Revista Colombiana de Filosofía de la Ciencia
Vol. 24 n.o 48 (2024): 161-194
ISSN: 0124-4620 (papel) & 2463-1159 (electrónico)

Fecha de recepción: 12/03/2023
Fecha de aprobación: 12/03/2024
Fecha de publicación: 03/07/2024
https://doi.org/10.18270/rcfc.4288

La posibilidad de explicación científica a partir de modelos basados en redes neuronales artificiales*

The possibility of scientific explanation from models based on artificial neural networks

Alejandro E. Rodríguez-Sánchez
Universidad Panamericana, Facultad de Ingeniería
Zapopan, México.
aerodriguez@up.edu.mx
https://orcid.org/0000-0003-3397-5261

RESUMEN

En inteligencia artificial, las redes neuronales artificiales son modelos muy precisos en tareas como la clasificación y la regresión en el estudio de fenómenos naturales, pero se consideran "cajas negras" porque no permiten explicación directa de aquello que abordan. Este trabajo revisa la posibilidad de explicación científica a partir de estos modelos y concluye que se requieren de otros esfuerzos para entender su funcionamiento interno. Esto plantea retos para acceder a la explicación científica a través de su uso, pues la naturaleza de las redes neuronales artificiales dificulta a primera instancia la comprensión científica que puede extraerse de estas.

Palabras clave: redes neuronales artificiales; explicación científica; explicabilidad; interpretabilidad; transparencia; cajas negras.

* Este artículo se debe citar: Rodríguez-Sánchez, Alejandro E. "La posibilidad de explicación científica a partir de modelos basados en redes neuronales artificiales". *Revista Colombiana de Filosofía de la Ciencia* 24.48 (2024): 161-194. *https://doi.org/10.18270/rcfc.4288*

ABSTRACT

In Artificial Intelligence, Artificial Neural Networks are very accurate models in tasks such as classification and regression in the study of natural phenomena, but they are considered "black boxes" because they do not allow direct explanation of what they address. This paper reviews the possibility of scientific explanation from these models and concludes that other efforts are required to understand their inner workings. This poses challenges to access scientific explanation through their use, since the nature of Artificial Neural Networks makes it difficult at first instance the scientific understanding that can be extracted from them.

Keywords: Artificial Neural Networks; Scientific Explanation; Explainability; Interpretability; Transparency; Black-boxes.

1. Introducción

En las últimas dos décadas, los sectores de la ciencia y de la tecnología han visto desarrollo y avance respecto a un paradigma del conocimiento que, entendido como una amalgama de tecnologías y métodos, representa una alternativa con la cual predecir y estimar fenómenos sociales o naturales. Este paradigma se refiere al aprendizaje automático (AA), el cual subsume diversos grupos de modelos y algoritmos empleados generalmente para la predicción, la retrodicción y la estimación de un fenómeno de interés (Alzubaidi et ál. 2021; Dhall, Kaur & Juneja 2020; Emmert-Streib et ál. 2020). A la luz de los reportes científicos publicados y de la tecnología generada a partir de ellos, es justo decir que el AA ha sido continuamente impulsado tanto por la comunidad científica como por el sector tecnológico a nivel global para alcanzar sus respectivas metas: ganar más conocimiento y generar valor para un segmento del mercado (Saxe, Nelli & Summerfield 2021; Soniya, Paul & Singh 2015; Tomašev et ál. 2020; Wang, Zhao & Pourpanah 2020). El desarrollo del AA es, pues,

inherente a los beneficios que cosecha en diversos sectores, como el tecnológico y el industrial (Tang et ál. 2019).

Desde un nivel taxonómico alto, el AA pertenece al campo de la inteligencia artificial (IA), ya que los algoritmos, los métodos y las técnicas que en ello se emplean basan su funcionamiento en procesos informáticos, y estos, a su vez, se inspiran en procesos biológicos observados en la naturaleza, como lo es el proceso de aprendizaje y la adaptación a nueva información (LeCun, Bengio & Hinton 2015). Entre ellos se destacan los *modelos basados en redes neuronales artificiales*, de los cuales su funcionamiento guarda semejanza con el procesamiento interno que ocurre en los cerebros biológicos de los animales y de los seres humanos (Aggarwal 2018; Janiesch, Zschech & Heinrich 2021).

Las redes neuronales artificiales (RNA) se usan en espacios técnico-científicos —universidades, centros de investigación y empresas de tecnología— para múltiples aplicaciones y tareas, entre las que vale la pena destacar la predicción, el diagnóstico y el análisis de fenómenos naturales (Goodfellow, Bengio & Courville 2016). Más aún, se sabe que, por su adaptabilidad, la aplicación de las RNA tiene lugar en disciplinas tan disimiles unas de otras como la meteorología (e.g., Abhishek et ál. 2012), la astronáutica o la ciencia de los materiales (Faller & Schreck 1996; Ramprasad et ál. 2017; Silvestrini & Lavagna 2022; Song, Rondao & Aouf 2022). Sobre esto, se puede decir que dichas aplicaciones confirman el potencial y el valor que tienen estos modelos para la ciencia, la tecnología y la sociedad, de ahí el fuerte compromiso para entender su naturaleza. Como se verá, dicho valor repara precisamente en su elemento predictivo más que explicativo con respecto tanto a los objetos que se analizan desde sus elementos funcionales, como a sus propios algoritmos.

En los reportes relacionados con el quehacer de la profesión de la ingeniería (Chen & Liu 2022; Shehab et ál. 2022; Singh et ál. 2009), los modelos

basados en RNA actualmente demuestran potencia de predicción[1] muy superior respecto a otros modelos tradicionales. Esto les posiciona como herramientas de predicción y estimación por excelencia, puesto que su capacidad de abstracción de orden, forma y tendencia de los datos resulta conveniente para algunas aplicaciones y usos donde o bien los modelos existentes tienen un error alto, o no existen modelos que estimen adecuadamente la respuesta estudiada de un fenómeno. Por esto, en la praxis tecnológica y de la ingeniería, las RNA tienden a ser preferidas por sobre otros modelos, ya que son excelentes en lo que se considera *desiderata no-teórica* (Cichy & Kaiser 2019); esto es, se ejecutan rápido en un ordenador, son baratas de construir en contraposición con el tiempo de modelado matemático tradicional para estudios multifactoriales, y constituyen un uso de memoria informática eficiente una vez que están creadas y validadas con datos experimentales. Sin embargo, es importante destacar que el proceso de entrenamiento de las RNA también puede requerir un alto costo computacional y de tiempo, especialmente cuando se trabaja con grandes conjuntos de datos y arquitecturas complejas, como los modelos de lenguaje de gran tamaño[2].

Sin embargo, aun con las primacías que presentan las RNA en cuanto a objetivos de predicción, existe una cuestión relevante en su uso que reside en la explicación desde la perspectiva científica que se puede extraer a partir de ellas. Esto supone un problema relevante en las ciencias pues parece existir cautela en cuanto al valor explicativo que se puede lograr por medio de modelos basados en RNA para tareas de diseño, optimización y exploración de sistemas físicos (Zednik 2021). Este problema nace a raíz de la preocupación latente de la gran potencia de estimación y predicción que ofrecen las RNA para predecir fenómenos naturales sin que necesariamente se entienda, a primera instancia, su estructura interna o la relación que esta guardaría como parámetro explicativo de la regularidad de

[1] Potencia precisada en términos de los valores de medidas estadísticas que relacionan la distancia entre las predicciones de un modelo con respecto a datos experimentales.

[2] Del inglés *large language models*.

un fenómeno (Rudin 2019). Sobre ello, se podría argumentar que la inquietud es legítima, pues no son pocas las ocasiones en las que en el ambiente de las ciencias los científicos se decanten por el uso de un modelo con menores parámetros en los elementos que le conforman, en favor de la explicación que ofrecen las descripciones más sencillas, en lugar del uso de un modelo muy preciso pero complejo como lo suelen ser las RNA (principio metodológico conocido como *navaja de Occam*). El quid se ubica, entonces, en que para esbozar una explicación es requisito conocer el funcionamiento interno de un constructo lógico o matemático (Verreault-Julien 2019), lo cual en las RNA suele ser una tarea difícil y complicada si solo se aborda su construcción y aplicación, como está reportado en documentos técnicos y libros de texto dedicados a la enseñanza de estas[3]. Incluso en ámbitos no afines a las disciplinas relacionadas con la construcción de modelos de tal clase es común hacer alusión al termino "caja negra" para referirse a la lógica de entrada-salida que permite condiciones y niveles de ingreso de uno o múltiples factores para emitir respuestas sin revelar su funcionamiento interno (Nathan 2021), como es el caso del *software* (Morin et ál. 2012).

