Universidad Nacional Autónoma de México

Posgrado en Filosofía de la Ciencia

Facultad de Filosofía y Letras

Instituto de Investigaciones Filosóficas

La necesidad de "comprensión" para dar cuenta de los problemas prácticos derivados del problema de la caja negra dentro de la Inteligencia Artificial Explicable

Tesis

Que para optar al grado de maestro de filosofía de la ciencia (Ciencias Cognitivas)

Presenta:

Lic. Juan Francisco Ortiz Moreno

Dra. Atocha Aliseda Llera, Instituto de Investigaciones Filosóficas (Codirectora)

Dra. Karen González Fernández, Universidad Panamericana (Codirectora)

Dr. Alejandro Vázquez del Mercado, Facultad de Filosofía y Letras UNAM

Dr. Andrea Onofri, Universidad Autónoma de San Luis Potosí

Dr. Francisco Hernández Quiroz, Facultad de Ciencias UNAM

México, CDMX, Diciembre de 2022

Agradecimientos

Esta tesis se hizo bajo condiciones extrañas. Mientras estaba en el proceso de admisión a esta maestría inició la pandemia de COVID-19. De los dos años escolarizados, año y medio fue a distancia. Los seis meses restantes, los pasé adaptándome a una nueva ciudad. Ahora más que nunca se resalta la naturaleza colaborativa de estos trabajos. Aunque sea mi nombre el que está en la portada y sea yo quien tiene que pararse frente a los y las sinodales, este texto no hubiese sido posible sin un muy amplio conjunto de personas e instituciones que hicieron posible el que me sentara a investigar y redactar. Desde quienes dieron las condiciones materiales, hasta quienes con una palabra o su mera compañía me permitieron continuar. Nombrar a todos y a todas es un trabajo imposible, sin embargo, me gustaría destacar a algunos y algunas.

Primero que nada, como siempre, este trabajo es un gran porcentaje de mi padre. Ha sido, es y será un pilar fundamental para mis metas académicas y vitales. Sin su apoyo económico y emocional no hubiese sido concebible que estudiase una licenciatura y menos un posgrado. Le estoy eternamente agradecido.

Agradecimientos a mi gata Mandy, fuente inagotable de serotonina. A les seres de luz Laura, Alejandra, Jorge y Bianca por el chismecito, por las risas, las cervezas, los cafés y por su amistad. Hicieron de esta travesía algo disfrutable. Mención honorifica para Bianca, compañera de traumas universitarios. Cerca del 60-70% de la tesis la escribí lleno de cafeína en su compañía. Si bien estábamos igual de perdidos, es más soportable estar perdidos juntos. Agradecimientos también a mi amiga y amigo, compañeros de piso, Alejandro y Valeria, cambiarse de ciudad con amigos siempre es más fácil. A los risk amigos, Rolando, David y Aarón, por las noches de juegos que eran un oasis en medio del estrés.

Gracias a los miembros del comité de titulación, Alejandro Vázquez del Mercado y Francisco Hernández Quiroz por la disposición para leer y dar comentarios a esta tesis. A mis directoras Atocha Aliseda y Karen Gónzalez. He aprendido muchísimo de cómo ser un profesional de la filosofía interactuando con ellas y los seminarios que dirigen han sido una gran fuente de inspiración. Muchas gracias por la diligencia y el esfuerzo que pusieron para que este texto sea lo mejor posible. No me olvido de Andrea Onofri, quién ha sido mi profesor/mentor desde la licenciatura, muchísimas gracias por toda su ayuda en mi naciente carrera académica.

Seguramente dejo fuera a muchos y a muchas, seguramente hay tantos que ni siquiera es contable. Ante la individualidad de la academia, que viva el colectivo. ¡Muchas gracias a todos y a todas!

Índice

Introducción

En el día a día, se utilizan herramientas de inteligencia artificial (AI) para ayudarnos a resolver una muy amplia gama de problemas: clasificar bases de datos, como los catálogos de Netflix o Spotify para recomendar qué ver o escuchar; hacer diagnósticos médicos para recomendar algún tratamiento; o analizar bases de datos para saber si es recomendable para una institución bancaria el otorgar un préstamo a una persona, por poner unos ejemplos. Muchas de estas herramientas están entrenadas con base en una técnica computacional llamada "aprendizaje de máquina", la cual consiste, a muy grandes rasgos, en recopilar un gran conjunto de datos, plantear un objetivo al que queramos llegar, y dejar que una red neuronal artificial detecte regularidades (dentro de dicho conjunto de datos) que nos ayuden a llegar al objetivo.

Tales algoritmos son muy exitosos en sus tareas, un ejemplo de los muchos que hay es *Alphafold* de *DeepMind*. Se trata de una AI que se dedica a predecir la estructura de nuevas proteínas a partir de su secuencia de aminoácidos. Ésta AI se ha utilizado con mucho éxito en la investigación en biología (Jumper et al., 2021) aunque, como muchos sistemas con base en aprendizaje profundo, es opaca, es difícil (por decir lo menos) saber por qué hace lo que hace, cómo llegan a tomar decisiones o a generar predicciones. Este problema se conoce como el problema de la caja negra y no es trivial dado que estos sistemas inciden en el mundo de forma significativa. Siguiendo a Zednik (Zednik, 2021), comparto que el problema tiene al menos tres aristas: práctico (es deseable que el usuario final confíe en el sistema y para esto ayuda que sea transparente, es decir, que se pueda saber cómo es que el sistema tomó una decisión, además, es deseable que los programadores sepan cómo funciona el sistema para mejorarlo o corregirlo); legal (es importante saber a quién le atribuiremos responsabilidad por las decisiones, y las consecuencias de éstas, tomadas por los agentes artificiales o con la ayuda de estos) y teórico (el problema de la caja negra no permite comparar la similitud que podrían tener las redes neuronales artificiales con los cerebros biológicos), por lo que requiere una respuesta.

Trabajos actuales en Inteligencia Artificial Explicable (xAI) intentan resolver este problema diseñando herramientas que permitan explicar la salida y/o el funcionamiento de los sistemas

opacos. Estas herramientas presuponen nociones de qué es una explicación y presuponen que generar explicaciones es lo que hará transparentes a las cajas negras.

En esta tesis, argumento que las técnicas actuales de xAI tienen dos problemas principales: 1) no hay una definición clara de "explicación" dentro del área de la xAI, lo cual se traduce en que no hay un objetivo claro de qué buscamos al hacer investigación en pos de resolver el problema de la caja negra y 2) las explicaciones que generan no nos son útiles para resolver los problemas prácticos dado que nos responden a la pregunta "¿cómo?" y no a la respuesta "¿por qué?" un sistema arrojó una salida y no otra. Para resolver estos problemas, propongo iniciar la investigación en xAI con un paso previo a las explicaciones de las salidas, la comprensión de un modelo suficientemente verdadero que nos permita generar explicaciones contextualistas, las cuáles, a su vez, nos permiten tratar los problemas prácticos que se generan a partir del problema de la caja negra.

Para hacerlo, desarrollaré el argumento en tres capítulos. En el capítulo 1, me dedicaré a dar el estado del arte dentro de la xAI. Expondré a grandes rasgos qué es el aprendizaje de máquina y las redes neuronales artificiales para, con base en ello, definir a los sistemas de aprendizaje profundo, los cuales son paradigmáticamente opacos. Para esta sección, utilizaré el texto clásico de Russell y Norvig (Russell & Norvig, 2020). Si bien como fuente única no ofrece un análisis a detalle de las problemáticas que involucran cada uno de los conceptos expuestos, da definiciones operativas que son un buen punto de partida y no impactan la rigurosidad de nuestro análisis. Posteriormente, definiré el problema de la caja negra y expondré las dos técnicas actuales que buscan resolverlo: la interpretabilidad *post hoc* y el diseño de cajas transparentes.

En el capítulo 2, analizaré el concepto de "explicación" desde la filosofía de la ciencia, en específico analizaré las explicaciones nomológicas-deductivas y las explicaciones causales-mecanicistas para argumentar que 1) las técnicas actuales de inteligencia artificial explicable presuponen, aunque sea implícitamente, nociones de explicación que no nos permiten resolver el problema de la caja negra y 2) las explicaciones contextualistas son el modelo de explicación que nos podría ayudar a tratar con los problemas prácticos derivados de los sistemas opacos.

Finalmente, en el capítulo 3, analizaré el concepto de “comprensión” a partir del trabajo de Catherine Elgin (Elgin, 2004, 2007). Iniciaré por argüir que la relación entre “explicación” y “comprensión” es que la segunda implica a la primera. Tras mostrar que esto es el caso, argumentaré que hemos de partir de comprender para poder generar explicaciones. Después, expondré el modelo de Elgin para argüir que, para lograr comprensión, hacemos uso de idealizaciones y proposiciones suficientemente verdaderas con referencia a un fin en específico, por lo que la comprensión también es contextual. Finalmente defenderé que es necesario partir de la comprensión de una teoría de trasfondo para poder generar explicaciones contextualistas que nos permitan tratar con los problemas prácticos que generar los sistemas opacos.

En este texto, todas las traducciones son mías.

1.0 Inteligencia artificial

Con “inteligencia artificial” (IA) me referiré a sistemas computacionales que realizan tareas similares a las de la cognición humana como resolver problemas, navegar por el mundo, razonar o aprender. De acuerdo al influyente texto *Artificial Intelligence: A Modern Approach*, de los investigadores Stuart Russell y Peter Norvig, hay dos aproximaciones con base en las cuales se han generado sistemas de inteligencia artificial: 1) humano vs racional y 2) pensamiento vs comportamiento(Russell & Norvig, 2020). De éstas se desprenden cuatro combinaciones posibles que corresponden a distintos campos de investigación dentro de la IA:

I) Sistemas que simulan pensar como humanos: Dedicados a tareas como resolución de problemas o aprendizaje. Sistemas de redes neuronales artificiales son un ejemplo.

II) Sistemas que simulan actuar como humanos: Dedicados a interactuar en el mundo. Un ejemplo son los autos autónomos o las aplicaciones de robótica.

III) Sistemas que simulan pensar racionalmente: Dedicados a simular el pensamiento humano. Usualmente se clasifican aquí a los sistemas expertos con base en lógica.

IV) Sistemas que simulan actuar racionalmente: Dedicados a emular el comportamiento humano al interactuar con el mundo. Una aplicación, por ejemplo, es diseñar robots que aprendan a utilizar artefactos.

Cada uno de estos campos de investigación, se dedica a resolver problemas dentro de las seis áreas prototípicas de la AI:

i) Procesamiento de lenguaje natural: comunicarse, producir o entender exitosamente el lenguaje humano.

ii) Representación del conocimiento: almacenar conocimiento y meta conocimiento mediante representaciones.

iii) Razonamiento automatizado: generar inferencias a partir del conocimiento previo y hacia conocimiento nuevo.

iv) Aprendizaje de máquina: adaptarse a nuevas circunstancias, así como detectar y extrapolar patrones cada vez mejores entre más experiencia tenga el sistema.
v) Visión computacional: percibir e interactuar con el mundo mediante el reconocimiento de imágenes.
vi) Robótica: manipular objetos y moverse en el mundo.

En este trabajo, me centraré en el área de aprendizaje de máquina y los problemas que nos genera.

1.1.0 Aprendizaje de máquina

Decimos que un agente o sistema aprende cuando su desempeño mejora con la experiencia. Cuando este aprendizaje sucede en una computadora, lo llamamos aprendizaje de máquina. A un sistema lo "alimentamos" con una base de datos, el sistema construye un modelo de estos datos y utiliza el modelo para generar hipótesis, predicciones y resolver problemas (Russell & Norvig, 2020).

