UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN FILOSOFÍA DE LA CIENCIA

Dirección General de Divulgación de la Ciencia

Facultad de Ciencias

Facultad de Filosofía y Letras

Instituto de Investigaciones Filosóficas

Campo de Estudio: Filosofía de las Ciencias Cognitivas

EL PROBLEMA COMUNICATIVO DE LA INTELIGENCIA ARTIFICIAL EXPLICABLE: UN MODELO CONVERSACIONAL

TESIS

QUE PARA OPTAR POR EL GRADO DE:

MAESTRO EN FILOSOFÍA DE LA CIENCIA

Presenta:

Alberto García Hernández

Tutora:

Dra. Atocha Aliseda Llera, IIF's - UNAM

Comité Revisor:

Dra. Claudia Lorena García Aguilar, IIF's - UNAM

Dra. Fernanda Samaniego Bañuelos, FFyL - UNAM

Dra. Ana Laura Fonseca Patrón, FCSyH - UASLP

Dra. Karen González Fernández, FF - UP

CD. MX., Marzo, 2022

Para Javier y Fernando.

Ustedes se merecen un tratado sobre Hubert Dreyfus, pero esto es todo lo que puedo ofrecer.

CONTENIDO

AGRADECIMIENTOS

Desarrollé y escribí este proyecto en un periodo de reclusión que duró esencialmente un año y medio. Fui muy feliz durante esos meses: reí, lloré, bebí muchísimo, tuve múltiples epifanías y, a pesar de que la reclusión me forzó a pasar las veinticuatro horas del día con mi familia —los amo, pero son gente muy ruidosa; a veces me gusta imaginar que vivo como monje de clausura—, e incluso cuando la salud de mis ojos ha empeorado de forma exponencial, hasta el punto en el que ya tengo la impresión de que me quedaré ciego en unos veinte o treinta años, pienso que el encierro me dio la facilidad de aprender muchísimo. A cada rato de silencio en mi habitación le seguía alguna lectura que, muy probablemente y de no ser por la pandemia, jamás me habría dispuesto a tocar. Me di cuenta además de que es mucho más agradable exponer y participar cuando hay una computadora y kilómetros de distancia entre mi persona y mis interlocutores. Hablar en clase es y seguirá siendo algo que me provoca dolores de estómago y una sensación pastosa y amarga en la boca, pero, desde casa, levantar la mano y preparar exposiciones resultaron ser experiencias menos atormentadoras. Estoy muy agradecido y muy satisfecho por todo lo que aprendí, y por las oportunidades que se me presentaron gracias a la cuarentena.

También estoy agradecido, desde luego, con la Doctora Atocha Aliseda, mi asesora desde la licenciatura, quien me ha dado muchísimas oportunidades para desarrollarme académicamente, como cuando me ofreció mi primer trabajo como su asistente en 2019, o cuando aceptó ser mi tutora en el programa de Estudiantes Asociados del IIF's. Me alegra haber escrito una segunda tesis con ella, por el cuidado que tuvo para leerme y por los seminarios que organizó para que sus otros alumnos y yo intercambiáramos comentarios, ambas cosas que resultaron indispensables para que yo lograra terminar este proyecto pero, particularmente, estoy feliz de haber continuado mis estudios de posgrado con la Dra. Atocha porque al escucharla comentar mi trabajo me percato de que aspiro a ser como ella, de que quiero hablar de

los temas que me interesan con el mismo cuidado y destreza que ella lo hace al presentarnos sus investigaciones; ese deseo por ser un filósofo profesional fue, innegablemente, la principal fuerza que me llevó a estudiar esta maestría.