Este trabajo presenta argumentación sobre los conceptos de *explicabilidad*, *interpretabilidad* y *transparencia* como ejes precursores de la explicación científica a partir de los modelos RNA del entorno del AA. Se abordarán los conceptos base de diferentes enfoques de explicación científica, siendo los modelos de cobertura legal de Hempel, el concepto de explicación de Van Fraassen (1980) y el modelo de causalidad de Pearl (2009) los mencionados para contextualizar la discusión de posibilidad de explicación científica a partir de modelos RNA. Con base en esto, el objetivo principal se centra en hacer una contribución a la educación en las ciencias y al sector de la tecnología respecto a lo que sucede en el

[3] Estas serían dos etapas de rutina muy generales en la práctica y uso de dichos modelos, i. e., cosecha y ordenamiento de los datos de un fenómeno para el entrenamiento, la validación y la prueba del modelo en cuestión —o, más precisamente, su construcción—, mientras que la aplicación se refiere al despliegue o puesta en producción de estos modelos en alguna plataforma de tecnología. Desde luego, como se verá, cada una de ellas entraña una serie de actividades y subetapas que lleva a cabo el ingeniero o investigador que las conduce.

ámbito del entorno de modelado de problemas y fenómenos por medio de modelos basados en RNA, ya que actualmente es común que dichos constructos dominen el funcionamiento de plataformas tecnológicas que son usadas en la cotidianidad, las cuales día a día son aún más sofisticadas dado su continuo desarrollo, pero casualmente no son comprendidas a un nivel explicativo y por ello se generan problemáticas como la concepción de la caja negra. Así, lo que se pretende es que lo contenido aquí sirva para dar a conocer las bases de lo que significan dichos modelos, sus objetivos y, dado el actual avance tecnológico, la posible semejanza que ya pueden tener con otros constructos lógicos más parsimoniosos y transparentes, como los modelos analíticos de las ciencias, en cuanto a su capacidad para explicar un fenómeno.

Para poder abordar con suficiencia los conceptos que aquí se mencionan, se hace una revisión a las bases fundamentales de los modelos en las ciencias; se recurre a la definición de modelo científico, y se enfatiza en las concepciones nómicas de la explicación en su grado más general y en términos del *modelo de cobertura hempeliano de explicación científica*. Además, se señalan los esfuerzos que en la actualidad se hacen para ofrecer modelos explicativos de las RNA desde una perspectiva de la heurística computacional. Finalmente, se dedica una sección a la teoría esencial que describe los elementos matemáticos y la semántica de las RNA con el fin de presentar las partes fundamentales con las que estas están construidas.

2. Modelos en la ciencia y la tecnología

La definición de los modelos científicos suele asociarse con réplicas de un objeto o con entidades de clase figurativa o formal de un fenómeno, y se construyen para entender parcelas de la realidad y del universo natural (Díaz 2005). En la actividad científica, estas vienen a ser representaciones abstractas e idealizadas

que consideran un rango de condiciones particulares para representar un fenómeno o un objeto (Frigg & Hartmann 2020). Así, en el ámbito de las ciencias, y como bien se apunta en el trabajo de Acevedo-Díaz et ál., "[…] el carácter ambiguo o polisémico de la idea de modelo, presente incluso dentro de la actividad científica en las diferentes disciplinas, […] induce una comprensión incompleta, cuando no errónea del significado de modelo científico" (2017 157). Por ello es claro e importante destacar la definición del concepto de modelo científico para el caso de estudio de las RNA, en lo posible encerrando la significación de su relación con las actividades científicas.

La acepción de modelo científico como una representación de un objeto, comportamiento o fragmento de la realidad que se busca entender y explicar es acertada (Giere 2004)[4]. Sin embargo, se puede complementar agregando el término de interpretación, puesto que es coherente con el contexto del entendimiento de la naturaleza; es decir, según Bailer-Jones (2009), un modelo se refiere a una descripción interpretativa de un fenómeno que facilita el acceso a este. También, y en términos de sus objetivos, un modelo, según como lo describe el filósofo Carl G. Hempel (2005), sirve para la predicción, retrodicción y explicación de las regularidades del flujo de acontecimientos en la naturaleza. Bajo estos términos y acepciones, entonces, un modelo para el caso que ocupa en las ciencias es una representación codificada y abstraída de la realidad, de tal suerte que, una vez construidos, estos tienden a ser económicos en su aplicación por cuanto pueden ser explotados para diversos fines, como la predicción de un fenómeno natural o social, pero esencialmente apoyan para la consecución de una meta (por ejemplo, predecir el producto interno bruto de un país). No obstante, es importante señalar que, aunque los modelos permiten estimar estados de la regularidad de aquello que

[4] Díaz en el 2005, por ejemplo, argumenta que la existencia de un modelo en las ciencias se da solo cuando hay una representación, en la que además se espera similitud con aquello que representa; esto es, "el modelo pretende constituirse en una analogía lo más cercana posible a una homología con las cosas, procesos o sistemas que interesan al estudioso, pero sin llegar a conseguirla" (Díaz 2005 13).

representan de la realidad, estos también cumplen la función de explicar cómo esa uniformidad da lugar a los hechos y las consecuencias que se desprenden de algunas condiciones observadas (Ladyman 2001).

Sobre la predicción y la retrodicción, estas corresponden con las actividades de estimar los estados futuros o pasados que presenta un fragmento de la realidad dentro de la regularidad que le supone. Así, por ejemplo, los modelos de la mecánica newtoniana nos permiten hacer inferencias de las posiciones de una roca después de ser arrojada desde lo alto de un acantilado a partir de su masa, sus condiciones iniciales y las condiciones de frontera del movimiento de esta (velocidades, aceleración, posición, región del movimiento, fuerzas que se le oponen). En el caso de las RNA, la retrodicción y predicción también se ciernen sobre los niveles de los factores o características de fenómeno.

El concepto de modelo en la tecnología se extiende a partir de las consideraciones previamente descritas desde la perspectiva en ciencia; es decir, los modelos en la tecnología también son constructos que formalizan la relación entre las condiciones y las respuestas de un fenómeno en particular, y, como tales, sirven para predecir y estimar alguna respuesta en concreto de este. La diferencia radica en que estos últimos se adhieren a fines que pueden ir más allá de la explicación de un fenómeno, como la predicción, el control o la intervención sobre este, y en el caso de la tecnología, dichos fines pueden incluir la generación de rentabilidad en una compañía. Además, aunque la taxonomía de modelos científicos y tecnológicos puede ser amplia, este artículo limita solo la mención y discusión de las RNA como "modelos computacionales" del tipo formal (Díaz 2005). Una descripción más profunda puede ser encontrada en el trabajo de Frigg y Hartmann (2020).

3. Redes neuronales artificiales

Las redes neuronales artificiales son modelos computacionales del tipo formal (Díaz 2005). Estos se fundamentan en el trabajo de McCulloch y Pitts (1943) relacionado con el estudio del modelado lógico del comportamiento de una neurona biológica en sus funciones de recibir, procesar y emitir información. En sí, es justo decir que una sola de estas unidades no es capaz de generar inteligencia propia para modelar un fenómeno (Kandel et ál. 2013)[5], de ahí que sea necesario contar con una red interconectada de estas para lograr el aprendizaje necesario y así representar datos como modelos científicos y tecnológicos (LeCun, Bengio & Hinton 2015). De hecho, las RNA deben su nombre a que son similares a los sistemas nerviosos de los seres vivos, en los que una red rica en neuronas conectadas entre sí da lugar a comportamientos y conductas complejas. En este contexto, la concepción de neurona artificial sería retomada con mayor énfasis y mejores resultados por Rosenblatt y otros investigadores durante las décadas de 1940 y 1950, como lo recopila Bishop (2015) en su revisión histórica. Rosenblatt llevó a cabo las primeras simulaciones de su "perceptrón" en la década de 1950, definido como un instrumento de procesamiento y aprendizaje de datos basado en entradas/salidas, y presentó un trabajo al respecto en 1958 (Rosenblatt 419-449). Posteriormente, en 1962, apareció su libro más importante sobre el tema, *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms* (Rosenblatt 1962). Consecuentemente, la forma moderna de las RNA ocurrió al incorporar el "algoritmo de retropropagación" de Rumelhart, Hinton y Williams (1986), y con dicho avance el perceptrón de Rosenblatt tendría la capacidad de

[5] McCulloch y Pitts (1943), en su estudio, hacen una introducción y desarrollo de la lógica computacional que describe por primera vez las neuronas artificiales, pero no establecen términos que definan las redes neuronales artificiales.

representar información contenida en un conjunto de datos al incorporar elementos que relacionan entradas y salidas[6].