Esta aproximación tiene la ventaja de que puede extrapolar lo conocido a escenarios nuevos, por ejemplo, un sistema para predecir la bolsa de valores debe de ser lo suficientemente flexible para adaptarse a los cambios drásticos que pueda haber. Si este sistema estuviese basado en lógica, sería demasiado rígido para poder adaptarse, cada que surja una nueva variable tendríamos que escribirla en el código. Como estas variables surgen en tiempo real y necesitamos el cambio en tiempo real, sistemas basados en lógica sería muy poco adaptables a estos cambios. Otra ventaja es que con el aprendizaje de máquina le podemos dar soluciones a problemas que quizá no sepamos cómo resolver con código, como el reconocimiento facial (Russell & Norvig, 2020), tarea que implica muchísimas variables y que, como humanos, parece que resolvemos intuitivamente y no siguiendo un algoritmo.

Hay tres técnicas de aprendizaje de máquina:

1) Aprendizaje supervisado.
2) Aprendizaje no supervisado.
3) Aprendizaje de refuerzo.

En el aprendizaje supervisado, el sistema observa la entrada y la salida y aprende la función que lleva del primero al segundo. Por ejemplo, en un sistema de clasificación de imágenes,

el sistema contaría con la imagen inicial y su etiqueta. Con la exposición a una base de datos de imágenes etiquetadas, aprenderá a hacer nuevas predicciones cuando se le presente una imagen nueva sin etiquetar.

En el aprendizaje no supervisado el sistema aprende patrones a partir de la entrada sin ningún tipo de retroalimentación explícita, mientras que, en el aprendizaje reforzado, el sistema aprende con base en una serie de refuerzos en forma de "recompensas" y "castigos" (Russell & Norvig, 2020). Un ejemplo son los sistemas que se dedican a juegos. Si un sistema que juega ajedrez pierde, ese será su "castigo" y tendrá que averiguar cuál de los pasos previos al refuerzo fue el responsable de la salida y tendrá que modificar las acciones para maximizar las recompensas en el futuro.

Si bien estas técnicas de aprendizaje son posibles en sistemas basados en lógica, al hablar de "aprendizaje de máquina" hablaré exclusivamente del aprendizaje de máquina que se implementa en sistemas con base en redes neuronales artificiales.

1.1.1 Redes neuronales artificiales

Muy a grandes rasgos, una red neuronal artificial es un modelo computacional que está inspirado en cómo se cree que funcionan los cerebros humanos. Sus componentes básicos son las neuronas artificiales, pequeños elementos computacionales que poseen un valor de activación numérico e influyen en los elementos vecinos. La conexión entre neuronas, sus "sinapsis" para seguir con la analogía neuronal, está cuantificada y se le llama peso. Entre más fuerte su conexión, mayor será su peso.

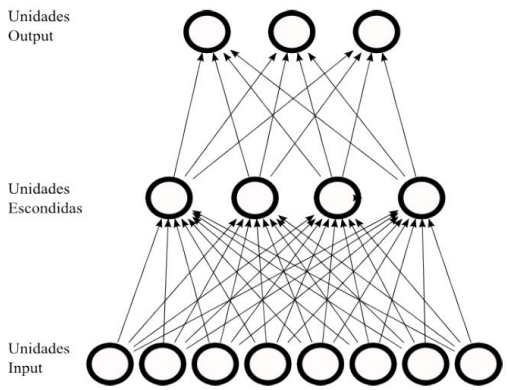

Ilustración 1 Red neuronal artificial, el *output* es la salida y el *input* la entrada.

Así, la influencia de una unidad I en una unidad J es, en este modelo, el valor de activación de I más la fuerza de la conexión de I y J. De este modo, si una unidad posee un valor de activación positivo, su influencia en la unidad vecina será positiva, si su peso hacia esa unidad es positivo, y negativa si el peso es negativo[1]. Siguiendo con la alusión neuronal, las conexiones con pesos positivos son llamadas excitatorias, mientras que las de peso negativo son llamadas inhibidoras. En estos modelos, la entrada del sistema es provisto al poner valores de activación en las unidades de entrada de la red, estos valores numéricos representan algún tipo de codificación o representación de la entrada. Por ejemplo, si ponemos una imagen como entrada, ésta es representada numéricamente en la capa de entrada. La activación en las unidades de entrada se propaga a través de las conexiones hasta que una configuración de valores de activación emerge en las unidades de salida, estos codifican la salida que el sistema ha computado a partir de la entrada (Smolensky, 1988).

Entre las unidades de entrada y las de salida hay otras unidades, las llamadas unidades ocultas, que participan sin representar ni la salida ni la entrada. El cómputo realizado por la red al transformar el patrón de actividad de la entrada depende de la configuración de los pesos de las conexiones. En este sentido, las fuerzas de conexión juegan el papel del programa en una computadora convencional, en ellas están codificadas las operaciones que se han de realizar para hacer un cómputo, sin embargo, a diferencia de un programa convencional, la información no está localizada y nunca hubo una instrucción explícita dada *a priori* por un programador.

Para lograr este tipo de cómputos, un modelo de redes neuronales artificiales es entrenado para desempeñar tareas específicas. Aprende a realizarlas configurando los pesos de las conexiones entre sus unidades. Si recordamos, decimos que un sistema "aprende" cuando con la experiencia, su desempeño mejora. Mucho del atractivo de las redes neuronales artificiales es precisamente este, que las redes se pueden "programar" a sí mismas.

[1] Esto puede no ser siempre el caso, dependiendo de qué tipo de unidad se trate, existen unidades cuyo valor de activación no es digital, sino analógico. Por ahora, baste con esta aproximación.

1.1.2 Sistemas de aprendizaje profundo: Breve introducción

"Aprendizaje profundo" refiere a una familia de técnicas de aprendizaje de máquina en las cuáles las hipótesis toman la forma de circuitos algebraicos complejos con fuerzas de conexión ajustables. Típicamente, estos circuitos están organizados en muchas capas, lo que implica que el cómputo tiene múltiples pasos. Son redes neuronales artificiales muy densas, es decir, con muchas neuronas y, sobre todo, muchas capas ocultas.

Típicamente, un modelo con base en aprendizaje profundo recibe como insumo una base de datos masiva previamente etiquetada y depurada llamada "cuerpo de entrenamiento", es entrenada para realizar una tarea específica con alguna técnica de las que resumí en la sección 1.1.0., haciendo uso de procesos estadísticos y estocásticos. Durante esta fase de entrenamiento, el sistema crea un modelo de los datos al ajustar sus pesos de conexión entre las unidades. A este modelo, no podemos acceder ni lo hemos programado, el sistema lo ha generado por su cuenta a partir de los datos. Finalmente, al completar el entrenamiento, el sistema es capaz de hacer predicciones en el campo para el que se le haya entrenado.

Por ejemplo, un sistema que emita recomendaciones en Twitter toma como cuerpo de entrenamiento los datos de las interacciones e intereses del usuario, al entrenarse, aprendería a encontrar correlaciones entre los temas de interés. Una vez que lo logra, emite una recomendación que podría ser interesante para el usuario. Sistemas como estos son utilizados en el día a día en disciplinas tan variadas como la medicina, la biología, la física o la economía. Sin embargo, el que el sistema genere autónomamente un modelo sobre los datos que le son suministrados crea el llamado "problema de la caja negra".

1.2.0 El problema de la caja negra

El problema de la caja negra consiste en que los sistemas con base en aprendizaje de máquina son opacos, es decir, no sabemos cómo es que los sistemas llegan a sus salidas ni cómo es que toman las decisiones que toman (Carabantes, 2020). Estos sistemas, además, entre más complejos se vuelven son más exitosos, generan mejores predicciones y toman mejores

decisiones (Bubeck & Sellke, 2021). Sin embargo, no sabemos cómo lo hacen. En su libro *Weapons of Math Destruction* Cathy O'Neil nos da ejemplos de cómo el problema de la caja negra impacta en el día a día dado que sistemas opacos se utilizan para tomar decisiones como dar o no un préstamo a una persona, o calcular el riesgo de reincidencia en el delito de una persona para decidir si se le concede o no libertad condicional (O'Neil, 2016). Una "arma de destrucción matemática" como la llama O'Neil, es una caja negra escalable que produce daño. Dado que es una caja negra, no podemos saber, por ejemplo, si está reproduciendo sesgos de género o raciales al emitir sus recomendaciones (las salidas nos parecen indicar que sí), el que sea escalable implica que su área de agencia es potencialmente cada vez mayor y genera daño al, por ejemplo, negarle un préstamo a una persona que lo requiere. El problema de la caja negra, por tanto, no es solo una preocupación teórica, impacta en la vida de muchas personas.

Por lo anterior, siguiendo a Zednik, afirmo que el que los sistemas con base en aprendizaje de máquina sean opacos es problemático en, al menos, tres sentidos (Zednik, 2021):

1) Práctico: Es deseable que los usuarios y los desarrolladores sepan con base en qué el sistema emite una recomendación. Esto para poder corregir e identificar problemas como sesgos raciales y de género u otro tipo de correlaciones espurias.
2) Legal: Es deseable saber a quién o a qué le atribuiremos responsabilidad por las decisiones tomadas por o con la ayuda de sistemas opacos.
3) Teórico: El problema de la caja negra no nos permite comparar la similitud entre las redes neuronales artificiales y los cerebros biológicos. En este sentido, el que los sistemas de redes neuronales sean cajas negras no nos permite utilizarlos como herramientas para estudiar la cognición humana[2].

Añadiría como problema teórico el que los sistemas opacos no nos permiten producir explicaciones científicas ni alcanzar comprensión de fenómenos como argumentaré a lo largo

[2] Esto no es un problema si, de entrada, rechazamos el conexionismo, la teoría de la arquitectura cognitiva que nos dice que las habilidades cognitivas de alto nivel tienen su base en redes neurales y la información está codificada de forma no-simbólica en los pesos de las conexiones de dicha red. Ni Zednik y yo lo rechazamos, pero no exploraré este problema en este trabajo. Para más información sobre el debate del conexionismo contra la postura clásica en ciencias cognitivas se puede consultar (Chalmers, 1993; J. A. Fodor & Pylyshyn, 1988; J. Fodor & McLaughlin, 1990; Smolensky, 1988)

de este trabajo. Para lidiar con el problema de la caja negra, en mayo de 2017, la Agencia de Proyectos de Investigación Avanzados de Defensa (DARPA, por sus siglas en inglés) de los EEUU, lanzó el programa de Inteligencia Artificial Explicable (xAI).

1.3.0 Inteligencia Artificial Explicable (xAI)

La inteligencia artificial explicable (xAI) es un proyecto del DARPA cuyo objetivo es resolver el problema de la caja negra al crear sistemas transparentes, en sus palabras: "sistemas de inteligencia artificial que utilicen nuevas o modificadas técnicas de aprendizaje de máquina que produzcan modelos explicables que, combinadas con técnicas efectivas de explicación, le permitan a los usuarios finales entender, confiar y gestionar adecuadamente a la generación emergente de sistemas de inteligencia artificial" (DARPA, 2016).

Desde la publicación del proyecto de investigación del DARPA se han producido una gran cantidad de artículos científicos que buscan dar explicaciones a los procesos o a las salidas de los sistemas opacos. Los esfuerzos para resolver el problema de la caja negra los podemos agrupar en dos grandes conjuntos: diseño de cajas transparentes e interpretabilidad *post hoc*.