Otra cosa que tengo que agradecerle a la Dra. Atocha es que me haya permitido integrarme, hace ya algunos años, a su Seminario sobre Epistemología de las Ciencias de la Salud. Ese espacio fue donde conocí a Cecilia, Fernanda y Laura, tres de las amigas más cercanas que he hecho en la UNAM; ellas han sido mis interlocutoras y lectoras desde la licenciatura, y doy gracias por todas las formas en las que me han apoyado en los últimos cuatro años. Otras tres lectoras, a las que conocí hasta iniciar el posgrado pero a las que también quiero mostrarles mi agradecimiento, son la Dra. Karen González, la Dra. Ana Laura Fonseca, y la Dra. Claudia Lorena García: todas ellas, a pesar de haber interactuado muy poco conmigo debido a la pandemia, me trataron siempre de manera amable y profesional, y me satisface mucho haberlas tenido en mi comité académico.

Por último, quiero decirle gracias a las amigas y amigos que estructuraron mi red de apoyo y soporte en los últimos tres años: Arturo, Dehilario, Cristina, Biani, Edgar, y mi mejor amigo Eric. Solo por ustedes no terminé como Jack Nicholson en *El Resplandor*.

*

Doy gracias al Consejo Nacional de Ciencia y Tecnología (CONACyT) por la beca que me otorgaron desde el 2019 hasta el 2021 para poder llevar a cabo mis estudios de Maestría en Filosofía de la Ciencia, en el campo de Filosofía de las Ciencias Cognitivas.

INTRODUCCIÓN

Esta es una tesis filosófica sobre inteligencia artificial (IA). Actualmente, la IA participa en muchos procesos de toma de decisiones: en el campo de la medicina hay sistemas de IA que ayudan a los médicos a realizar diagnósticos; en algunas empresas se utiliza inteligencias artificiales al momento de decretar una contratación; muchas compañías de seguros usan programas de IA para decidir qué clientes son candidatos ideales para recibir sus servicios. En estos casos, diferentes tipos de usuarios, la mayoría de ellos sin un conocimiento alguno sobre ciencias de la computación, interactúan directamente con inteligencias artificiales. Y, en muchos de estos escenarios, las decisiones tomadas por los sistemas de AI pueden cambiar la vida de los usuarios. Es por esta razón que los usuarios necesitan entender, verificar y confiar en las decisiones en las que los programas intervienen. Desafortunadamente, muchos de estos sistemas funcionan como cajas negras, esto es, ni siquiera los desarrolladores saben cómo es que llegan de un *input* a un *output*.

Dado que los usuarios necesitan saber cómo es que los sistemas producen **X** o **Y** decisión, científicos y desarrolladores de IA han hecho un esfuerzo por crear sistemas explicables: programas de inteligencia artificial que puedan explicar por qué toman cierta decisión. Este fenómeno es similar al que se produjo en los años ochenta con la popularización de los ordenadores [Sears & Jacko, 2009]. Cuando los ordenadores empezaron a producirse de forma masiva y a dirigirse a usuarios no expertos, surgió la necesidad de mejorar la interacción entre personas y computadoras. Del mismo modo, ahora está surgiendo la necesidad de hacer que la IA sea comprensible y fiable para los usuarios que no tienen los mismos conocimientos que un computólogo o un ingeniero en computación. La línea de investigación dedicada a crear programas que puedan explicar sus propias decisiones recibe el nombre de **inteligencia artificial explicable** o, por sus siglas en inglés, **XAI** [Villone y Lungo, 2020; Barredo Arrieta, et. al., 2019]

Ahora bien, a pesar de que una de las intenciones centrales de la XAI es la de crear sistemas que el usuario pueda entender, básicamente todos los programas de inteligencia artificial explicable que existen actualmente producen explicaciones altamente técnicas, en un lenguaje especializado y que presuponen una formación matemática. Esto no quiere decir que el proyecto de la XAI en su totalidad sea un fracaso: hay inteligencias artificiales explicables exitosas, pero las explicaciones que generan no son para todos. Este es un problema al que he denominado como "el problema comunicativo de la XAI".