De McCulloch, Pitts, Rosenblatt y Rumelhart se aduce que las RNA son modelos computacionales, ya que aprenden a representar aquello que se contiene en conjuntos de datos de información y sus propiedades son inherentes de la informática (los algoritmos que las constituyen se hacen posibles a través de procesamiento computacional). Aunque los trabajos de estos últimos autores mencionados refieren a los fundamentos y bases que dieron lugar a las RNA modernas, desde los años setenta se han visto diversas etapas de desarrollo, madurez y estacionalidad en cuanto a la ciencia que entraña el estudio de estas (LeCun, Bengio & Hinton 2015). En este artículo, la discusión y argumentación se centran en los modelos de RNA con "aprendizaje supervisado" a través del método de entrenamiento por retropropagación (O'Shea & Nash 2015; Schmidt 2019). En aras del estudio y la argumentación que se presenta en las próximas secciones, se prescinde de modelos RNA más sofisticados como los mencionados en los trabajos de Goodfellow, Bengio y Courville (2016) y Aggarwal (2018).

3.1. Breve descripción matemática de las RNA

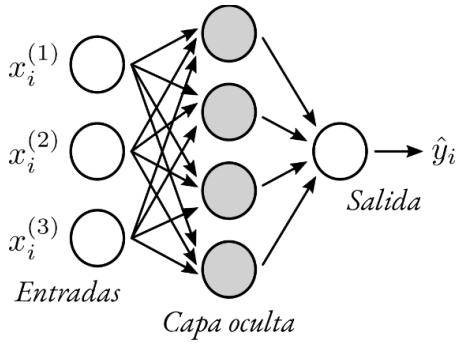
Las RNA no son solamente instrumentos y modelos que producen salidas a partir de entradas[7]. Como modelos formales, sus elementos se describen por medio de expresiones matemáticas y lógicas. Además, estas poseen unidades básicas llamadas "neuronas artificiales", cada una de las cuales recibe y emite información si se alcanza un umbral definido por una función de activación. Precisamente, el

[6] Dentro del contexto del AA y las RNA, no es para nada trivial que los textos de teoría especializados empleen la palabra inglesa *respresentation*, puesto que una RNA se asume propiamente como un modelo de aquella información que representa

[7] Es particularmente importante resaltar que el entendimiento de estas en esos términos es el detonante de problemáticas del tipo caja negra en las RNA.

término red proviene de la conexión de varias neuronas, ya que por sí sola una no es capaz de capturar y representar las complejas relaciones y patrones presentes en los fenómenos que se busca modelar, de ahí que se requieran redes de estas (Kandel et ál. 2013). Así, el modelado mediante RNA implica la construcción de representaciones abstractas de los fenómenos estudiados a través del aprendizaje de patrones en los datos.

En una RNA de retropropagación, las neuronas artificiales están dispuestas por grupos llamados "capas" (del anglicismo *layers*), donde se alojan neuronas artificiales, y, a su vez, estas se conectan con otras capas a través de coeficientes que representan la intensidad de conexiones entre neuronas, también llamados "pesos". La primera y la última capa se denominan "capa de entrada" y "capa de salida", que, respectivamente, contienen las entradas y salidas de información de datos que pretenden procesar. A su vez, las capas intermedias se denominan "capas ocultas", las cuales son las neuronas en las que sus pesos se ajustan para lograr un error de predicción aceptable en una RNA del tipo de retropropagación (figura 1).



*Nota: los superíndices indican la posición de una neurona y el subíndice i se refiere al *i-ésimo* ejemplar contenido en un conjunto de datos C.

Figura 1. Diagrama de una red neuronal artificial de tres entradas, cuatro neuronas ocultas y una salida.
Fuente: elaboración propia.

La figura 1 muestra un esquema de una RNA con una sola capa oculta del tipo de retropropagación. Se describen las entradas por medio de un *vector de características*:

$$\mathbf{x}_i = \begin{bmatrix} x_i^{(1)} \\ x_i^{(2)} \\ \vdots \\ x_i^{(m)} \end{bmatrix}, \mathbf{x}_i \in \mathbb{R}^m, \tag{1}$$

donde cada superíndice indica una característica del fenómeno que se pretende representar; el subíndice i representa el *i-ésimo* vector de características y $\mathbb{R}$ se refiere al conjunto de los números reales (para el caso de la figura 1, $m = 3$). En este contexto, las características se entienden como las variables independientes sobre las cuales se da una correspondencia de la forma

$$f_{RNA}: \mathbb{R}^m \to \mathbb{R}^n, \tag{2}$$

donde f_{RNA} especifica un mapeo funcional desde $\mathbb{R}^m$ hacia $\mathbb{R}^n$ (para el caso concreto de la figura 1, $n = 1$). Esta correspondencia es precisamente la que un modelo de RNA exitoso representa, y se da a partir del aprendizaje de un conjunto de datos $\mathcal{C}$:

$$\mathcal{C} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_i^N, \mathbf{x}_i \in \mathbb{R}^m; \quad \mathbf{y}_i \in \mathbb{R}^n, \tag{3}$$

donde N es el tamaño del conjunto $\mathcal{C}$ y y_i es la respuesta o variable dependiente que se interpreta como la salida que se pretende modelar:

$$\mathbf{y}_i = \begin{bmatrix} y_i^{(1)} \\ y_i^{(2)} \\ \vdots \\ y_i^{(m)} \end{bmatrix}, \mathbf{y}_i \in \mathbb{R}^n. \tag{4}$$

Además, una RNA se construye cuando las conexiones internas entre cada una de sus neuronas, conocidas como "pesos", se entrenan o calibran para efectivamente reducir el error existente entre sus salidas. Esto se hace a través de un proceso iterativo conocido como "descenso de gradiente", el cual puede ser descrito por la siguiente expresión:

$$w_{l,k}^{(j)} \leftarrow w_{l,k}^{(j)} - \frac{\alpha}{N} \sum_{i=1}^{N} \left(\frac{\partial L}{\partial w_{l,k}^{(j)}} \right)_i, \tag{5}$$

donde $w_{l,k}^{(j)}$ se refiere al peso existente de la neurona k de la capa l con la neurona (j) de la capa previa $(l-1)$, esto es, define la relevancia de si una neurona recibe información para ser procesada. Además, α es una *tasa de aprendizaje* y la derivada parcial del lado derecho es el gradiente del error L para cada conexión:

$$L = \frac{1}{Nn} \sum_{\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}}^{N} \underbrace{\sum_{j=1}^{n} \left(y_i^{(j)} - \hat{y}_i^{(j)} \right)^2}_{C_i}, \tag{6}$$

donde C_i es un error individual para cada *i-ésimo* par en el conjunto de datos, y las predicciones o estimaciones de una RNA se alojan en el siguiente vector:

$$\hat{\mathbf{y}}_i = \begin{bmatrix} \hat{y}_i^{(1)} \\ \hat{y}_i^{(2)} \\ \vdots \\ \hat{y}_i^{(n)} \end{bmatrix}, \hat{\mathbf{y}}_i \in \mathbb{R}^n. \tag{7}$$

Las ecuaciones 1 a 7 describen los elementos de una RNA, por lo que, sucintamente, estas pueden ser circunscritas como un problema de optimización, es decir, el propósito de minimizar L a través de una heurística computacional, tal que

$$\min_{w} L\,(\mathbf{w}); \quad L(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^{N} C_i(\mathbf{w}), \tag{8}$$

donde w se refiere a la forma vectorizada de los pesos. De esta manera, en las RNA, el aprendizaje ocurre cuando los pesos[8] se ajustan para reducir el error entre los datos y las predicciones a través de algoritmos iterativos. Uno de estos algoritmos es el descenso de gradiente, como el que se define por la ecuación 4.[9]

Como tal, los parámetros de una RNA no revelan nada del fenómeno que esta podría representar. Esto genera un problema desde la perspectiva científica pues, tal como se menciona en el estudio de Zednik del 2021, la preocupación se centra en que los parámetros de un modelo no ofrecen elementos para generar interpretación y explicación de los fenómenos que se buscan modelar a partir de un conjunto de datos. Es decir, contrario a un modelo matemático formal, aquí, por ejemplo, una matriz de pesos **w** no contiene necesariamente valor epistémico

[8] No son, sin embargo, los únicos parámetros que se modifican de un modelo; dependiendo de su arquitectura, una RNA podría también ver ajuste, por ejemplo, en otros parámetros *b* llamados "sesgos".