1.3.1 Diseño de cajas transparentes

La estrategia de diseñar cajas transparentes consiste en incluir, en los sistemas opacos, otros algoritmos tales como árboles de decisiones, conjuntos de reglas booleanas o modelos aditivos generalizados (Durán, 2021). Se presupone que estos nuevos algoritmos son transparentes para los agentes especializados en el área de interés, como los médicos en medicina o los científicos computacionales en ciencias computacionales. Estos agentes pueden explicarse la salida del sistema rastreando la cadena causal que nos llevó a dicho resultado, así como qué variables estuvieron en juego al realizar el cómputo. La forma en que estos sistemas explican una salida O es a través de mostrar la secuencia de pasos que llevaron a la salida O.

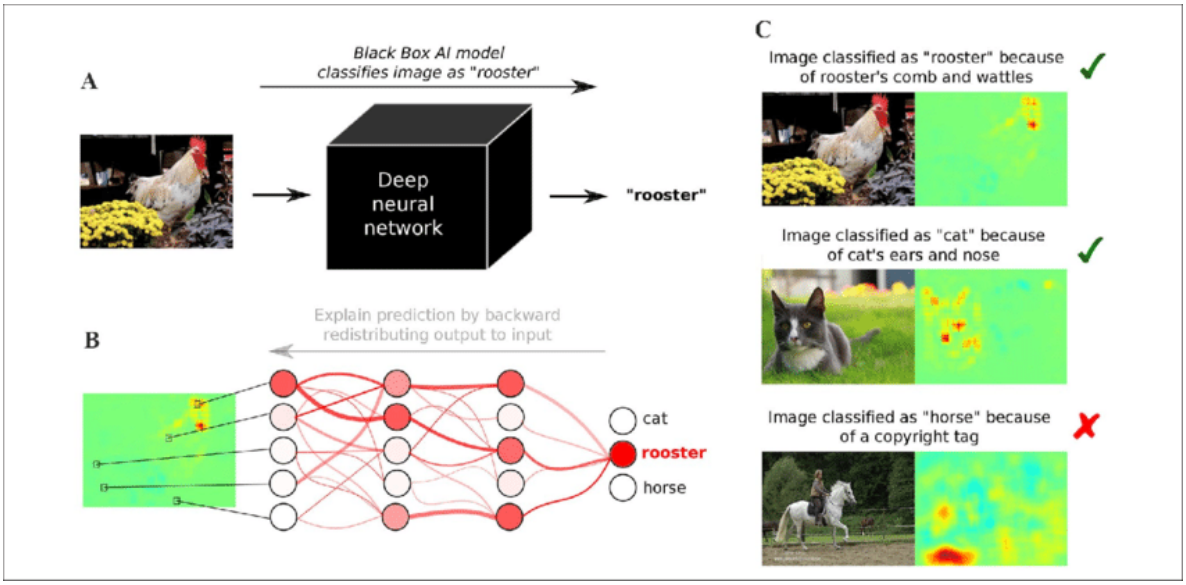

Ilustración 2 Ejemplo de aplicación del diseño de cajas transparentes (Schwendicke et al, 2020)

Un ejemplo lo podemos ver en la ilustración 2. En ella, buscamos explicar cómo es que un sistema de clasificación de imágenes emite sus predicciones. A partir de rastrear cómo se llegó a la salida desde la entrada, este sistema infiere qué partes de la imagen fueron las más relevantes para emitir la predicción y da como resultado final la predicción y el mapa de calor que muestra las regiones relevantes en la imagen. Si preguntamos "¿por qué el sistema clasificó la imagen de un gallo como un gallo?" Podríamos responder que los hizo por la cresta y por la papada, de acuerdo al mapa de calor. En cambio, en un error potencial, como el caso del caballo, podríamos explicar qué salió mal, generando un modelo transparente, al menos en teoría.

1.3.2 Interpretabilidad *post hoc*

La segunda técnica para alcanzar sistemas transparentes es la interpretabilidad *post hoc*. Ésta consiste en generar un nuevo modelo, llamado "modelo intérprete", que sirva como intermediario entre el sistema opaco y el usuario. El modelo intérprete se encarga de analizar cuáles han sido los parámetros relevantes tomados por el modelo opaco para llegar a una decisión y luego, explicar al usuario la salida del modelo opaco (Durán, 2021). En el nivel

más básico, hace lo mismo que el diseño de cajas transparentes, mostrar la secuencia de pasos que llevaron al algoritmo a la salida O.

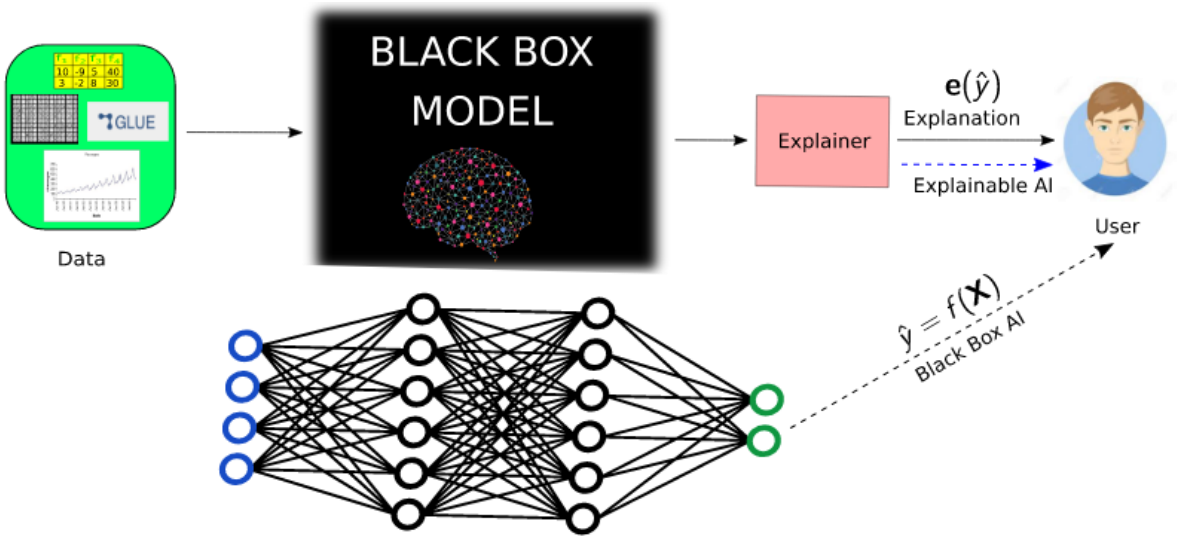

Ilustración 3 Ejemplo de interpretabilidad *post hoc* (Maveli, 2021).

En el diagrama de la ilustración 3 vemos un ejemplo. Ahí, a partir de los datos, el sistema genera un modelo de caja negra y emite predicciones. Para que esta caja sea transparente, usamos un modelo nuevo que sirve de interprete cuya misión es generar una explicación de la caja negra. Esta explicación se da de forma visual o textual (Carabantes, 2020) y, de nuevo, la presuposición es que, para los agentes adecuados, esta explicación será transparente.

1.4.0 Problemas con las técnicas actuales de xAI

Estas dos técnicas poseen problemas de fondo que avanzaré ahora para profundizar en el siguiente capítulo. A saber, presuponen definiciones de "explicación", y buscan dar explicaciones con ellas, que no son aplicables a los modelos de aprendizaje profundo. Además, confunden qué tipo de pregunta queremos que nos respondan los sistemas opacos para resolver los problemas prácticos que nos generan. Las técnicas de xAI solo pueden explicar "cómo" se llegó a la salida O, no "por qué" y, para resolver el problema de la caja negra, necesitamos responder a la pregunta "por qué" y no solo a la pregunta "cómo" como argumentaré más adelante.

En el próximo capítulo, analizaré el concepto de “explicación” desde la filosofía de la ciencia para justificar las afirmaciones hechas en el párrafo anterior. Posteriormente, expondré un modelo de explicación, las explicaciones contextuales, que podrían ayudarnos a esclarecer qué tipo de explicaciones necesitamos en la xAI y qué debemos tomar en cuenta para alcanzarlas.

2.0 Explicación

En filosofía de la ciencia, existen múltiples caracterizaciones de qué es una explicación. La más cercana al problema de la caja negra es la visión argumental. En ella, una explicación es un argumento que subsume un fenómeno a un marco teórico más general (de Regt, 2009) y contiene, al menos, tres partes esenciales: *explanans, explanandum* y una relación explicativa[3]. *Explanans* es la unidad que porta la información relevante para la explicación, el *explanandum* la unidad a ser explicada y la relación explicativa es la relación de dependencia entre *explanans* y *explanandum*.

Hay dos grandes conjuntos de tipos de explicación científica: objetivista y pragmática. Los objetivistas suelen creer que hay un criterio fijo e independiente de las mentes de los sujetos que cumplen todas las explicaciones. Estos criterios son universales en varios sentidos: aplican a todas las explicaciones científicas, no incorporan presuposiciones empíricas específicas que puedan ser hechas por científicos de distintas áreas, y son independientes de los intereses de audiencias particulares (Durán, 2021).

Los pragmáticos, en cambio, consideran que la psicología de quienes dan explicaciones y de quienes las reciben tienen un rol en la explicación misma. Dentro del conjunto pragmático se suelen ofrecer explicaciones contextuales. Una explicación contextual es el tipo de explicación que, dada la exposición de un cuerpo específico de información a cierta audiencia, producirá un sentido de inteligibilidad condicionada al conocimiento previo y a los intereses de la audiencia.

Dentro de las distintas teorías objetivistas de la explicación se encuentran los modelos nomológico-deductivo e inductivo-estadístico de Hempel, la explicación causal-mecanicista de Salmon y sus derivados. Estas teorías de la explicación asumen los supuestos de las visiones objetivistas y, como argumentaré, son el tipo de explicaciones que las técnicas actuales de xAI buscan proveer.

[3] Durán añade que esta relación explicativa es objetiva, yo creo que no, el argumento lo desarrollo en el capítulo 3.

2.1.0 Explicaciones científicas

Asumimos que la ciencia da explicaciones y no solo descripciones de los fenómenos que estudia. Bajo esta asunción, cuando damos una explicación científica, buscamos responder a la pregunta "¿por qué sucede x?" donde x puede ser un evento particular o alguna regularidad[4]. La tarea de una teoría o modelo de las explicaciones científicas es, entonces, caracterizar la estructura de tales explicaciones.

Algunos de los primeros autores en caracterizar la estructura de las explicaciones científicas fueron Carl Hempel y Paul Oppenheim. De acuerdo a los autores, la marca principal del conocimiento científico es su naturaleza objetiva. Las explicaciones que genera el conocimiento científico son independientes de las mentes de los científicos, por lo que los filósofos y filósofas de la ciencia han de ofrecer una visión objetivista de la ciencia (independiente de la mente de los sujetos que proveen y reciben explicaciones), y de la explicación científica en particular, ignorando aspectos que dependen de las mentes de los sujetos. Para lograrlo, Hempel propone el modelo nomológico-deductivo(Hempel, 1958).

2.2.0 Modelo nomológico-deductivo

De acuerdo al modelo nomológico deductivo, una explicación consta de un *explanandum* y un *explanans*. El *explanandum* es una afirmación que describe el fenómeno a ser explicado y el *explanans* es el conjunto de afirmaciones que dan cuenta del fenómeno por explicar (Hempel, 1965). El *explanans*, a su vez, tiene dos subconjuntos: 1) el conjunto de enunciados C_1, C_2, …, C_k que enuncian las condiciones iniciales y 2) el conjunto de enunciados L_1, L_2, …, L_k que representan leyes generales.

Si una explicación dada es sólida, cumple con ciertas condiciones empíricas y lógicas de adecuación. Las condiciones lógicas son:

i) El *explanandum* ha de ser consecuencia lógica del *explanans*.

ii) El *explanans* ha de incluir al menos una ley nomológica y ésta es requerida necesariamente para derivar el *explanandum*.