La pregunta que guía esta investigación es la de si existe un modelo sobre la comunicación de explicaciones del que se pueda partir para resolver el problema comunicativo de la XAI. Mi hipótesis es que sí, y que el modelo en cuestión es el propuesto por el psicólogo Denis Hilton en su artículo "Conversational Processes and Causal Explanation" [1990]. La defensa de esta hipótesis se divide en tres capítulos. En el primer capítulo presento una serie de conceptos que me resultan indispensables para delinear cuidadosamente el problema comunicativo de la XAI. Estos conceptos son "inteligencia artificial", "aprendizaje de máquinas" e "inteligencia artificial explicable".

Luego de presentar el problema comunicativo de la XAI, en el segundo capítulo expongo el modelo conversacional de Denis Hilton. Como mencioné en el párrafo anterior, Hilton es un psicólogo, específicamente un psicólogo social, sin embargo, su teoría se basa fuertemente en la filosofía del lenguaje de Paul Grice [1975]. Por lo tanto, la segunda sección de esta tesis estará dedicada a hablar, primero, de las máximas conversacionales de Grice, y después, de la manera en la que Hilton retoma las máximas griceanas para construir un modelo *descriptivo* sobre la manera en la que los humanos comunicamos nuestras explicaciones.

En el último capítulo hago una revisión exhaustiva sobre la literatura que hay disponible actualmente en el área de la XAI. Esta revisión me permite llegar a la conclusión de que, para resolver el problema comunicativo de la

XAI es necesario partir de un modelo: 1) que tome en consideración el contexto (y el estado epistémico) en el que tiene lugar el intercambio de explicaciones; 2) que proponga máximas que indican cómo reaccionar ante la situación epistémica del interlocutor; y 3) que entienda que las explicaciones se hacen en respuesta a una pregunta. Explico cómo el modelo conversacional de Hilton cumple con estos tres incisos y concluyo que al adoptarlo estaríamos resolviendo el problema comunicativo de la XAI.

Antes de proceder con el resto del escrito, necesito dejar en claro que esta tesis es puramente teórica y normativa. No pretendo —ni puedo— ofrecer una explicación de cómo es que el modelo de Hilton podría tomarse de lo teórico y aplicarse directamente en la creación de inteligencia artificiales explicables. Me limitaré a construir un argumento sobre cuáles son las condiciones necesarias que un modelo debería de cumplir para resolver el problema comunicativo de la XAI. No voy a mostrar en lo más mínimo cómo es que un teórico de la inteligencia artificial podría tomar la propuesta de Hilton y usarla para construir una inteligencia artificial explicable.

I. EL PROBLEMA COMUNICATIVO DE LA XAI: TRES CONCEPTOS

Esta es una tesis sobre una cuestión que surge desde una parte muy específica de la Inteligencia Artificial. Existen muchas discusiones filosóficas en torno a la IA y, ciertamente, en este trabajo no puedo cubrirlas todas. No hablaré, por ejemplo, de las consecuencias que podrían tener los avances más recientes en IA para el debate entre el empirismo y el racionalismo [Buckner 2018]. Asimismo, pasaré por alto las —claramente importantes— controversias respecto al uso y desarrollo de vehículos auto-conducidos y armas letales autónomas [Nyholm & Smids, 2016; Lin et al., 2017]. Ahora bien, sí me enfocaré —quizás de manera superficial— en las preocupaciones éticas y epistemológicas que emergen ante la existencia de inteligencias artificiales cuyo comportamiento refleja sesgos racistas o sexistas. En los últimos años se han desarrollado una serie de técnicas computacionales que, supuestamente, sirven para explicar el comportamiento de una IA y así justificar sus decisiones, mejorar su desempeño y, desde luego, verificar que no exhiba un sesgo. Este conjunto de técnicas computacionales conforman una línea de investigación nombrada ***Inteligencia Artificial Explicable*** o ***XAI***. Es de la XAI de donde nace la pregunta central de mi escrito, y es en este capítulo donde explicaré cómo emerge dicha pregunta y cuál es la hipótesis con la que pretendo responder.