[9] Los algoritmos de actualización de los parámetros de una RNA se disponen en una taxonomía que les distingue como *heurísticos* y *metaheurísticos*; los primeros se basan en descripciones formales, similares a la ecuación 5, mientras que los segundos carecen de esto y recurren a lógica por analogía a algún proceso observado en la naturaleza, como los algoritmos genéticos (Yang 2010).

del fenómeno, aun si la predicción que se genera a partir de ello es muy precisa[10]. Esto adquiere mayor grado de opacidad cuando se dan arquitecturas de RNA del tipo profundo con más capas ocultas o neuronas de las que se muestran en la figura 1, es decir, modelos de RNA del *deep learning* (LeCun, Bengio & Hinton 2015) que requieren de millones de parámetros para su funcionamiento interno, pero que ninguno revela o dice algo sobre el fenómeno que aborda.

4. El problema de la caja negra de los modelos RNA

El término "caja negra", en el caso semántico y práctico del modelado de fenómenos por RNA, implica que estas pueden cumplir con el objetivo de predecir y modelar fenómenos a partir de conjuntos de datos sin revelar lo que supone su

estructura interna para llevarlo a cabo (Zednik 2021). Esto es, las RNA no ofrecen a primera instancia interpretabilidad o transparencia en el proceso de predicciones o estimaciones, dificultando así la actividad de la explicación científica. Arriba se vio que los pesos y otros parámetros de una RNA nada dicen del fenómeno sobre el cual se construyen a partir de sus datos. Esto supone que el funcionamiento interno de estos modelos no permite establecer vínculo con la regularidad del fenómeno, provocando un problema de mayor trascendencia en las actividades dentro de las ciencias.

Según Rudin (2019), un modelo de caja negra bien puede ser una función que es demasiado complicada para ser entendida por un humano o un modelo reservado del cual no se tiene acceso a los mecanismos que le hacen funcionar. Las RNA caben dentro de esta definición a partir de lo que expresado

[10] En este caso, me refiero a aquello que, en un modelo matemático tradicional, como lo es el caso de la masa m en la ecuación $\mathbf{F} = m\mathbf{a}$, está explícito y representa algo de aquello que tiene nombre, relevancia, e influencia del problema o del fenómeno que se estudia.

en la ecuación 2. Asociado a esto, tanto Rudin como otros autores señalan los riesgos de operar bajo estas condiciones, entre los que se destacan la posible toma de decisiones erróneas en instituciones públicas o en firmas especializadas conectadas a la economía (Barredo Arrieta et ál. 2020; Ghassemi et ál. 2021; Slack et ál. 2020). Esto se adhiere al principal problema que entraña la cuestión del término y concepto de caja negra, puesto que los modelos sirven, esencialmente, para tomar decisiones.

Muchos algoritmos de AA han sido nombrados modelos de caja negra debido a su inescrutable funcionamiento interno y la opacidad que existe en ellos (Barredo Arrieta et ál. 2020). En este sentido, dichos modelos no pueden servir para llevar a cabo explicaciones ya que cuando se busca una explicación lo que se pretende es entender por qué $\hat{y}_i$ sucede a partir de ciertas condiciones establecidas en x_i; es decir, y según Van Fraassen (1980), las explicaciones son respuestas a preguntas del tipo *¿por qué?*[11]. Con base en esto, en un modelo que efectiva y positivamente es transparente, y que no se adhiere a la problemática de la caja negra (como lo es el caso de la *ley de los gases ideales*), sus elementos permiten positivamente establecer relaciones del tipo causa-efecto para responder a esta clase de interrogantes; por ejemplo, al aumentar la temperatura de un gas bloqueando la variable de volumen, se encuentra la razón de por qué se aumentaría la presión dentro de este[12]. En esos términos, se podría plantear la idea de que en la actividad científica la legitimidad de un modelo de RNA tendría que ver con dos objetivos esenciales: (1) predecir y (2) explicar aquello que modela; en caso contrario, el problema de la caja negra se hace presente si no se interpreta y se da

[11] Un análisis más profundo de esta concepción la arroja el trabajo de Diéguez (1994), en el que se estudian los elementos que conforman una pregunta del tipo por qué: asunto, la clase de contraste y la relación de relevancia, siendo esta última lo que determina lo que contará como un factor explicativo a dicha pregunta.

[12] La forma más conocida de dicha ley está cifrada en la expresión matemática $PV = \upsilon RT$, donde P es la presión, V el volumen, υ es la cantidad de moles del gas, R la constante universal de los gases y T es la temperatura.

explicación de la relación entre las entradas, las salidas, y la relación causa-efecto de estas en un modelo.

Es importante enfatizar que, aunque a primera instancia las RNA adolecen de la parte explicativa, estas son excepcionales en la parte predictiva, más aún en la exploración y el mapeo de datos de los fenómenos que abordan (con ello es posible generar respuestas sintéticas de condiciones no testadas en las entradas de un modelo). Así, debe tenerse en cuenta el valor que aportan estos modelos, por lo que es esencial recurrir a métodos precursores que permitan a los investigadores llevar a cabo, primero, la tarea de transparentar el funcionamiento interno de un modelo, y, posteriormente, el trabajo de construcción de la explicación en términos científicos. Para abordar este problema es importante entender qué es la explicación en las ciencias, al menos desde un marco de entendimiento que haya sido o sea aceptado en la historia de la ciencia y la filosofía[13].

5. Modelos de explicación en las ciencias

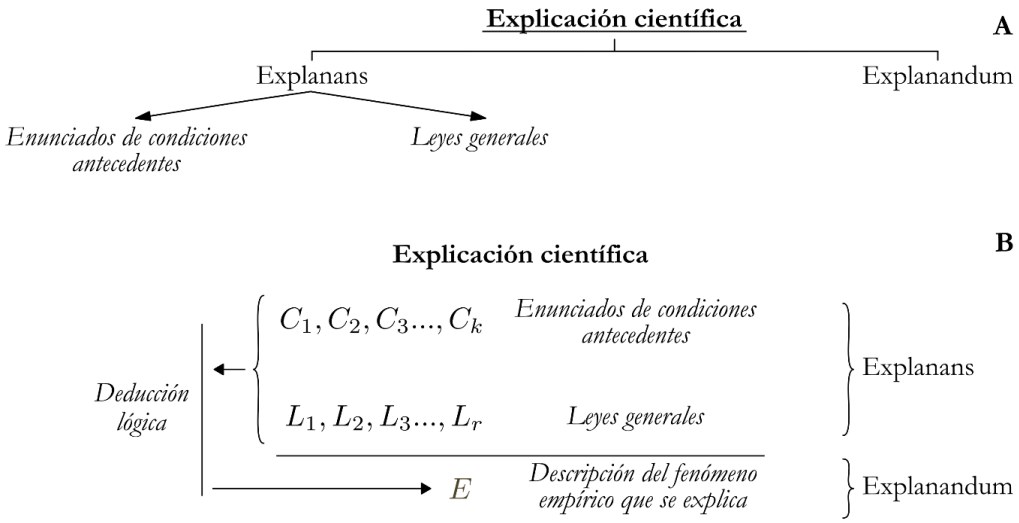
Como se ha planteado en la sección anterior, la explicación en las ciencias es responder a preguntas del tipo "por qué". Aun cuando esa es, grosso modo, la forma sucinta y actualmente más aceptada del concepto, no es exclusiva de los modelos de explicación que en el pasado han servido de soporte a los científicos como marco de referencia para explicar las regularidades de un fenómeno en concreto. En esta sección se presentarán dos de las acepciones que han tenido mayor peso en la historia de la ciencia, lo cual permitirá abordar la discusión central de la existencia de explicación científica para el caso de las RNA.