[4] Es cierto que existen actividades científicas puramente descriptivas o taxonómicas y que esto haría que esta afirmación no sea verdadera en todos los casos. Sin embargo, aún en las actividades puramente taxonómicas, estoy asumiendo que tienen como propósito a nivel macro responder una pregunta por qué.

iii) El *explanans* ha de tener contenido empírico, es decir, ha de ser capaz de ser corroborado o falseado mediante la experimentación o la observación.

Por otro lado, la condición empírica es:

iv) Las oraciones que constituyen el *explanans* deben ser verdaderas.

A esta última condición se le conoce también como la condición de factividad[5].

Podemos ver el modelo nomológico-deductivo en el siguiente esquema:

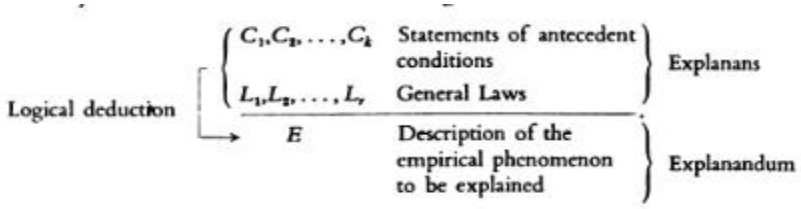

Ilustración 4: Modelo nomológico-deductivo (Hempel, 1965)

Como podemos ver, para Hempel, una explicación toma la estructura de un argumento deductivo sólido[6], donde el *explanandum* se sigue como la conclusión de las premisas en el *explanans* y todas las premisas son verdaderas. De ahí lo "deductivo" del modelo. En ii podemos ver el componente nomológico.

Una explicación siguiendo el modelo nomológico-deductivo se puede dar en inteligencia artificial. Supongamos que tenemos un sistema experto que se dedica a actualizar una base de datos bancaria que determina si puedes o no obtener un préstamo en ese banco. Esta base de datos cuenta con tu historial crediticio, datos sobre tus ingresos y sobre tu situación socio económica en general, y, como ley[7], cuenta con la oración "A todo S que tenga un historial crediticio negativo, no se le otorgará un préstamo", entre otras leyes que determinan cuándo sí se otorgará dicho préstamo. Queremos explicar la salida:

[5] No hay un término adecuado, que pudiese encontrar, para traducir *factivity*. Usaré factividad.
[6] Un argumento deductivo sólido es aquel que es válido y todas sus premisas son verdaderas.
[7] Esta no es estrictamente una ley nomológica, pero cumple con las condiciones de ser parte central de la explicación y de ser contrastable empíricamente. La noción de "ley" es problemática en filosofía de la ciencia y hay un amplio debate abierto sobre ella. No entraremos en ese debate. Para profundizar en él, véase: (Armstrong, 1983; Carroll, 1994; Shumener, 2019)

F: No se te otorgará un préstamo.

Dado que el sistema está cerrado bajo consecuencia lógica y sabemos que no hemos pagado nuestras tarjetas de crédito, podemos explicar la salida del siguiente modo:

Sea f=Francisco

L_1: Todo sujeto S que no page las tarjetas de crédito, tendrá un historial crediticio negativo.

L_2: A todo S que tenga un historial crediticio negativo, no se le otorgará un préstamo.

C_1: f no pagó sus tarjetas de crédito.

_____

E: No se le otorgará un préstamo a f.

Nuestro modelo, por tanto, es transparente. Sabemos por qué la salida es E dadas las condiciones iniciales y las leyes relevantes. No solo eso, sino que además la conclusión es esperable. Una explicación nomológica-deductiva responde a la pregunta "¿por qué ocurrió el fenómeno descrito en el *explanandum*?" al mostrar que el fenómeno es el resultado de ciertas condiciones específicas en conexión con las leyes relevantes. Al hacer esto, el argumento nos muestra que era esperable que E ocurriera. Hempel afirma que es en este sentido en el que la explicación nos permite comprender por qué ocurrió el fenómeno.

2.2.1 Problemas con las explicaciones nomológico-deductivas en la AI actual

En la historia del campo de la Inteligencia Artificial, estos tipos de explicaciones se han ofrecido con éxito para sistemas de la llamada *Good Old Fashion Artificial Intelligence* (GOFAI) (Haugeland, 1985). Dichos sistemas tienen como base la lógica clásica.

El problema para extender este tipo de explicaciones a los actuales modelos de IA con base en aprendizaje de máquina, es que los sistemas de inteligencia artificial basados en lógica de la GOFAI son radicalmente distintos de los sistemas opacos con base en aprendizaje de máquina. Los sistemas de aprendizaje profundo aprenden a detectar regularidades en grandes

bases de datos, al hacerlo, van modificando su arquitectura interna para que se vaya aproximando cada vez más a los parámetros adecuados para resolver una tarea.

Si recordamos, un sistema de aprendizaje profundo, tiene como base una red neuronal artificial que tiene distintos valores numéricos entre las distintas "sinapsis", estos valores numéricos, llamados "pesos", en su conjunto codifican el tipo de cómputo que realiza el sistema. El tipo de cómputo que realizan estos sistemas no se basa en manipulación simbólica siguiendo reglas lógicas ni hay un código escrito por un programador sobre los pasos a seguir, sino que se trata de un procedimiento estadístico en paralelo.

Volviendo al ejemplo previo, para que un sistema de aprendizaje profundo nos dé una recomendación sobre si otorgar o no un crédito a un sujeto, lo único que tenemos que hacer es darle una base de datos, el objetivo, y dejar que alcance la solución óptima. Con la experiencia, el sistema va ajustando sus pesos de conexión hasta que cumple satisfactoriamente la tarea, a esta fase de ajuste de pesos se le llama "entrenamiento". El sistema final, una vez entrenado, ha ajustado por su cuenta los pesos de conexión de modo que no podemos rastrear la decisión como en los sistemas de GOFAI, por lo que no cumplimos con la condición ii y iii del modelo nomológico-deductivo: El *explanans* no incluye al menos una ley nomológica (o general) que sea requerida necesariamente para derivar el *explanandum* y el *explanans* no es capaz de ser corroborado o falseado mediante la experimentación o la observación.

Sobre todo, los sistemas de aprendizaje profundo son estocásticos, es decir, no deterministas. En el proceso de selección de datos y en el entrenamiento se seleccionan parámetros aleatorios con el fin de volver al sistema más robusto y más preciso (Páez, 2019), lo que implica que el resultado es impredecible hasta que el sistema se ejecuta, por lo que tampoco podemos aplicar la condición de factividad.

Resumiendo, dado que los modelos basados de aprendizaje profundo son estocásticos, no es posible garantizar la transmisión de verdad de las premisas a la conclusión y, dada su diferencia en funcionamiento en comparación con los sistemas de AI de la GOFAI, tampoco podemos cumplir con la estructura del modelo nomológico-deductivo de Hempel: no sabemos cuáles son las leyes (si las hay) ni cuáles son las condiciones específicas, por lo que las explicaciones nomológicas deductivas no pueden resolver el problema de la caja negra.

2.3.0 Explicaciones causales-mecanicistas

Otra caracterización de explicación que se ha utilizado con éxito en la historia de la inteligencia artificial, y que las técnicas de interpretabilidad *post hoc* y diseño de cajas transparentes parecen adoptar para resolver el problema de la caja negra, son las explicaciones causales-mecanicistas.

En 1978, el filósofo Peter Railton (Railton, 1978) propone el modelo nomotético-deductivo. Este modelo afirma que las explicaciones describen causas. En ocasiones, la secuencia de eventos que derivan en el evento por ser explicado pueden ser poco probables, por lo que, en lugar de buscar que la explicación nos dé que el evento sea esperable, como pedía Hempel, hemos de buscar que la explicación dé cuenta del mecanismo operante.

Salmon (Salmon, 1984) sigue la idea de Railton y habla de "nexo causal", una amplia red de procesos causales en interacción. Las explicaciones científicas explican eventos al mostrar cómo estos se acoplan en la estructura causal del mundo a través de procesos causales. Un proceso es una entidad que mantiene una estructura persistente a través del tiempo y el espacio (Glennan, 2002). Un proceso causal es un proceso que es capaz de transmitir los cambios en su estructura y una interacción causal es la intersección entre procesos causales en los cuales ocurre una alteración de las propiedades persistentes de tales procesos.

Para Salmon, las explicaciones no son argumentos, sino descripciones de las propiedades de una realidad independiente de las mentes: la estructura causal del mundo. A esto se le conoce como la concepción óntica de la explicación. En ella, el foco está en las relaciones causales entre las partes de un mecanismo y no entre las relaciones lógicas entre las proposiciones del argumento que da la explicación. La explicación es a nivel mundo y no a nivel semántico.

Ahora bien, para volver inteligibles estas relaciones causales utilizamos modelos mecanicistas. Un modelo mecanicista es el vehículo para una explicación mecánica. A partir de mostrar cómo una secuencia de eventos actuando e interactuando en cierta configuración espacio-temporal nos lleva a un resultado, estos modelos muestran cómo un fenómeno se produce o se mantiene y cuáles son los mecanismos relevantes para ello.

Veamos un ejemplo. Tenemos un sistema de AI que juega ajedrez y recomienda jugar para las blancas Qxb8+[8] en la siguiente posición:

Ilustración 5: Qxb8

El sistema nos dice que si jugamos Qxb8 ganamos, queremos explicar por qué ganamos con Qxb8. En este modelo de explicación, buscaremos cuál es el mecanismo que nos lleva a que necesariamente ganemos si jugamos ese movimiento. Las reglas del ajedrez juegan el papel de las leyes de la física y la posición de las piezas es la configuración espacio-temporal de las partes del mecanismo. Representamos esta ubicación con la notación estándar del ajedrez y creamos un árbol de decisiones, un modelo mecanicista para dar cuenta de las relaciones causales operantes. El árbol es el siguiente:

[8] En la notación estándar del ajedrez, "Q" representa a la dama, cada una de las casillas es nombrada por su coordenada y, si una pieza captura a otra en una casilla se denota con el nombre de la pieza, una "x" y la coordenada. Si, además, el rey enemigo queda en jaque, se le añade un signo de más al final. En el ejemplo, "Qxb8+" denota que la dama captura a la pieza en la casilla d8 y da un jaque.

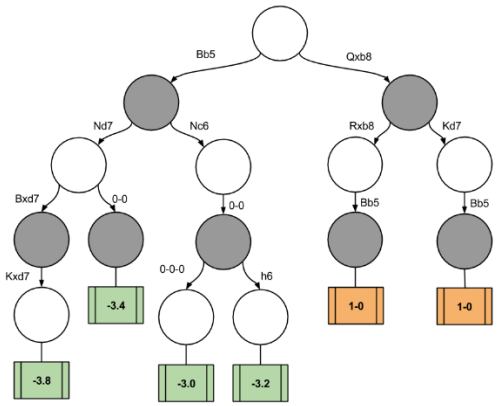

Ilustración 6: Árbol de decisiones de Qxb8

Podemos ver así, que si jugamos Qxb8 inicia una cadena causal que, dada la configuración espacio-temporal de las piezas y las reglas del ajedrez, nos lleva a ganar. La explicación mecanicista a la pregunta "¿por qué ganamos con Qxb8?" nos diría algo como lo siguiente:

Qxb8→Rxd8 v Kd7

If Rxd8

Then Bd5

If Bd5

Then 1-0

If Kd7

Then Bd5

If Bd5

Then 1-0

Qxb8

En esta secuencia representamos las interacciones causales que nos llevan a que, si jugamos Qxb8, ganemos. Sin embargo y como bien critica Duran (Durán, 2021), mostrar cómo se dio la salida no es lo mismo que responder por qué se dio la salida. En el problema de la caja negra lo que queremos saber es por qué un sistema nos da tal o cual recomendación para

atender a los problemas prácticos, legales y teóricos que mencioné en el primer capítulo. Para ello, necesitamos aún más que solo mostrar el camino causal que llevó a la salida.