En la XAI se busca explicar por qué una inteligencia artificial se comporta de cierta forma. Tengo que resaltar que no todas las inteligencias artificiales son entidades enigmáticas, cuyo funcionamiento supera incluso a aquellos que tienen una formación técnica altamente sofisticada. Los métodos existentes de XAI se aplican específicamente a inteligencias artificiales "opacas". Eventualmente aclararé qué es lo que se quiere decir cuando se afirma que un sistema de AI es opaco, pero por ahora solo diré que una inteligencia artificial opaca es aquella cuyo comportamiento no puede ser entendido en su totalidad, *ni siquiera por una persona educada en las ciencias*

de la computación. Argumentaré que, si las técnicas desarrolladas dentro de la XAI son exitosas, lo son solo de manera parcial. Quiero decir: tal vez es el caso que los métodos de XAI actuales sí nos permiten obtener explicaciones sobre el funcionamiento de algunos sistemas opacos, pero las explicaciones resultantes sólo pueden ser interpretadas por un grupo sumamente limitado de personas, un conjunto compuesto únicamente por computólogos e informáticos. En este contexto, la pregunta a la que me enfrento es la de si existe algún modelo en el cúal la XAI debería basarse para poder producir explicaciones que puedan ser entendidas por una audiencia mucho más amplia. Mi respuesta —o más bien, mi hipótesis— es que sí, y que el modelo en cuestión es el "modelo conversacional de la explicación" propuesto por Denis Hilton.

A continuación presento una exposición más orgánica y refinada sobre el problema que recién describí. Sin embargo, para dar cuenta de esta pregunta es necesario definir tres conceptos que no he precisado hasta ahora:

1. Inteligencia Artificial;
2. Aprendizaje de Máquinas;
3. Inteligencia Artificial Explicable.

Estos tres conceptos tienen más de un significado y, para prevenir alguna confusión, me es obligatorio explicitar el sentido que le atribuiré a cada uno. Iniciaré exponiendo cómo entenderé la frase "inteligencia artificial". Luego de describir algunos ejemplos históricos de inteligencias artificiales, pasaré a concretar el concepto de aprendizaje profundo que adoptaré. Posteriormente, traeré a cuenta un problema propio del aprendizaje profundo: el *problema de la opacidad*. Hablar de la opacidad en el aprendizaje profundo me permitirá introducir la noción de inteligencia artificial explicable. Una vez expuestos estos tres conceptos podré expresar, con mucho más cuidado, cómo surge el problema al que me quiero enfrentar a lo largo del escrito.

CONCEPTO 1: INTELIGENCIA ARTIFICIAL

El término "inteligencia artificial" o "IA" tiene, por lo menos, dos significados: 1) uno específico, que se refiere a un sistema al que se le atribuye la cualidad de ser inteligente; y 2) uno genérico, que denota un campo de investigación. El primer significado de "inteligencia artificial" puede resultar metafísicamente problemático. Por un lado, cuando hablamos de un sistema en inteligencia artificial podemos estarnos refiriendo a una entidad matemática, una estructura abstracta que —dependiendo de qué tan nominalista sea nuestra ontología— existe fuera del espacio y el tiempo (sin ubicación geográfica, sin peso, etc.). Por otro lado, la frase "inteligencia artificial" también puede referirse a un objeto físico, específicamente a una máquina o computadora, un conjunto de cables y chips de silicio donde una estructura matemática ha sido implementada. La distinción entre la entidad matemática y su implementación física es filosóficamente interesante, pero no es relevante para esta tesis. A lo largo del texto, usaré las frases "inteligencia artificial", "sistema", "algoritmo" y "programa" para referirme, de manera indiscriminada, tanto a objetos físicos como a algoritmos y programas matemáticos.