[13] Aquí debo hacer una aclaración: en la historia de la filosofía de la ciencia, la explicación científica se ha reconocido como una actividad fundamental en las ciencias. Existen múltiples abordajes y modelos mismos de esta clase de explicación. Sin embargo, los marcos que aquí se presentan establecen elementos que son válidos en muchos modelos matemáticos, lo que permite poner referencias para argumentar sobre la problemática de la caja negra en los modelos basados en RNA.

La figura 2(A) presenta un recuadro construido a partir del trabajo de Hempel, el cual refiere a su célebre modelo nomológico-deductivo (Hempel 2005). En dicho recuadro se pueden identificar dos partes esenciales del modelo de explicación en cuestión: (1) enunciados explanantes (*explanans*) y (2) enunciado *explanandum*. Estos dos elementos se describen a continuación:

1. El *explanandum* es la oración que describe el fenómeno a explicar (y no al fenómeno mismo).
2. El *explanans* son las clases de aquellas oraciones que se aducen para dilucidar el fenómeno.

Según Hempel (2005), las explicaciones son argumentos de dos tipos: (1) aquellos que indican las condiciones existentes y que se manifiestan antes de la explicación de un fenómeno, y (2) los que expresan o se subsumen en ciertas leyes generales del fenómeno. Además, según Hempel, la explicación en las ciencias debe cumplir dos requisitos esenciales para adquirir su estatus (2005): (i) tiene que ser relevante en términos explicativos (requisito de relevancia explicativa), y (ii) debe ser empíricamente contrastable (esto es, debe cumplir con el requisito de contrastabilidad).



(A) Descripción de sus elementos. (B) Deducción lógica de la explicación
Figura 2: Modelo de cobertura de la explicación científica de Hempel
Fuente: Hempel 2005.

Muchas explicaciones de fenómenos físicos caben en el modelo nomológico-deductivo, ya que es por medio de modelos matemáticos construidos a través de leyes generales físicas que es posible explicar el fenómeno que estos abordan. Esto se observa en la figura 2(B), en la que se señala que los *explanans* son las condiciones de las características de un fenómeno, las cuales se cifran y se relacionan entre sí a través de la instancia de un modelo, que es construido a partir de una o varias leyes generales. Sobre esto, tómese, por ejemplo, el caso particular de un cilindro de gas expuesto a altas temperaturas; aquí la ley y las condiciones, como *explanans*, permitirían explicar el comportamiento del incremento en su presión interna (*explanandum*).

Por otra parte, aunque pareciese tener una cobertura amplia, el mismo modelo hempeliano de explicación encuentra límites, como lo describe su creador, pues en su trabajo (Hempel 2005) se menciona que muchos fenómenos no solo se subsumen en una primera regularidad definida sobre alguna ley física, sino que también lo hacen en regularidades propias de la estocasticidad. En este sentido, el modelo inductivo-estadístico de Salmon (1970) surge como una alternativa al modelo de Hempel, al considerar que las explicaciones científicas pueden basarse en leyes estadísticas y no solo en leyes deterministas. Según Salmon, una explicación estadística es adecuada si el fenómeno a explicar se puede subsumir bajo una ley estadística de alto nivel. Es importante mencionar esto ya que toda RNA es construida a partir de datos que en muchas ocasiones siguen una tendencia general, al mismo tiempo que presentan una marcada variabilidad en su contenido, lo que las hace más compatibles con un enfoque estadístico de la explicación científica.

Existe otro abordaje más reciente de la explicación en las ciencias que viene soportado por los *modelos causales de explicación científica*, los cuales fueron desarrollados a finales del siglo XX e inicios del siglo XXI en pleno auge de los modelos de la inferencia estadística y la computación. Estos modelos son marcos que se usan para representar las relaciones entre las características de un sistema y las causas y los efectos de los cambios que estas sufren cuando hay un cambio en ellas. Como tal, la idea principal de dichos modelos es comprender la estructura de causas en un sistema para hacer predicciones, entender los fenómenos que se estudian y así poder dar explicarlos. El marco de modelos de causalidad más relevante es el "modelo de causalidad estructural" propuesto por Pearl y Mackenzie (2018). Este proporciona un conjunto de reglas y algoritmos para inferir la estructura de causas en un sistema. Según los autores, la explicación surge cuando los seres humanos hacen razonamiento contrafactual, el cual genera escenarios del tipo "¿qué podría ocurrir o habría ocurrido si…". En ellos se puede entender las posibles consecuencias de haber seguido o considerado diferentes

niveles en las características de un sistema, aun si no se cuenta con datos de este (de ahí que sean contrafácticos)[14].

Aunque el modelo nomológico-deductivo y los modelos de causalidad no son a primera instancia compatibles, sí es importante señalar que ambos, incluso en épocas distintas, han gozado de aceptación por su acometido para el desarrollo de la explicación de un fenómeno. Así, aunque de valor epistémico distinto, estos enfoques ponen en relieve dos situaciones clave para el caso de la explicación: (1) para ejercer una explicación del tipo científico se requieren leyes que cubran el fenómeno; (2) se necesitan llevar a cabo razonamientos del tipo contrafactual para establecer las hipótesis de la causa de un fenómeno. Ambas situaciones parecieran no ser posibles cuando se tiene un modelo RNA, puesto que el mero hecho de que permita predecir con excelente precisión una manifestación natural no significa que a partir de este se puedan plantear contrafactuales, ni que sus elementos básicos pertenecen a un conjunto dado de datos que modelan para construir un *explanandum*; esto es, las neuronas artificiales y algoritmos de entrenamiento no pertenecen a una ley que lo abarca todo (tómese por ejemplo el caso de si se modelara una instancia cubierta por la *ley de los gases ideales*). En esencia, esto es lo que no establece un vínculo posible para que de una RNA poder llevar a cabo el proceso de la explicación en un primer intento. De ahí que en tiempos recientes se han depositado esfuerzos desde diferentes marcos y frentes para establecer métodos y herramientas como precursores de la interpretación de un modelo opaco como lo son las RNA.

[14] En el caso del cilindro de gas, es interesante preguntarse qué habría pasado con la temperatura si esta no se hubiera incrementado.

6. Explicabilidad, interpretabilidad y transparencia en el AA

Los filósofos llevan siglos discutiendo y ampliando el concepto de explicación en las ciencias (Woodward & Ross 2021), pero en el campo de los datos y el AA recién se comienzan a desarrollar términos sui generis que dan fundamento para tratar de explicar modelos como las RNA (Linardatos, Papastefanopoulos & Kotsiantis 2021). Como se ha visto, un modelo de este tipo es ciertamente opaco, y para poder aspirar a llevar a cabo ejercicios contrafactuales o de desarrollo de *explanandum* de un fenómeno, es necesario primero entender qué mecanismos de su interior hacen que sea funcional.

Existen tres conceptos clave dentro del AA que aluden a explicación, pero que no lo son si se limita este término al caso de explicar el fenómeno que aborda un modelo:

- *Interpretabilidad*. Los científicos de datos e ingenieros del AA le llaman "interpretabilidad" a la capacidad de explicar o presentar en términos comprensibles la información interna de un modelo a un humano (Doshi-Velez & Kim 2017). Según Vilone y Longo (2021), la interpretabilidad es el mapeo de un concepto abstracto (por ejemplo, una clase estimada) en un dominio al que el humano puede dar sentido. Un modelo interpretable podría ser un modelo de árbol de decisiones simple, del cual puede entenderse su funcionamiento tan solo con observar su estructura. De esta manera, se identifica que la interpretabilidad corresponde a (1) un examen del propio modelo o (2), alternativamente, a un examen de la respuesta que presenta un modelo a estímulos planificados que se llevan a cabo al considerar diferentes niveles en sus entradas (incluso de naturaleza estocástica).