2.3.1 Problemas con las explicaciones mecanicistas-causales en la AI actual

Supongamos que de hecho basta con dar el cómo se llegó a la decisión. Un primer problema es el que ya mostró Hanson en *Patterns of Discovery* (Hanson, 1958), no basta con mostrar la secuencia causal si no tenemos una teoría de trasfondo[9] que nos diga por qué las interacciones descritas en la cadena causal derivan en el fenómeno a explicar. Un segundo problema es que los sistemas de aprendizaje profundo son sistemas que, entre más complejos sean, mejores son. Modelos como GTP-3, un procesador de lenguaje natural de OpenAI, contienen cerca de 100 trillones de sinapsis y 10 mil millones de neuronas artificiales haciendo cómputos en paralelo (Brown et al., 2020). Rastrear la cadena causal es bastante complicado, por decir lo menos, por lo que si adoptamos modelos mecanicistas-causales clásicos de la explicación el sistema seguirá siendo opaco.

Las explicaciones nomológicas-deductivas y mecanicistas-causales son muy exitosas en muchas áreas, incluida la AI basada en lógica, pero no son buenos modelos para volver explicables a los sistemas de aprendizaje profundo por las razones que he dado. Sin embargo, no todo está perdido. Si bien he argumentado que explicaciones objetivistas no nos son útiles para resolver el problema de la caja negra, quedan por revisar las explicaciones pragmáticas.

2.4.0 Explicaciones pragmáticas

Recapitulando, existen dos grandes conjuntos de tipos de explicación científica, la objetivista, que ya hemos revisado, y la pragmática. Dentro del conjunto pragmático se suelen ofrecer explicaciones contextuales. Autores y autoras como van Fraasen (Fraassen, 1989), de Regt (de Regt, 2009, 2019) y Elgin (Elgin, 2004, 2007; Elgin, 2017) argumentan a favor de la posición pragmática. Los y las autoras afirman que toda explicación involucra elementos "pragmáticos" como lo son: intereses, creencias u otros componentes psicológicos de quienes

[9] *Background theory*.

dan y reciben explicaciones y/o refieren al contexto en el cual se da una explicación como un factor irreductible para proporcionar explicaciones.

En el contexto de las explicaciones científicas, "pragmático" tiene al menos dos acepciones: 1) El contexto local y la psicología de quienes dan y reciben explicaciones es un factor irreducible para la explicación y 2) la utilidad de una explicación para alcanzar algún objetivo de interés humano es la consideración pragmática (Woodward & Ross, 2021).

En la segunda acepción, las explicaciones pragmáticas afirman que las consideraciones psicológicas y contextuales tienen un rol epistémico en la explicación, pero no son parte de la estructura central de las explicaciones. La estructura central sigue siendo relaciones causales objetivas embebidas de información pragmática como los intereses y objetivos de los científicos o las comunidades científicas.

En este trabajo, hablaré de explicaciones pragmáticas en el primer sentido. Una explicación pragmática afirma que el hecho de que se produzca una comprensión de un cuerpo de información, al explicarlo a una audiencia particular A, depende del conocimiento previo de A y del contexto local. Es decir, una explicación será exitosa para lograr comprensión en una audiencia solo si toma en cuenta el conocimiento previo y el contexto local de A. Conocimiento previo y contexto son necesarios para tener explicaciones exitosas, las explicaciones que se generan a partir de este modelo de explicación se llaman explicaciones contextualistas.

Ahora bien, una de las críticas que existen contra las explicaciones objetivistas es que, para lograr producir explicaciones, es necesario tener antes una comprensión de una teoría de trasfondo que nos permita ligar los fenómenos que conforman la cadena causal en una explicación. Esto es cierto también para las explicaciones contextualistas. Filósofos dentro de esta tradición, como de Regt, afirman que para lograr explicaciones se requiere de una condición previa: comprender un cuerpo de información (de Regt, 2009).

Volvamos al ejemplo del ajedrez. El árbol de decisiones que nos daba la explicación causal de por qué el movimiento Qxb8 ganaba solo será explicativo, dirán los teóricos pragmáticos, en un contexto C donde la audiencia tenga, al menos, comprensión básica de 1) la notación estándar del ajedrez, 2) las reglas del juego, y 3) el cómo interpretar el árbol de decisión. Si

nos encontramos en un contexto C' donde la audiencia no tiene al menos uno de estos puntos como información de trasfondo, tendríamos que adecuar la información que ofrecemos para que la información que compartamos a la audiencia sea de hecho explicativa. En este ejemplo, tendríamos que añadir información acerca del punto que ignore la audiencia para que la explicación de hecho funcione. Así, en el escenario C' tendríamos una explicación pragmática. Una condición sin la cual no sería explicativa nuestra "explicación" es el conocimiento previo de a quién le estoy explicando, la teoría del ajedrez como trasfondo para poder acoplar el fenómeno "Qxb8 gana en esta posición" y comprender el árbol de decisión que se ofrece como evidencia.

Resumiendo, una explicación contextualista es una en la cual el conocimiento previo y el contexto local son condiciones irreducibles de la explicación y, además, se requiere de la comprensión de un cuerpo de información, que hemos llamado información de trasfondo, para poder producir explicaciones.

En este capítulo he expuesto cómo se han intentado, aunque sea de forma implícita, hacer explicaciones objetivistas en el proyecto de inteligencia artificial explicable siguiendo el modelo nomológico-deductivo o el modelo mecanicista-causal clásico. He criticado la implementación del modelo nomológico-deductivo como modelo de explicación porque no se acopla a las particularidades de los sistemas con base en aprendizaje de máquina en al menos dos puntos 1) estos sistemas utilizan procesos estocásticos durante el entrenamiento, por lo que no podemos cumplir con la condición de factividad y 2) no es claro que existan leyes para poder hacer encajar el funcionamiento del sistema en un argumento deductivo sólido como lo pide el modelo nomológico-deductivo.

Posteriormente, he analizado el modelo causal-mecanicista clásico de explicación y también lo he rechazado como estándar de explicación en el contexto de la inteligencia artificial explicable porque i) nos da respuestas sobre el cómo se ha llegado a una salida O, pero buscamos que nos respondan por qué se ha dado O y ii) en los sistemas con base en aprendizaje profundo la complejidad de operaciones es tan amplia, que reconstruirlo en una cadena causal inteligible no es posible.

Por último, he expuesto a las explicaciones contextualistas como el modelo de explicación que podría permitirnos tener acceso epistémico a las salidas de los sistemas opacos. Para

lograr este tipo de explicaciones, tomamos al contexto local y al conocimiento previo de la audiencia como factores irreducibles para lograr explicaciones exitosas, aquellas que logran generar un sentido de comprensión en la audiencia. He argumentado también que este tipo de explicaciones tiene como pre condición la comprensión de una teoría de trasfondo que nos permita hacer sentido de la información ofrecida en la explicación.

En el siguiente capítulo, analizaré el concepto de comprensión con base en el trabajo de Catherine Elgin (Elgin, 2004, 2007, 2017) para argumentar que, en el contexto de la xAI, "comprensión" es un concepto que necesario para lograr producir explicaciones contextualistas que nos den acceso cognitivo a las salidas de los sistemas opacos y nos permitan ocuparnos de los problemas prácticos que estos nos generan.

3.0 Comprensión como condición necesaria para generar explicaciones en la xAI

El término "comprensión" se ha utilizado en filosofía de la ciencia y epistemología en un sentido y en la filosofía no-analítica en otro. En la filosofía continental, se habla de comprensión humanista (Grimm, 2021) al referirse a que, para comprender a otros seres humanos y sus artefactos, se requiere una metodología y unas habilidades distintas a las que necesitamos para comprender fenómenos de la naturaleza. La idea general es que hay algún sentido de comprensión de otros seres humanos que solo podemos alcanzar al reconstruir sus perspectivas "desde dentro", al hacer un análisis hermenéutico. Autores como Vico (Vico, 2002) o Dilthey (Dilthey, 1991) son ejemplos de comprensión humanista.

En este texto, dejaremos a un lado la comprensión humanista para hablar solamente de comprensión desde la epistemología y la filosofía de la ciencia. En esta área de la filosofía, a su vez, hay al menos tres sentidos en que se habla de comprensión (de Regt, 2009):

> **FU[10]:** Sentimiento de comprensión; las experiencias psicológicas subjetivas que acompañan a una explicación.
>
> **UT[11]:** Comprender una teoría; ser capaz de usar la teoría.
>
> UP[12]: Comprender un fenómeno; tener una explicación apropiada de un fenómeno.

Un ejemplo de FU es el sentimiento de "eureka" cuando creemos que hemos logrado asir información que no teníamos antes. Por ejemplo, al recibir una explicación de un divulgador de la física sobre un fenómeno, podré creer que comprendo el fenómeno, aunque se me escapen detalles importantes para que un físico pueda afirmar que comprende el mismo fenómeno. La comprensión que yo adquiera será del tipo FU.

UT involucra ser capaz de manipular una teoría para hacer distintas cosas como dar explicaciones, generar argumentos, razonar con base en ella, entre otras. En este sentido, "comprensión" es una habilidad para realizar juicios que se desarrolla con la práctica, como elaboraré más adelante.

---

[10] *Feeling of understanding*.
[11] *Understanding a theory*.
[12] *Understanding a phenomenon*.

Por último, UP es meramente tener una explicación adecuada de un fenómeno. Hempel argumenta a favor de este sentido de comprensión al ser ésta derivada de la explicación (Hempel, 1965). A continuación, revisaré cuál es la relación que se da entre estos dos conceptos.

3.1 Relación entre comprensión y explicación

Al igual que con la explicación, se suele clasificar la "comprensión" en dos grandes conjuntos: comprensión objetivista y comprensión pragmática. Dependiendo de cómo nos posicionemos en el debate, la comprensión será un producto de la explicación o la explicación será un producto de la comprensión.

3.1.1 La visión objetivista de la comprensión

Como hemos visto, los y las filósofas objetivistas en relación a la explicación suelen afirmar que la explicación científica nos brinda comprensión del mundo. Es decir, basta con tener una explicación adecuada de un fenómeno para comprender tal fenómeno. A esta creencia la llamaremos la visión objetivista de la comprensión y podemos identificar a Hempel (Hempel, 1958, 1965) y a Trout (Trout, 2002, 2007)como sus mayores exponentes.

Recordando, de acuerdo a Hempel, la marca principal del conocimiento científico es su naturaleza objetiva. Las explicaciones que genera el conocimiento científico son independientes de las mentes de los científicos, por lo que los filósofos y filósofas de la ciencia han de ofrecer una visión objetivista de la misma, es decir, independiente de la mente de los sujetos que proveen y reciben explicaciones. La explicación científica debe de ser objetiva y ha de ignorar aspectos que dependan de las mentes de los sujetos, como la comprensión.

De acuerdo a Hempel, "comprensión" es un término pragmático porque su uso requiere hacer referencia a las personas involucradas en el proceso de explicar (Hempel, 1958) por lo que es relativo: una explicación E será comprensible para alguna persona p pero no para otra p'. Dada su naturaleza subjetiva, la comprensión no tiene cabida en la visión objetivista de la explicación científica.