Existe una literatura filosófica sobre la inteligencia artificial de la que me gustaría distanciarme. En un trabajo influyente, John Searle [1980] distinguió entre dos proyectos de la inteligencia artificial: la "IA fuerte" y la "IA débil". La inteligencia artificial fuerte tiene como meta crear agentes que realmente puedan poseer los mismos estados mentales que tienen los humanos. En el proyecto débil el objetivo es el de diseñar máquinas con la capacidad de emular algunos aspectos del pensamiento humano. Searle asevera que el proyecto de la inteligencia artificial fuerte es imposible y, para sostener esta afirmación, recurre a un experimento mental conocido como "el cuarto chino": John Searle se encuentra encerrado dentro de un cuarto; afuera hay hablantes nativos de chino que no están enterados de la presencia de Searle. Los hablantes de chino introducen tarjetas a la habitación mediante una ranura; en dichas tarjetas hay preguntas escritas en chino. Searle no sabe

chino, pero a su alcance hay un libro: el texto, formado por una serie de reglas semánticas, le indica cómo responder a las preguntas de las tarjetas con precisión. A partir del experimento se argumenta que:

1. Las computadoras usan reglas sintácticas para manipular cadenas de símbolos sin comprender su significado.
2. De lo anterior, los estados internos de una computadora son puramente sintácticos. En contraste, los estados mentales de los humanos cuentan, además, con un carácter semántico. Dicho de otro modo, nuestros pensamientos no son mera sintaxis, sino que tienen significados.
3. La habitación del experimento posee una sintaxis —porque hay unas reglas sobre cómo operar los símbolos—; la habitación no dispone de una semántica —porque no hay nadie en ella que entienda lo que significan los símbolos. Por lo tanto, contar con una sintaxis no es una condición suficiente para tener una semántica.
4. Los programas computacionales no tienen semántica. Los programas son solo sintaxis, pero la sintaxis es insuficiente para la semántica. La mente humana cuenta con una semántica. Y luego, contra los deseos y expectativas de los adeptos al programa fuerte de la IA, ningún programa podría poseer los estados mentales de los humanos.

Luego de la publicación del texto de Searle han aparecido centenares de críticas y de defensas al experimento de la habitación china. Sin embargo, estas discusiones —así como las distinciones entre AI fuerte y débil— no serán retomadas dentro de esta tesis. Si fuese imprescindible dar una clasificación, los sistemas que me interesa abordar se catalogarían como parte del proyecto débil de la inteligencia artificial. Los programas de los que hablaré fueron creados con la intención de ayudar a los humanos en la resolución de problemas. Esto es, se diseñaron como herramientas de apoyo, instrumentos para asistir a un agente humano en una tarea específica. De ahora en adelante

daré por sentado que toda inteligencia artificial de la que hable es un producto del proyecto débil. Por ejemplo, si hablo de una IA usada en medicina, daré por sentado que su propósito no es el de sustituir a los médicos, sino el auxiliarlos en su labor.

Los primeros programas exitosos en inteligencia artificial tomaban como base a la lógica deductiva. En 1956, Allen Newell y Herbert Simon desarrollaron un programa llamado *Logic Theorist*, un sistema pretendía imitar la capacidad de resolución de problemas de un ser humano y que se considera el primer programa de inteligencia artificial. El *Logic Theorist* fue capaz de demostrar 38 de los primeros 52 teoremas de los Principia Mathematica de Whitehead y Russell, además de encontrar pruebas más cortas para algunos de ellos [McCorduck, 2004]. En 1959, Newell y Simon presentaron el *General Problem Solver*, un programa destinado a funcionar como una máquina universal de resolución de problemas [Newell & Simon, 1959].

