org/10.1017/psa.2025.160

Durán, J. M. (2024). Against epistemic transparency of algorithms. *Minds and Machines*, 34, 45. https://doi.org/10.1007/s11023-024-09681-3

Durán, J. M. (2025). Beyond transparency: Computational reliabilism as an externalist epistemology of algorithms. En J. M. Durán & G. Pozzi (Eds.), *Philosophy of science for machine learning: Core issues and new perspectives.* Springer. https://doi.org/10.1007/978-3-031-03083-2_4

Durán, J. M., & Formanek, N. (2018). Grounds for trust: Essential epistemic opacity and computational reliabilism. *Minds and Machines,* 28(4), 645–666. https://doi.org/10.1007/s11023-018-9481-6

Faries, F., & Raja, V. (2022). *Black boxes and theory deserts: Deep networks and epistemic opacity in the cognitive sciences* [Preprint]. PhilSci-Archive. https://philsci-archive.pitt.edu/20492/

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.

Hughes, R. I. G. (1999). The Ising model, computer simulation, and universal physics. En M. Morrison & M. S. Morgan (Eds.), *Models as mediators: Perspectives on natural and social science* (pp. 97–145). Cambridge University Press.

Humphreys, P. (2004). *Extending ourselves: Computational science, empiricism, and scientific method.* Oxford University Press.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Morrison, M., & Morgan, M. S. (1999). Models as mediating instruments. En M. Morrison & M. S. Morgan (Eds.), *Models as mediators: Perspectives on natural and social science* (pp. 10–37). Cambridge University Press.

Müller, V. C. (2025). Deep opacity in AI. En M. Hähnel & R. Müller (Eds.), A companion to applied philosophy of AI. Wiley-Blackwell. https://doi.org/10.1002/9781394238651.ch6

Ortmann, J. (2025). Of opaque oracles: Epistemic dependence on AI in science poses no novel problems for social epistemology. *Synthese*, 205, Article 80. https://doi.org/10.1007/s11229-025-04930-x

Raleigh, T., & Knoks, A. (2025). Clarifying the opacity of neural networks. *Minds and Machines*, 35(4), 1–30. https://doi.org/10.1007/s11023-025-09745-w

Reichenbach, H. (1938). *Experience and prediction: An analysis of the foundations and the structure of knowledge.* University of Chicago Press.

Royal Swedish Academy of Sciences. (2024a). *The Nobel Prize in Chemistry 2024: Scientific background*. NobelPrize.org. https://www.nobelprize.org/prizes/chemistry/2024/advanced-information/

Royal Swedish Academy of Sciences. (2024b). *The Nobel Prize in Chemistry 2024: Popular information.* NobelPrize.org. https://www.nobelprize.org/prizes/chemistry/2024/popular-information/

San Pedro, I. (2024). Degrees of epistemic opacity. *Teorema: Revista Internacional de Filosofía, 43*(2), 5–21. https://dialnet.unirioja.es/servlet/articulo?codigo=9760968

Søgaard, A. (2024). On the opacity of deep neural networks. *Canadian Journal of Philosophy*, *53*(3), 224–239. https://doi.org/10.1017/can.2024.1

Sullivan, E. (2022). Understanding from machine learning models. *The British Journal for the Philosophy of Science, 73*(1), 109–133. https://doi.org/10.1093/bjps/axz035

Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology, 31*(2), 841–887. https://doi.org/10.2139/ssrn.3063289

Winsberg, E. B. (2010). *Science in the age of computer simulation*. University of Chicago Press.

Zerilli, J. (2022). Explaining machine learning decisions. *Philosophy of Science, 89*(1), 1–19. https://doi.org/10.1017/psa.2021.13