- *Explicabilidad*. Se refiere a una colección de artefactos visuales o interactivos, como los gráficos, que proporcionan al usuario de un modelo de IA una descripción suficiente de su comportamiento; en los términos expuestos por Saleem et ál. (2022), la explicabilidad se trata de explicar el proceso interno de un modelo para responder a la pregunta "¿Cómo toma determinadas decisiones el modelo de caja negra de la IA?".
- *Transparencia*. Según Roscher et ál. (2020), se dice que un enfoque de AA se hace transparente si los procesos que extraen los parámetros del modelo a partir de los datos de entrenamiento y generan etiquetas a partir de los datos de prueba pueden ser descritos y motivados por el diseñador del enfoque. Como tal, la transparencia se refiere a la arquitectura general de una RNA (o, en general, de un modelo de AA), sus componentes individuales del modelo, el algoritmo de entrenamiento que se usó para calibrarla y cómo se obtienen la soluciones mediante su uso (Sokol et ál. 2022).

Es importante exponer que los tres conceptos de explicabilidad tienen una connotación y trasfondo eminentemente informáticos. Esto es, a excepción del uso práctico que ofrece la interpretabilidad, estos conceptos no tienen un significado más allá de la explotación y consecuente aplicación de un modelo RNA que se da sobre una plataforma computacional. Más aún, en la ausencia de ello pierden sentido, pues estos se constituyen solo cuando existe y se da un modelo que ocupa espacio tangible en memoria electrónica. Por ello es importante aclarar que estos conceptos más bien se refieren a metamodelos[15] de una RNA —

[15] Según Van Gigch (1991), un metamodelo es un modelo que incorpora los elementos y las condiciones de otro modelo cuando este último posee características que lo hacen de difícil acceso; es decir, un metamodelo es un modelo sustituto de un modelo. Según este autor, los metamodelos se utilizan cuando los resultados de un modelo son demasiado complejos e intrincados, demasiado costosos de producir o no es factible trabajar con ellos. En el caso de los metamodelos de explicación, estos son modelos sustitutos que se usan para interpretar un modelo intrincado, como una RNA.

o de alguna otra clase de modelo informático del AA —[16]. Estos tres conceptos, además, se alejan de toda formalidad como la que ofrece Hempel en su modelo nomológico-deductivo. Esto no es un rasgo mismo de la disciplina en la que ven su desarrollo, más bien es una consecuencia del avance en construcción de nuevas tecnologías; es decir, son modelos informáticos que explican los modelos de IA, como las RNA. De esta manera, un metamodelo de explicabilidad es un modelo que explica el interior de un modelo opaco, no un metamodelo que explica la regularidad de un fenómeno.

Los términos de interpretabilidad, explicabilidad y transparencia tampoco se adhieren a los modelos de causalidad puesto que al ejecutarlos para transparentar y entender solo el funcionamiento de un modelo en concreto estos métodos no son para explicar aquello que abarca el modelo; en términos de Pearl y Mackenzie (2018 6), los "datos son mudos"[17] y, dado que un modelo RNA se construye a partir de estos, no se podría esperar que los métodos relativos a la explicabilidad sean sinónimo de la explicación científica de la regularidad que describen tales datos. Lo que se entiende como explicabilidad, en el contexto informático, consiste, entonces, en un modelo mismo de los parámetros internos de una RNA a suerte de traducción de lo que ocurre en su interior. Así, las acepciones de interpretabilidad, explicación y transparencia en la ciencia de datos están ligadas al análisis de la relación entradas/salidas de un modelo, generalmente mediante técnicas como las librerías LIME y los modelos agnósticos (Ribeiro, Singh & Guestrin 2016); los modelos de SHAP (Lundberg & Lee 2017); las gráficas de efecto local acumulado ALE (Apley & Zhu 2020); la importancia de características (Guyon & Elisseeff 2003), o técnicas de graficado por dependencia

[16] Al constituirse como herramientas informáticas, tanto los metamodelos de explicabilidad como los modelos de interpretabilidad tienen lugar toda vez que primero se ajusta y entrena un modelo de AA, en este caso una RNA, a través de sus predicciones. En el caso de la transparencia, esta se constituye si y solo si existe un modelo ya construido y validado.

[17] Traducido por el autor directamente de la frase "data is profoundly dumb", expuesta en el trabajo *The Book of Why* (Pearl & Mackenzie 2018 6).

parcial (Greenwell 2017). Estas herramientas son técnicas y modelos de modelos (metamodelos), de ahí su relevancia en cuestiones prácticas.

7. Posibilidad de explicación a partir de un modelo RNA

Como se ha visto en secciones precedentes, una RNA se constituye a partir de conjuntos de datos, por lo que, con precisión, estos modelos también son modelos de datos[18], esto es, representan un conjunto de datos que se obtiene a partir de la observación y la experimentación. Sin embargo, las RNA no son interpretables a primera instancia y se requieren metamodelos de explicabilidad que desentrañen su funcionamiento interno. Además, estos modelos tienen un valor práctico notable, de ahí su auge, pues la utilidad que proveen es para predecir y estimar datos con una elevada exactitud. Sin embargo, a la luz de lo planteado en las secciones anteriores, no se puede explicar un fenómeno a través de las RNA si solo se las tiene en cuenta, como sí ocurre en el caso de otros modelos en las ciencias (por ejemplo, en los modelos matemáticos).

El término de explicación en el contexto de explicabilidad a través de metamodelos, entonces, no hace una referencia directamente a explicar el fenómeno que una RNA ajusta para las tareas y los propósitos ya mencionados. El término de explicación, en el contexto de las ciencias, aborda algo más amplio y trascendente que el solo hecho de desentrañar el interior de un modelo, esto es, entendimiento científico, y es lo que plantean Krenn et ál. (2022). Por esto, también es justo por extensión decir que los metamodelos de explicación no hacen

[18] Aquí es importante destacar que me refiero a la definición tal como se menciona en el trabajo de Frigg y Hartmann (2020), y que refiere al concepto de Suppes (1962), según el cual un modelo de datos es una versión corregida, rectificada, reglamentada y en la mayoría de los casos idealizada de los datos que se obtienen a partir de la observación inmediata, es decir de los llamados datos brutos.

ni generan por sí mismos contrafactuales para poder ofrecer la explicación del fenómeno que aborda un modelo RNA. Como tal, la explicabilidad y la interpretabilidad se pueden entender como "métodos precursores" que sirven de soporte para una posterior explicación científica si se habla de su uso en redes neuronales. Sin embargo, tanto las RNA como estos metamodelos pueden no ser suficientes por sí solos para alcanzar un entendimiento científico; la explicabilidad solo es la capacidad de describir la relación entre la entrada y la salida de un modelo de forma comprensible para un ser humano, y la interpretabilidad se refiere a la capacidad de comprender su funcionamiento interno. Por esto, el mero hecho de poder explicar o interpretar el comportamiento interno de una RNA, incluso de manera no formal (ausencia de descripción matemática), no garantiza una explicación científica, ya que los propios modelos no se basan en los principios y las relaciones funcionales que tratan de aprender o representar, y estos pueden no coincidir con conocimientos científicos previos. Un reducto excesivo es el siguiente: incluso esos modelos de explicabilidad requieren ser interpretados por alguien. Como siempre lo ha sido, es tarea de los científicos interpretar los constructos que se erigen a partir de observaciones y luego explicar a partir de estos. Nada de eso ha cambiado en la actualidad.

Como se plantea en un trabajo reciente de Krenn et ál. (2022), y siguiendo las ideas planteadas por De Regt y Dieks (2005), para que un modelo de IA pudiese generar entendimiento científico y, por extensión, una explicación de tal categoría, este tendría que cumplir dos condiciones específicas:

1. Reconocer cualidades y características de una teoría sin realizar cálculos exactos, así como utilizarlas en un nuevo contexto.
2. Adquirir conocimientos científicos y transferirlos a un humano.

Según dichos autores, esto quiere decir que solo se está en condiciones de comprender un fenómeno cuando existe tal teoría inteligible de este que los cien-

tíficos reconocen cualitativamente consecuencias y características de esta sin necesariamente profundizar en cálculos exactos (como ejemplo téngase en cuenta el caso de la *ley de los gases ideales* ya descrita arriba, en la que no hubo necesidad de generar nuevos datos a partir de experimentación para llegar a la conclusión mencionada). Con base en lo anterior, una explicación científica está ligada al entendimiento científico, y para que esto ocurra debe existir una teoría (o los cimientos de esta)[19], cosa que puede estar ausente en un modelo de datos, como ya se ha descrito en las secciones precedentes. Cuando eso sucede, ningún método automático computacional podría brindar una plena explicación científica sin un usuario que lo explote y lo interprete, ni sin una teoría. Es decir, es tarea del usuario de estos métodos generar una explicación de la regularidad que en principio pretende modelar con un modelo basado en redes neuronales artificiales, y que es precisamente el objetivo que se persigue tras recurrir a los métodos precursores de la explicabilidad en la mayoría de los modelos del AA.