Los objetivistas afirman que la explicación científica nos provee de comprensión científica solo en el sentido en que muestra como un fenómeno se acopla a un sistema de regularidades representadas por leyes empíricas o principios teóricos (Hempel, 1958). En Hempel los modelos nomológico-deductivos y estadístico-inductivos hacen precisamente esto, subsumen fenómenos particulares a leyes nomológicas o a regularidades del mundo. Por otro lado, Trout afirma que la comprensión científica se infiere solo si se tiene una afirmación en modo de *explanandum* que es aproximadamente verdadera; un agente con suficiente información relevante sobre la afirmación que funge como *explanandum* y la creencia de que se tiene una explicación se produce mediante un proceso fiable (Trout, 2007). La comprensión es un producto de la explicación.

De estas afirmaciones se sigue que la explicación científica es condición suficiente de la comprensión científica y la comprensión es, a lo más, condición necesaria de la explicación. Dicho de otro modo, la explicación implica la comprensión, pero la comprensión no implica a la explicación. Así, un modelo objetivista de la relación entre comprensión (C) y explicación (E) sería:

E→C

3.1.2. La visión pragmática de la comprensión

En su artículo *The Epistemic Value of Understanding*, Henk de Regt hace una crítica a la visión objetivista de la comprensión y afirma que en toda explicación hay un componente pragmático: las habilidades y los juicios de los científicos que las producen. Por lo que la comprensión no puede ser un producto de la explicación científica.

Es de Regt quien distingue entre los tres modos de entender "comprensión" que vimos en la presentación de este capítulo. Recordando, son:

FU: las experiencias psicológicas subjetivas que acompañan a una explicación.

UT: ser capaz de usar la teoría (comprensión pragmática).

UP: tener una explicación apropiada de un fenómeno.

La visión objetivista corresponde a UP, sin embargo, de Regt afirma que UT es una condición *sine qua non* se logra UP.

De Regt parte de la visión argumental de las explicaciones. Una explicación es un argumento que subsume un fenómeno en un marco teórico más general (de Regt, 2009), de acuerdo a Hempel, esto se logra al deducir el *explanandum* de leyes nomológicas y condiciones de frontera (*boundary conditions*). Por ejemplo, se puede explicar que un avión vuela al deducirlo del principio de Bernoulli y las condiciones ambientales relevantes. Sin embargo, arguye de Regt, para construir una explicación no basta con solo conocer el principio de Bernoulli y las condiciones ambientales relevantes, se requiere algo más. Ese "algo más" es la **habilidad** para construir deducciones con base en el conocimiento que se posee y no es solo un añadido, sino una condición sin la cual no se podrían construir explicaciones. En el proceso de aprender una disciplina, uno aprende a hacer **juicios** no-algorítmicos sobre cómo proceder para generar explicaciones, es decir, aprendemos a realizar razonamientos heurísticos, razonamientos plausibles sin reglas infalibles para resolver problemas (Fonseca Patrón, 2019).

Dado que las habilidades y los juicios de los científicos son necesarios para establecer y evaluar relaciones entre las teorías y los fenómenos, se sigue que las explicaciones son posibles solo si se cumplen algunas condiciones pragmáticas particulares (de Regt, 2009). Autores como Hanson, en *Patterns of Discovery* (Hanson, 1958), y Elgin, en *True Enough* (Elgin, 2004), han llegado a la misma conclusión.

A la afirmación de que se requieren factores pragmáticos para dar cuenta de explicaciones la llamaremos la visión pragmática de la comprensión. En ella, la comprensión de una teoría T (comprensión entendida como UT) es condición suficiente para generar explicaciones dentro de T. El modelo, por tanto, es el siguiente:

C→E

Qué condiciones pragmáticas particulares son las relevantes para generar explicaciones es donde los y las diversas autoras dentro de la filosofía de la ciencia difieren. Para Hanson, una explicación solo es posible si se tiene comprensión de un cuerpo coherente de información que nos permita establecer las relaciones entre el *explanans* y el *explanandum*, donde los

factores pragmáticos relevantes son los sesgos de formación: científicos que conozcan distintas teorías verán cosas distintas y generarán explicaciones distintas. Para de Regt, como hemos visto, las condiciones pragmáticas relevantes son las habilidades y el juicio de los científicos para generar explicaciones. Para Elgin, también el factor pragmático relevante es la habilidad de utilizar la teoría, sin embargo, lo que diferencia su posición de la de de Regt es que, además de poder utilizar una teoría, comprender conlleva tener acceso epistémico a los fenómenos mediante idealizaciones. Argumentaré a favor de la visión de comprensión de Elgin.

3.2.0 Idealizaciones y comprensión

En nuestras prácticas científicas utilizamos modelos e idealizaciones para obtener acceso epistémico a los fenómenos que estudiamos. Suavizar la curva al analizar datos, experimentar bajo condiciones ideales, afirmaciones *ceteris paribus*, o argumentos a *fortiori* son algunos ejemplos. Aceptamos estas técnicas, aunque sepamos que no son estrictamente verdaderas, y a partir de ellas inferimos conocimiento, explicaciones y comprensión de fenómenos (Elgin, 2004). A estas prácticas y modelos que no son verdaderos, pero nos sirven para alcanzar a comprender fenómenos Catherine Elgin los llama *felicitous falsehoods*[13].

[13] No encontré una traducción adecuada que, a la vez, mantuviese el sentido del original y no disparase malos entendidos en español. Una traducción aproximada sería "falsedades benignas" o "falsedades oportunas". Por la falta de una buena traducción, mantendré el término en el inglés original.

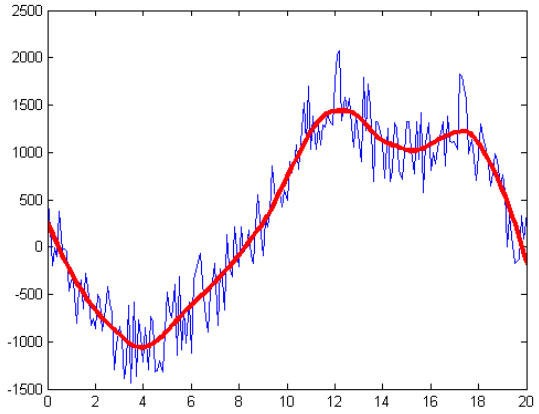

Ilustración 7: Ejemplo de felicitous falsehood, suavizar la curva. La línea roja es la que utilizamos para analizar datos, la línea azul, los datos que de hecho obtenemos del mundo.

Las *felicitous falsehoods* nos sirven, a su vez, para crear afirmaciones suficientemente verdaderas que podamos utilizar en nuestros modelos científicos para crear explicaciones. Argumentaré que el uso de *felicitous falsehoods* y afirmaciones suficientemente verdaderas nos permitirá generar modelos que sean lo suficientemente verdaderos para lograr acceso epistémico a las salidas que hasta ahora nos son opacos en los sistemas de aprendizaje profundo.

3.2.2 Suficientemente verdadero

Ahora bien, del hecho de que aceptemos afirmaciones que no consideremos estrictamente verdaderas, no se sigue que aceptemos falsedades indiscriminadamente. Aceptamos una afirmación cuando creemos que es suficientemente verdadera para algún fin determinado, cuando consideramos que su desviación de la verdad, si hay alguna, es despreciable (Elgin, 2004).

Pensemos en la constante de la aceleración de la gravedad, "g= 9.8m/s^2", será lo suficientemente verdadera con respecto a: I) para qué queramos usar la fórmula; y 2) qué tan preciso necesitamos que sea el cálculo o la explicación que derivemos de ella. Toda afirmación es lo suficientemente verdadera con respecto a la pregunta "¿suficientemente verdadera para qué?" Por lo que los objetivos son un constreñimiento a estos enunciados. Si aceptamos una afirmación o no, dependerá de a qué fin sirve su aceptación.

Las oraciones suficientemente verdaderas no tienen un significado ni un propósito aisladas, pertenecen y juegan un rol en cuerpos de discurso más amplios como argumentos, explicaciones o teorías. Al aceptar una oración suficientemente verdadera, la aceptamos en un contexto jugando un rol en un cuerpo discursivo que busca alcanzar un fin. "$g=9.8m/s^2$" será suficientemente verdadera dependiendo de si el cuerpo discursivo en el cual aparece cumple con su función cognitiva, es decir, si logra producir comprensión en el dominio que nos interesa.

Pongamos un ejemplo. Supongamos que queremos explicar en clase de física cómo funciona la caída libre con base en la mecánica clásica a un grupo de estudiantes de secundaria. Este es nuestro objetivo. El cuerpo discursivo es una explicación. Así, la oración "$g=9.8m/s^2$" es lo suficientemente verdadera en este contexto y con este fin. Si los fines o el contexto cambian, los criterios de aceptabilidad cambiarán también. Supongamos ahora que queremos calcular qué tanta oposición representará la gravedad al lanzar un satélite al espacio. Como el fin y el contexto cambian, ya no será aceptable "$g=9.8\ m/s^2$" necesitaremos un enunciado más cercano a la verdad, uno que tome en cuenta más factores como la latitud, altitud, profundidad, resistencia del aire sobre el objeto acelerado y las posibles anomalías gravitacionales del sitio para hacer un cálculo más preciso.

La comprensión, por tanto, es primariamente acerca de cuerpos de información, es holística, no es acerca de enunciados individuales (Elgin, 2007). La comprensión en los enunciados individuales se deriva de una comprensión de un cuerpo de información más grande que incluye a estos enunciados individuales. Siguiendo el ejemplo de Elgin, si comprendo que "Atenas venció a Persia en la batalla de Maratón" es porque comprendo cómo este enunciado se acopla, contribuye a, y es justificado por un cuerpo de información más grande que la incluye (Elgin, 2007). Así mismo, comprender "el sistema recomienda el tratamiento x" en el campo de la inteligencia artificial explicable implica comprender cómo este enunciado se acopla, contribuye a y es justificado por un cuerpo de información que lo incluye.

Resumiendo, la comprensión para Elgin es una habilidad de utilizar una teoría y un tipo de éxito cognitivo. Decimos que comprendemos algo cuando: 1) podemos utilizar una teoría para distintos fines y 2) tenemos acceso epistémico a un cuerpo de información.

3.3.0 El modelo de Elgin

La comprensión es una habilidad y un tipo de éxito cognitivo. Decimos que comprendemos un cuerpo de información cuando podemos captar las relaciones entre las distintas proposiciones de dicho cuerpo de información y tenemos la habilidad de utilizar esta información para nuestros fines. Que tan bien estén captadas esas relaciones entre proposiciones es cuestión de grados, por lo que la comprensión también se da en grados.

Por ejemplo, comprender el álgebra involucra más que saber todos los axiomas, los teoremas y cómo derivar nuevas fórmulas. Quien comprende álgebra debe además ser capaz de usar esa información, aplicarla o razonar con base en ella. Estas habilidades se darán en mayor o menor medida por lo que se comprenderá más o menos el álgebra.

La comprensión no es indiferente a la verdad, pero no es fáctica (Elgin, 2007). Es común utilizar oraciones suficientemente verdaderas y *felicitous falsehoods*. Las *felicitous falsehoods* son idealizaciones efectivas para las cuales no hay correspondencia en el mundo, por lo que, como descripciones, son falsas, pero son *felicitous* porque nos permiten tener acceso epistémico a cuestiones de hecho que de otro modo son difíciles o imposibles de alcanzar. Ejemplos de estas *felicitous falsehoods* son los experimentos en laboratorio, donde se crean condiciones artificiales para acceder a fenómenos que no se pueden obtener en condiciones reales.