8. Reflexiones finales

La explicación en la ciencia de datos y el AA por medio de técnicas y modelos computacionales está en pleno desarrollo, lo cual inexorablemente incluye a las RNA. Aunque en la actualidad hay reportes que dan evidencia de la posibilidad de simbolización matemática para interpretabilidad de modelos RNA mediante otras técnicas de metamodelos (ejemplos muy concretos de esto pueden ser consultados en los trabajos de Abdusalamov et ál. 2023, Alaa & Van der Schaar 2019, Vedantam et ál. 2019), lo cierto es que no hay una regla, una técnica o un método general aplicable a todos los casos (estos avances son atomizados, por

[19] Aquí me refiero a modelos incipientes que no abarcan múltiples condiciones, también a la taxonomía y clases que definen a un problema, al igual que reportes de aquello sobre lo que se espera se logre la madurez para construir una teoría.

área específica y los métodos que reportan no son generalizables a otros casos todavía). Esta situación plantea la cuestión de si es posible o no alcanzar una regla o método general en ese campo, y si este desafío es exclusivo de la ciencia de datos y el AA o si se presenta también en otros ámbitos científicos. Si bien dicha pregunta merece un abordaje más profundo que excede el alcance del presente trabajo, es importante tener en cuenta su relevancia al considerar los desafíos y las limitaciones actuales de la explicación en modelos computacionales. A pesar de considerarse formales, no se debe perder la noción de que un modelo computacional de datos conlleva opacidad y requiere de un esfuerzo extra de sus usuarios para poder establecer ese vínculo que solo la explicación científica permite en la comprensión de un fenómeno.

Por otra parte, para efectivamente considerar el estatus de explicación científica a partir de modelos RNA, se requiere de un conjunto de técnicas y métodos precursores, una teoría del fenómeno que se aborda (o los cimientos de ella) y un entendimiento científico de este. Dichos elementos son fundamentales para alcanzar tal condición en las ciencias, como se ha discutido en los estudios citados a lo largo de las secciones precedentes.

Las redes neuronales artificiales y los modelos más complejos que se construyen con estas, aunque presentan problemas para una explicación científica plena, sí son prácticos, pues, como se plantea en Krenn et ál. (2022 763), se aducen como "microscopios computacionales", los cuales "permiten investigar objetos o procesos que no pueden visualizarse o sondearse de ninguna otra manera, por ejemplo, procesos biológicos, químicos o físicos que ocurren a escalas de longitud y tiempo no accesibles en los experimentos". Este valor se reconoce en la actualidad, y ha llevado al sector tecnológico a adoptarlas para resolver problemas de clase multifactorial, donde generalmente es difícil encontrar un modelo tradicional para resolver problemas (Ahmed, Wahed & Thompson 2023), o en el caso de las ciencias, para explorar soluciones de la respuesta de un sistema natural y, por subsiguiente, tratar de explicarlo como actividad humana.

Referencias

Abdusalamov, Rasul et ál. "Automatic Generation of Interpretable Hyperelastic Material Models by Symbolic Regression". *International Journal for Numerical Methods in Engineering* (2023): 1-12.

Abhishek, Kumar et ál. "Weather Forecasting Model Using Artificial Neural Network". *Procedia Technology* 4 (2012): 311-318. <https://doi.org/10.1016/j.protcy.2012.05.047>

Acevedo-Díaz, José Antonio et ál. "Modelos científicos: significado y papel en la práctica científica". *Revista Científica* 30.3 (2017): 155-166. <https://doi.org/10.14483/23448350.12288>

Aggarwal, Charu C. *Neural Networks and Deep Learning*. Cham: Springer, 2018. <https://doi.org/10.1007/978-3-319-94463-0>

Ahmed, Nur, Wahed, Muntasir y Thompson, Neil C. "The growing influence of industry in AI research". *Science* 379 (2023): 884-886. <https://doi.org/10.1126/science.ade2420>

Alaa, Ahmed M. y Van der Schaar, Mihaela. "Demystifying black-box models with symbolic metamodels". *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY: Curran Associates Inc., 2019. 1-11.

Alzubaidi, Laith et ál. "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions". *Journal of Big Data* 8.1 (2021): 53.

Apley, Daniel W. y Zhu, Jingyu. "Visualizing the effects of predictor variables in black box supervised learning models". *Journal of the Royal Statistical Society Series B: Statistical Methodology* 1 (2020): 1059-1086.

Bailer-Jones, Daniela M. *Scientific Models in Philosophy of Science*. Pittsburgh: University of Pittsburgh Press, 2009.

Barredo Arrieta, Alejandro et ál. "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". *Information Fusion* 58 (2020): 82-115. <https://doi.org/10.1016/j.inffus.2019.12.012>

Bishop, J. Mark. "History and philosophy of neural networks". *Computational Intelligence, Volume I*. Hisao Ishibuchi (ed.). París: EOLSS Publications, 2015.

Chen, Jie y Liu, Yongming. "Fatigue modeling using neural networks: A comprehensive review". *Fatigue & Fracture of Engineering Materials & Structures* 45.4 (2022): 945-979.

Cichy, Radoslaw M. y Kaiser, Daniel. "Deep neural networks as scientific models". *Trends in Cognitive Sciences* 23.4 (2019): 305-317. <https://doi.org/10.1016/j.tics.2019.01.009>

De Regt, Henk y Dieks, Dennis W. "A contextual approach to scientific understanding". *Synthese* 144.1 (2005): 137-170. <https://doi.org/10.1007/s11229-005-5000-4>

Dhall, Devanshi, Kaur, Ravinder y Juneja, Mamta. "Machine learning: A review of the algorithms and its applications". *Intelligent Computing and Applications*, Pradeep Kumar Singh et ál. (eds.). Cham: Springer International Publishing, 2020. 47-63.

Díaz, José Luis. "Modelo científico: conceptos y usos". *El modelo en la ciencia y la cultura*. Alfredo López-Austin (ed.). Ciudad de México: Siglo XXI Editores/Universidad Nacional Autónoma de México, 2005. 11-28.

Diéguez, Antonio J. "Explicando la explicación". *Daimon Revista Internacional de Filosofia* 8 (1994): 83-108. <https://doi.org/10.58680/ej19947514>

Doshi-Velez, Finale y Kim, Been. "Towards a rigorous science of interpretable machine learning". *arXiv: Machine Learning* (2017): 1-13.

Emmert-Streib, Frank et ál. "An introductory review of deep learning for prediction models with big data". *Frontiers in Artificial Intelligence* 3 (2020): 4.

Faller, William E. y Schreck, Scott J. "Neural networks: Applications and opportunities in aeronautics". *Progress in Aerospace Sciences* 32.5 (1996): 433-456. <https://doi.org/10.1016/0376-0421(95)00011-9>

Frigg, Roman y Hartmann, Stephan. "Models in Science". *The Stanford Encyclopedia of Philosophy*. Edward N. Zalta. (ed.). Spring/Metaphysics Research Lab, Stanford University, 2020. *Online*. <https://plato.stanford.edu/archives/spr2020/entries/models-science/>

Ghassemi, Marzyeh et ál. "The false hope of current approaches to explainable artificial intelligence in health care". *The Lancet Digital Health* 3.11 (2021): 745-750. <https://doi.org/10.1016/S2589-7500(21)00208-9>

Giere, Ronald N. "How models are used to represent reality". *Philosophy of Science* 71.5 (2004): 742-752. <https://doi.org/10.1086/425063>

Goodfellow, Ian, Bengio, Yoshua y Courville, Aaron. *Deep Learning*. Cambridge: The MIT Press, 2016.

Greenwell, Brandon M. "pdp: An R package for constructing partial dependence plots". *The R Journal* 9 (2017): 421-436.

Guyon, Isabelle y Elisseeff, André. "An introduction to variable and feature selection". *Journal of Machine Learning Research* 3 (2003): 1157-1182.

Hempel, Carl G. *La explicación científica: estudios sobre la filosofía de la ciencia*. Barcelona: Paidós Surcos, 2005.