Hay ocasiones en las que las *felicitous falsehoods* son preferibles a proposiciones verdaderas. Un caso paradigmático son las ecuaciones que se utilizan en dinámicas de flujos para describir el comportamiento de un fluido en las capas límite (*boundary layers*). Podemos derivar una ecuación parcial de segundo orden para describir exactamente cómo se comportan los fluidos en las capas límite. Sin embargo, como no es una ecuación lineal, no le podemos dar una solución analítica. Podemos dar la ecuación, pero no resolverla. Quienes trabajan en dinámica de flujos prefieren una ecuación diferencial parcial de primer orden que aproxima la verdad y admite una solución analítica (Elgin, 2004). Esta ecuación da resultados literalmente falsos, aunque lo suficientemente verdaderos para aplicarlos en tecnología, como en las alas de los aviones, además de avanzar en la comprensión de la teoría de dinámica de flujos. La ecuación que describe con verdad el comportamiento de los fluidos en estas circunstancias es irresoluble e inútil, por lo que preferimos otra suficientemente verdadera

que es útil y tiene solución. Esto no implica que no nos importe la verdad, por esto es que los enunciados suficientemente verdaderos tienen restricciones para poder ser aceptados, lo único que esto implica es que, en ocasiones, pedir la verdad solo la verdad y nada más que la verdad es pedir demasiado.

La comprensión de un cuerpo de información tiene sus bases en los hechos, responde debidamente a la evidencia y nos permite hacer inferencias no triviales, argumentos e incluso acciones con referencia al tema sobre el cual es la información de dicho cuerpo. Este tipo de comprensión entra en la categoría de UT de de Regt y es paso previo para lograr UP, que es lo que buscamos en la xAI.

3.4 La necesidad de la comprensión en inteligencia artificial explicable

En el campo de la xAI, habría que dar el paso previo de UT para lograr generar explicaciones contextualistas que nos permitan tener acceso epistémico (UP) a las salidas de los sistemas opacos y resolver algunos de los problemas prácticos que estos generan, por lo que tenemos que tomar en cuenta la "comprensión" a la hora de generar explicaciones. Aún queremos explicaciones, dado que queremos resolver problemas prácticos como "por qué este sistema reproduce sesgos raciales" si no damos explicaciones, nos quedamos en mostrar cómo o en saber cómo, pero no podemos pasar a la acción.

Hasta ahora, he argumentado que en la investigación actual en inteligencia artificial explicable hay un problema principal, se apunta por explicaciones que no responden la pregunta "¿por qué?" sino la pregunta "¿cómo?" Este problema podría deberse a la falta de una definición en la literatura de las ciencias computacionales sobre qué es una explicación y qué se requiere para ofrecerla, esto, a su vez, provoca una visión de túnel que busca dar explicaciones para comprender fenómenos, como las salidas de los sistemas opacos, sin dar el paso previo y necesario de comprender una teoría. Para resolver este problema, propongo tomar como modelo de explicación a las explicaciones contextualistas. Éstas nos permiten dar explicaciones satisfactorias al problema de la caja negra al incluir factores contextuales necesarios para explicar adecuadamente fenómenos.

3.4.1 Explicaciones contextualistas en la xAI

Cuando buscamos generar explicaciones en la xAI, hay al menos dos preguntas por hacernos, ¿qué cuenta cómo una explicación satisfactoria? Y ¿qué requerimos para lograrla? La respuesta a la primera pregunta nos permite saber qué es lo que buscamos y la respuesta a la segunda qué necesitamos cumplir para alcanzar nuestro objetivo. Entiendo por "explicación satisfactoria" un tipo de explicación que nos permita comprender el porqué de las salidas de los sistemas opacos para poder resolver los problemas prácticos que surgen a partir de los sistemas de caja negra. Para poder dar cuenta de los problemas prácticos, hemos de tomar en cuenta a quién le vamos a dar la explicación, en dónde, y con qué fin. Es decir, hemos de saber quiénes son los diferentes usuarios, en qué contexto les ofrecemos explicaciones y cuál es el objetivo de estos usuarios e incluir esta información en nuestras explicaciones.

En resumen, buscamos explicaciones que nos permitan tener acceso cognitivo a las salidas de los sistemas opacos con el fin de resolver los problemas prácticos que nos ocasionan. Para lograrlo, tomamos en consideración el conocimiento previo de quienes reciben las explicaciones, en qué contexto las ofrecemos y cuál es la finalidad del agente que recibe la explicación, es decir, ¿para qué quiere la explicación? Buscamos, por tanto, explicaciones contextualistas.

3.4.2 Factores contextuales necesarios para dar explicaciones en la xAI

Como dijimos en el capítulo 1, el objetivo del proyecto de la xAI es "permitir a los usuarios finales comprender, confiar e interactuar de forma efectiva con la generación emergente de sistemas de inteligencia artificial" (DARPA, 2016). Estos usuarios finales son al menos de tres tipos: desarrolladores, usuarios expertos en el dominio y usuarios no expertos. Por ejemplo, en un sistema opaco para hacer diagnóstico médico, nuestras partes interesadas en que el algoritmo sea transparente son los desarrolladores del algoritmo (para poder corregir errores si estos se presentan, así como para optimizar el modelo a futuro), los médicos (para poder confiar en la predicción del sistema) y los pacientes (para saber por qué se les diagnostica lo que se les diagnostica). El tipo de explicación que se ha de dar a cada uno de ellos es distinta, dado que su conocimiento previo y sus fines son distintos. Los desarrolladores querrán explicaciones sobre cómo procesa información el algoritmo, los

médicos sobre cuáles son las bases médicas para que el algoritmo genere la predicción que genera y los pacientes querrán saber por qué el diagnóstico dado es el mejor, por dar un ejemplo.

La finalidad de una explicación es lograr que una audiencia particular comprenda un fenómeno y, para esto, se requiere una teoría de trasfondo a la cual poder vincular la explicación. Para poder explicar por qué un sistema opaco ha decidido darle o no un crédito a una persona, no basta con que un sistema intérprete nos diga "dados los parámetros i, j, k, entonces se da la salida O", sino que se requiere comprender un cuerpo de información que dé cuenta de las relaciones entre i, j y k entre ellas y con la salida O. De igual modo la explicación "dados los parámetros i, j, k, entonces se da la salida O" no será una buena explicación para todos los usuarios, dado que no en todos logrará que el usuario comprenda el fenómeno. Para dar buenas explicaciones se requiere tomar en cuenta factores contextuales.

3.4.3 Comprensión de teoría en la xAI

El problema de la caja negra es análogo al problema de las capas límite en dinámica de flujos que expuse en la sección 3.3.0. En él sabemos las ecuaciones que describen con verdad cómo se comportan los fluidos en las capas límite, pero no tenemos solución analítica para ellas, por lo que nos son inútiles. En cambio, utilizamos ecuaciones que describen de forma suficientemente verdadera el comportamiento de los fluidos en las capas límite y nos permiten hacer aplicaciones tecnológicas. En el caso de la xAI, sabemos cómo funcionan las redes neuronales artificiales, cómo representan conocimiento, qué tipo de modelos generan a partir de los datos dependiendo de sus distintas arquitecturas entre muchas otras cosas. Podemos describir matemáticamente cómo es que funcionan los sistemas opacos. Sin embargo, no tenemos acceso a por qué arroja un resultado y no otro, este es el problema de la caja negra.

La descripción matemática de una red neuronal artificial es estéril a la hora de resolver problemas prácticos cómo la radicalización de las visiones políticas de las personas causada por el algoritmo de Twitter (Wolfowicz et al., 2021), el crecimiento de las brechas económicas y sociales entre la población marcada por la división entre quienes controlan los

modelos y los datos y quienes somos los datos que el modelo explota para vendernos distintos bienes en la publicidad digital (O'Neil, 2016), los sesgos de género y raza de los algoritmos utilizados para dar acceso a universidades, créditos o préstamos bancarios que impiden a mujeres y personas de color el acceso a ellos (O'Neil, 2016), entre muchos otros.

Al igual que en dinámica de flujos, el hecho de que no podamos ofrecer una respuesta verdadera, en un sentido fuerte, al problema de la caja negra, no significa que no podamos dar respuestas suficientemente verdaderas que nos permitan tener acceso cognitivo a las salidas e idear respuestas a los problemas prácticos antes mencionados pero, para lograrlo, hemos de partir de la comprensión de una teoría de fondo que nos permita decir cuáles afirmaciones son lo suficientemente verdaderas para nuestros fines en nuestra área.

Partir de comprender en lugar de partir de explicar a los sistemas de AI basados en aprendizaje de máquina nos permitiría saltear los problemas que tienen las técnicas actuales de xAI: 1) Nos permite tener una guía clara de investigación al tener un objetivo bien definido; 2) se podría dar cuenta del funcionamiento del sistema sin importar que éste sea estocástico, 3) nos permite avanzar respuestas a los problemas prácticos que nos aquejan; y 4) dado que la comprensión viene en grados, podríamos dar cuenta de los factores contextuales relevantes para ofrecer buenas explicaciones a los distintos usuarios finales al ajustar las explicaciones al nivel de conocimiento y a los intereses de los usuarios, lo cual a su vez permitiría hacer transparentes a los modelos para todos los usuarios.

Los problemas prácticos derivados del problema de la caja negra se podrían tratar si partimos de utilizar las teorías de trasfondo relevantes para el área, a saber: ciencias computacionales para describir el modelo; matemáticas para comprender los procesos estadísticos y estocásticos que se utilizan al entrenarlo; la teoría del área en que se utilice el sistema, economía, medicina, biología, entre otras. Además, requeriríamos información acerca de quienes serán los agentes que reciban las explicaciones y cuáles son sus fines. Partiendo de esto, podemos generar idealizaciones usando proposiciones suficientemente verdaderas acerca del funcionamiento o la toma de decisiones de los sistemas opacos para, con base en esto, poder generar explicaciones contextualistas que permitan a los usuarios finales comprender las salidas del sistema para satisfacer sus fines prácticos como corregir el sistema, razonar contrafácticamente o justificar decisiones. Las técnicas actuales de la xAI

no nos permiten tratar los problemas prácticos hasta abrir la caja negra y, como argumenté en el capítulo 2, los modelos de explicación que subyacen a estas técnicas no nos lo permiten. Es por esto que, si nuestro objetivo en la xAI es resolver problemas prácticos, debemos partir de generar un modelo suficientemente verdadero que nos permita crear explicaciones contextualistas utilizando como teoría de fondo la de las ciencias computacionales, las matemáticas y las del dominio en el cual se está utilizando el sistema.

Conclusiones.

A lo largo de este trabajo he defendido que es necesario tomar en cuenta factores contextuales como a quién le vamos a ofrecer explicaciones, en qué contexto y con qué fin para generar buenas explicaciones a los modelos opacos. Una buena explicación es aquella que nos dé acceso cognitivo a las salidas, es decir, que genera comprensión en los distintos usuarios y les permite tratar los problemas prácticos que derivan de los sistemas opacos. Argüí que para lograr incorporar estos factores contextuales en la explicación y ofrecer buenas explicaciones a los distintos usuarios, hemos de comprender algunas teorías de trasfondo que hagan que la explicación tenga sentido. Es decir, no podemos dar cuenta del problema de la caja negra ignorando los factores pragmáticos a la hora de ofrecer explicaciones.

He criticado la implementación del modelo nomológico-deductivo como modelo de explicación porque no se acopla a las particularidades de los sistemas con base en aprendizaje de máquina en al menos dos puntos 1) estos sistemas utilizan procesos estocásticos durante el entrenamiento, por lo que no podemos cumplir con la condición de factividad y 2) no es claro que existan leyes para poder hacer encajar el funcionamiento del sistema en un argumento deductivo sólido como lo pide el modelo nomológico-deductivo.

Posteriormente, he analizado el modelo causal-mecanicista clásico de explicación y también lo he rechazado como estándar de explicación en el contexto de la inteligencia artificial explicable porque i) nos da respuestas al cómo se ha llegado a una salida O, pero buscamos que nos respondan por qué se ha dado O y ii) en los sistemas con base en aprendizaje profundo la complejidad de operaciones es tan amplia, que reconstruirlo en una cadena causal inteligible no es posible.

Por último, he expuesto a las explicaciones contextualistas como el modelo de explicación que podría permitirnos tener acceso epistémico a las salidas de los sistemas opacos. Para lograr este tipo de explicaciones, tomamos al contexto local y al conocimiento previo de la audiencia como factores irreducibles para lograr explicaciones exitosas. He argumentado también que este tipo de explicaciones tiene como pre condición la comprensión de una teoría de trasfondo que nos permita dotar de sentido a la información ofrecida en la explicación.

En el capítulo tres, expuse los tres sentidos en que se puede entender “comprensión” en la filosofía de la ciencia: sentimiento de comprensión (FU); comprensión de una teoría (UT) y comprensión de un fenómeno (UP). UT implica UP siendo UT la capacidad de utilizar una teoría para razonar, argumentar, generar tecnología, entre otras cosas, y UP el tener una explicación apropiada de un fenómeno.

Al distinguir entre la visión objetivista y la pragmática de la comprensión, mostré cómo la visión objetivista afirma que la explicación implica comprensión. Argumenté contra esta visión. En toda explicación hay componentes pragmáticos como las habilidades y juicios de quienes las producen, por lo que la comprensión no puede ser producto de la explicación. Dado que las habilidades y los juicios son necesarios para establecer y evaluar relaciones entre las teorías y los fenómenos, se sigue que las explicaciones solo son posibles si se cumplen algunas condiciones pragmáticas particulares. Por tanto, es necesario tener UT para lograr UP, es decir, es la comprensión la que implica la explicación. Primero comprendemos y luego explicamos.

Afirmé, siguiendo a Elgin, que la comprensión es una habilidad para manipular una teoría y es un tipo de éxito cognitivo que nos permite captar relaciones entre un cuerpo de información. Para lograr comprender fenómenos, echamos mano de una teoría de trasfondo que le da sentido y, para generar esta teoría, utilizamos idealizaciones que no son necesariamente verdaderas como suavizar la curva al analizar datos o utilizar afirmaciones *ceteris paribus*. Con base en estas idealizaciones, generamos afirmaciones suficientemente verdaderas que utilizamos en nuestros modelos científicos para crear explicaciones, inferir conocimiento y hacer aplicaciones tecnológicas.

Estas afirmaciones son suficientemente verdaderas con relación a algún fin determinado. Las oraciones suficientemente verdaderas no tienen un significado ni un propósito aisladas, pertenecen y juegan un rol en cuerpos de discurso más amplios como argumentos, explicaciones o teorías. Al aceptar una oración suficientemente verdadera, la aceptamos en un contexto jugando un rol en un cuerpo discursivo que busca alcanzar un fin.

De lo anterior se sigue que la comprensión es primariamente acerca de cuerpos de información, es holística, no es acerca de enunciados individuales. La comprensión en los enunciados individuales se deriva de una comprensión de un cuerpo de información más grande que incluye a estos enunciados individuales.

Pasando al caso de la xAI, el objetivo del proyecto de la xAI es permitir a los usuarios finales comprender, confiar e interactuar de forma efectiva con la generación emergente de sistemas de inteligencia artificial. Mostré que estos usuarios finales son de al menos tres tipos: desarrolladores, usuarios expertos en el dominio y usuarios no expertos. El tipo de explicación que se ha de dar a cada uno de ellos es distinta, dado que su conocimiento previo y sus fines son distintos. La finalidad de una explicación es lograr que una audiencia particular comprenda un fenómeno y, para esto, se requiere una teoría de trasfondo a la cual poder vincular la explicación.

Ahora bien, las explicaciones las queremos para un fin: resolver los problemas prácticos derivados de los sistemas opacos. Es por esto que la comprensión sola no nos basta. Si bien es posible que cumplamos con la definición de comprensión de Elgin de modo que tengamos acceso cognitivo a fenómenos y tengamos la habilidad de hacer cosas con una teoría, no es necesario que lo hagamos. Puedo comprender el álgebra y no razonar algebraicamente, o dar explicaciones en términos algebraicos. Sin embargo, si hacemos esto en el campo de la xAI, nos quedaríamos en el nivel teórico, pero no podríamos hacer aplicaciones sin dar explicaciones a usuarios en distintos niveles de experiencia. Dado que queremos resolver problemas prácticos, es necesario que ofrezcamos explicaciones y dado que los usuarios tienen diferentes grados de experiencia y distintos intereses es que debemos dar explicaciones pragmáticas. Por lo anterior es que mi conclusión general es que, si nuestro objetivo en la xAI es resolver problemas prácticos, debemos de partir de producir un modelo suficientemente verdadero que nos permita generar explicaciones contextualistas utilizando como teoría de fondo las de las ciencias computacionales, las matemáticas y las del dominio en el cual se está utilizando el sistema. Solo así podemos tratar los problemas prácticos derivados del problema de la caja negra.

Las ventajas de hablar de comprensión en el campo de la xAI es que nos permiten esclarecer el problema de volver explicable a un sistema de aprendizaje profundo al ofrecernos una

meta de investigación clara al definir precisamente qué explicaciones queremos, qué necesitamos y cómo las podemos generar. Además, partir de comprensión para llegar a las explicaciones contextualistas, nos permite evitar los problemas que poseían los modelos de explicación objetivista y que critiqué en el capítulo 2.

Cómo implementar computacionalmente la propuesta queda fuera del objetivo de este trabajo, requeriría un trabajo interdisciplinar a futuro. Desde la filosofía, una posible propuesta es generar sistemas que modelen el funcionamiento de las cajas negras en un cuerpo coherente de proposiciones, en un modelo, en el sentido de Elgin.

Lo anterior implicaría una reforma al proyecto de interpretabilidad *post hoc*, en donde en lugar de entrenar un nuevo modelo opaco para explicar al primer modelo opaco, generaríamos un modelo que contenga una teoría del primer modelo y nos permita integrar la salida que este arroja en un cuerpo coherente de proposiciones que lo justifiquen y, además, permita modificar la información que se ofrece a los distintos usuarios con base en los fines de los mismos.

Probablemente nunca podamos abrir la caja negra, pero eso no quiere decir que no podamos hacer nada para tratar los problemas que surgen de la toma de decisiones con la ayuda de modelos de aprendizaje profundo. Esta propuesta es una posible ruta de acción.

Bibliografía:

Armstrong, D. M. (1983). *What is a Law of Nature?* Cambridge University Press. https://doi.org/10.1017/CBO9781139171700

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language Models are Few-Shot Learners*.

Bubeck, S., & Sellke, M. (2021). *A Universal Law of Robustness via Isoperimetry*.

Carabantes, M. (2020). Black-box artificial intelligence: an epistemological and critical analysis. *AI and Society*, *35*(2), 309–317. https://doi.org/10.1007/s00146-019-00888-w

Carroll, J. W. (1994). *Laws of Nature*. Cambridge University Press. https://doi.org/10.1017/CBO9780511619908

Chalmers, D. J. (1993). Connectionism and compositionality: Why Fodor and Pylyshyn were wrong. *Philosophical Psychology*, *6*(3), 305–319. https://doi.org/10.1080/09515089308573094

DARPA. (2016). *Broad Agency Announcement Explainable Artificial Intelligence (XAI)*.

de Regt, H. W. (2009). The epistemic value of understanding. *Philosophy of Science*, *76*(5), 585–597. https://doi.org/10.1086/605795

de Regt, H. W. (2019). From Explanation to Understanding: Normativity Lost? *Journal for General Philosophy of Science*, *50*(3), 327–343. https://doi.org/10.1007/s10838-019-09477-3

Dilthey, W. (1991). *Introduction to the human sciences*. Princeton University Press.

Durán, J. M. (2021). Dissecting scientific explanation in AI (sXAI): A case for medicine and healthcare. *Artificial Intelligence*, *297*. https://doi.org/10.1016/j.artint.2021.103498

Elgin, C. (2004). True Enough. *Philosophical Issues*, *14*, 113–131.

Elgin, C. (2007). Understanding and the facts. *Philosophical Studies*, *132*(1), 33–42. https://doi.org/10.1007/s11098-006-9054-z

Elgin, C. Z. (2017). *True Enough* (1a ed.). The MIT Press.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1–2), 3–71. https://doi.org/10.1016/0010-0277(88)90031-5

Fodor, J., & McLaughlin, B. P. (1990). Connectionism and the problem of systematicity: Why Smolensky's solution doesn't work. *Cognition*, *35*(2), 183–204. https://doi.org/10.1016/0010-0277(90)90014-B

Fonseca Patrón, A. L. (2019). *Cognición humana, razonamiento y racionalidad: Los retos de la investigación empírica a la visión estándar de la racionalidad*. Universidad de Guanajuato.

Fraassen, B. C. van. (1989). *Laws and Symmetry*. Oxford University PressOxford. https://doi.org/10.1093/0198248601.001.0001

Glennan, S. (2002). Rethinking Mechanistic Explanation. *Philosophy of Science*, *69*(S3), S342–S353. https://doi.org/10.1086/341857

Grimm, S. (2021). Understanding. En *The Stanford Encyclopedia of Philosophy*.

Hanson, N. R. (1958). *Patterns of Discovery: An inquiry into the conceptual foundations of science*. Cambridge University Press.

Haugeland, J. (1985). *Artificial Intelligence; The very idea*. MIT Press.

Hempel, C. (1958). Aspects of Scientific Explanation. En *Aspects of Scientific Explanation and Other Essaysin the Philosophy of Science* (pp. 331–496). The Free Press.

Hempel, C. (1965). Studies in the logic of explanation. En *Aspects of scientific explanation and others essays in the philosophy of science* (pp. 245–290). The Free 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

O'Neil, C. (2016). *Weapons of Math Destruction*. Crown Publishing Group.

Páez, A. (2019). The Pragmatic Turn in Explainable Artificial Intelligence (XAI). *Minds and Machines*, *29*(3), 441–459. https://doi.org/10.1007/s11023-019-09502-w

Railton, P. (1978). A deductive-nomological model of probabilistic explanation. *Philosophy of Science*, *45*(2), 206–226.

Russell, S., & Norvig, P. (2020). *Artificial Intelligence A Modern Approach Fourth Edition*.

Salmon, W. C. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton University Press.

Shumener, E. (2019). Laws of Nature, Explanation, and Semantic Circularity. *The British Journal for the Philosophy of Science*, *70*(3), 787–815. https://doi.org/10.1093/bjps/axx020

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences*, *11*(1), 1–23. https://doi.org/10.1017/S0140525X00052432

Trout, J. D. (2002). Scientific explanation and the sense of understanding. *Philosophy of Science*, *69*(2), 212–233. https://doi.org/10.1086/341050

Trout, J. D. (2007). The Psychology of Scientific Explanation. *Philosophy Compass*, *2*(3), 564–591. https://doi.org/10.1111/j.1747-9991.2007.00081.x

Vico, G. (2002). *The first new science*. Cambridge University Press.

Wolfowicz, M., Weisburd, D., & Hasisi, B. (2021). Examining the interactive effects of the filter bubble and the echo chamber on radicalization. *Journal of Experimental Criminology*. https://doi.org/10.1007/s11292-021-09471-0

Woodward, J., & Ross, L. (2021). Scientific Explanation. *The Stanford Encyclopedia of Philosophy*.

Zednik, C. (2021). Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence. *Philosophy and Technology*, *34*(2), 265–288. https://doi.org/10.1007/s13347-019-00382-7