Janiesch, Christian, Zschech, Patrick y Heinrich, Kai. "Machine learning and deep learning". *Electronic Markets* 31.3 (2021): 685-695. <https://doi.org/10.1007/s12525-021-00475-2>

Kandel, Eric R. et ál. *Principles of Neural Science*. 5ª ed. Nueva York: McGraw-Hill Education, 2013.

Krenn, Mario et ál. "On scientific understanding with artificial intelligence". *Nature Reviews Physics* 4.12 (2022): 761-769. <https://doi.org/10.1038/s42254-022-00518-3>

Ladyman, James. *Understanding Philosophy of Science*. Londres: Routledge, 2001.

LeCun, Yann, Bengio, Yoshua y Hinton, Geoffrey. "Deep learning". *Nature* 521.7553 (2015): 436-444. <https://doi.org/10.1038/nature14539>

Linardatos, Pantelis, Papastefanopoulos, Vasilis y Kotsiantis, Sotiris. "Explainable AI: A review of machine learning interpretability methods". *Entropy* 23.1 (2021): 18. <https://doi.org/10.3390/e23010018>

Lundberg, Scott M y Lee, Su-In. "A unified approach to interpreting model predictions". *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Red Hook, NY: Curran Associates Inc., 2017. 4768-4777.

McCulloch, Warren S y Pitts, Walter. "A logical calculus of the ideas immanent in nervous activity". *The Bulletin of Mathematical Biophysics* 5.4 (1943): 115-133. <https://doi.org/10.1007/BF02478259>

Morin, Alexander et ál. "Shining light into black boxes". *Science* 336.6078 (2012): 159-160. <https://doi.org/10.1126/science.1218263>

Nathan, Marco J. *Black Boxes: How Science Turns Ignorance into Knowledge*. New York: Oxford University Press, 2021.

O'Shea, Keiron y Nash, Ryan. "An introduction to convolutional neural networks". *arXiv abs/1511.08458* 1 (2015): 1-12.

Pearl, Judea. *Causality: Models, Reasoning and Inference*. 2ª ed. Nueva York: Cambridge University Press, 2009.

Pearl, Judea y Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Nueva York: Basic Books, Inc., 2018.

Ramprasad, Rampi et ál. "Machine learning in materials informatics: recent applications and prospects". *npj Computational Materials* 3.1 (2017): 54.

Ribeiro, Marco Tulio, Singh, Sameer y Guestrin, Carlos. "Model-agnostic interpretability of machine learning". *arXiv 1602.04938* (2016): 1-5.

Roscher, Ribana et ál. "Explainable machine learning for scientific insights and discoveries". *IEEE Access* 8 (2020): 42200-42216. <https://doi.org/10.1109/ACCESS.2020.2976199>

Rosenblatt, Frank. "Two theorems of statistical separability in the perceptron". *The Mechanisation of Thought Processes: Proceedings of a Symposium Held at the National Physical Laboratory*. Vol. 1. Londres: HMSO, 1958. 419-449.

_____. *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Ann Arbor: Spartan Books, 1962.

Rudin, Cynthia. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead". *Nature Machine Intelligence* 1.5 (2019): 206-215.

Rumelhart, David E., Hinton, Geoffrey E. y Williams, Ronald J. "Learning representations by back-propagating errors". *Nature* 323.6088 (1986): 533-536. <https://doi.org/10.1038/323533a0>

Saleem, Rabia et ál. "Explaining deep neural networks: A survey on the global interpretation methods". *Neurocomputing* 513 (2022): 165-180. <https://doi.org/10.1016/j.neucom.2022.09.129>

Salmon, Wesley C. "Statistical explanation". *The Nature and Function of Scientific Theories*. Robert Colodny (ed.). Pittsburgh: University of Pittsburgh Press, 1970. 173-231.

Saxe, Andrew., Nelli, Stephanie., y Summerfield, Christopher. "If deep learning is the answer, what is the question?". *Nature Reviews Neuroscience* 22.1 (2021): 55-67. <https://doi.org/10.1038/s41583-020-00395-8>

Schmidt, Robin M. "Recurrent Neural Networks (RNNs): A gentle introduction and overview". *arXiv abs/1912.05911* (2019): 1-16.

Shehab, Mohammad et ál. "Artificial neural networks for engineering applications: a review". *Artificial Neural Networks for Renewable Energy Systems and Real-World Applications*. Ammar H. Elsheikh y Mohamed Elasyed Abd Elaziz (eds.). Academic Press, 2022. 189-206.

Silvestrini, Stefano y Lavagna, Michèle. "Deep learning and artificial neural networks for spacecraft dynamics, navigation and control". *Drones* 6.10 (2022). <https://doi.org/10.3390/drones6100270>

Singh, Yogesh et ál. "Application of neural networks in software engineering: A review". *Information Systems, Technology and Management*. Sushil K. Prasad et ál. (eds.). Berlín/Heidelberg: Springer Berlin Heidelberg, 2009. 128-137.

Slack, Dylan et ál. "Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods". *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. Nueva York, NY: Association for Computing Machinery, 2020. 180-186.

Sokol, Kacper et ál. "What and how of machine learning transparency: Building bespoke explainability tools with interoperable algorithmic components". *Journal of Open Source Education* 5.58 (2022): 175. <https://doi.org/10.21105/jose.00175>

Song, Jianing, Rondao, Duarte y Aouf, Nabil. "Deep learning-based spacecraft relative navigation methods: A survey". *Acta Astronautica* 191 (2022): 22-40. <https://doi.org/10.1016/j.actaastro.2021.10.025>

Soniya, Paul, Sandeep y Singh, Lotika. "A review on advances in deep learning". *2015 IEEE Workshop on Computational Intelligence: Theories, Applications and Future Directions (WCI). IEEE* (2015): 1-6. <https://doi.org/10.1109/WCI.2015.7495514>

Suppes, Patrick. "Models of data". *Logic, Methodology and Philosophy of Science: Proceedings of the 1960 International Congress*. Ernest Nagel et ál. (eds.). Stanford University Press, 1962. 252-261.

Tang, Binhua et ál. "Recent advances of deep learning in bioinformatics and computational biology". *Frontiers in Genetics* 10 (2019): 214.

Tomašev, Nenad et ál. "AI for social good: Unlocking the opportunity for positive impact". *Nature Communications* 11.1 (2020): 2468. <https://doi.org/10.1038/s41467-020-15592-3>

Van Gigch, John P. *System Design Modeling and Metamodeling*. Nueva York: Springer Science, 1991.

Van Fraassen, Bas C. *The Scientific Image*. Oxford: Clarendon Press, 1980. <https://doi.org/10.1093/0198244274.001.0001>

Vedantam, Ramakrishna et ál. "Probabilistic neural symbolic models for interpretable visual question answering". *Proceedings of the 36th International Conference on Machine Learning*. Vol. 97. Kamalika Chaudhuri y Ruslan Salakhutdinov (ed.). Red Hook, NY: Curran Associates Inc., 2019. 6428-6437.

Verreault-Julien, Philippe. "How could models possibly provide how-possibly explanations?". *Studies in History and Philosophy of Science Part A* 73 (2019): 22-33. <https://doi.org/10.1016/j.shpsa.2018.06.008>

Vilone, Giulia y Longo, Luca. "Notions of explainability and evaluation approaches for explainable artificial intelligence". *Information Fusion* 76 (2021): 89-106. <https://doi.org/10.1016/j.inffus.2021.05.009>

Wang, Xizhao, Zhao, Yanxia y Pourpanah, Farhad. "Recent advances in deep learning". *International Journal of Machine Learning and Cybernetics* 11.4 (2020): 747-750. <https://doi.org/10.1007/s13042-020-01096-5>

Woodward, James y Ross, Lauren. "Scientific explanation". *The Stanford Encyclopedia of Philosophy*. Edward N. Zalta. (ed.). Summer/Metaphysics Research Lab, Stanford University, 2021. *Online.* <https://plato.stanford.edu/archives/sum2021/entries/scientific-explanation/>

Yang, Xin-She. *Engineering Optimization: An Introduction with Metaheuristic Applications*. Hoboken: Wiley, 2010.

Zednik, Carlos. "Solving the black box problem: A normative framework for rxplainable artificial intelligence". *Philosophy & Technology* 34.2 (2021): 265-288.