Los programas de IA que recién mencioné funcionaban en dominios limitados y únicamente podían resolver problemas relativamente simples. Esto llevó a muchos investigadores a centrar sus esfuerzos en desarrollar sistemas que solucionaran problemas más difíciles en dominios especializados. La mayoría de estos sistemas se creó con la intención de emular la capacidad de resolución de problemas que tendría un humano experto en un área de conocimiento específica. Por esta razón estos programas se conocen como *sistemas expertos*. Los sistemas expertos tienen dos componentes básicos: una base de conocimientos y un motor de inferencia. La información de la base de conocimientos se obtiene entrevistando a personas expertas en el área en cuestión. El entrevistador organiza la información obtenida de los expertos en una colección de reglas, típicamente estructuradas como condicionales de la forma "si α – entonces ß". El motor de inferencia permite al sistema experto hacer deducciones con las reglas de la base de conocimientos. Por ejemplo, si la base de conocimientos contiene reglas "si X entonces Y" y "si Y entonces Z", el motor de inferencia puede deducir "si X

entonces Z". El sistema experto podría entonces preguntar a su usuario "¿es X verdadero en la situación que estamos considerando?" y, si la respuesta es afirmativa, el sistema procederá a inferir Z.

Algunos sistemas expertos exitosos son: DENDRAL [Lindsay et al., 1980], un sistema para analizar espectrogramas de masas en química; XCON [McDermott, 1982], un sistema para configurar computadoras; y ACRONYM [Brooks, 1981], un sistema de apoyo visual. MYCIN, uno de los sistemas expertos más estudiados, se desarrolló en la Universidad de Stanford, como la tesis doctoral de Edward Shortliffe [1975]. Este sistema fue diseñado para recomendar diagnósticos y terapias a pacientes con enfermedades sanguíneas infecciosas. En *Expert Systems: Design and Development*, John Durkin [1995, pp. 131–62] dedica un capítulo completo a MYCIN y proporciona algunas ideas muy interesantes sobre los antecedentes, métodos y desempeño del sistema. Durkin señala que en ese momento —la década de 1970— había un uso indebido y excesivo de antibióticos dentro de la práctica médica, y que "el 66% de las terapias seleccionadas por los médicos eran desaconsejables y, de éstas, más del 62% utilizaba combinaciones inapropiadas de antibióticos" (p.132). Además, había una escasez de conocimientos especializados en el ámbito de las enfermedades sanguíneas. La creación de un programa como MYCIN, en este contexto, resultó ser una gran idea. Según Durkin, [pp. 134-140], los siguientes son algunos rasgos que caracterizan a MYCIN y que lo hacen resaltar sobre otros sistemas expertos:

- Recuerda las sesiones que ha tenido con sus pacientes, como lo haría cualquier experto humano en el área de la medicina.
- Incorpora meta-reglas —es decir, reglas a cerca de otras reglas—: estas meta-reglas le permitían a MYCIN saber cuándo romper ciertas reglas para casos o situaciones especiales.
- Se adapta al usuario: es fácil de usar y transparente para el usuario/médico.
- Interactúa usando lenguaje natural.

- Proporciona explicaciones: MYCIN puede explicar CÓMO y POR QUÉ ha llegado a una determinada conclusión.
- Puede ofrecer recomendaciones alternativas: MYCIN trata de ofrecer alternativas para que el médico elija por su cuenta. De este modo, el programa coopera y auxilia al médico, en lugar de dirigir sus decisiones; esto evita que el médico sienta que el programa se le está imponiendo.

Ahora bien, aunque programas como MYCIN tuvieron bastante éxito durante su momento, los sistemas expertos tienen una limitación importante: carecen de "sentido común". No saben para qué sirven, ni cuáles son los límites de su aplicabilidad, ni cómo encajan sus recomendaciones en un contexto más amplio. Si se le dijera a MYCIN que un paciente que ha recibido una herida de bala se está desangrando, el programa intentaría diagnosticar una causa bacteriana para los síntomas del paciente [Durkin, op. cit., p. 140].

En 1984, Douglas Lenat, un informático afincado en Texas, empezó a trabajar en un sistema llamado CYC. El objetivo de Lenat era construir una base de conocimientos que contuviera un porcentaje significativo de los conocimientos que conforman el sentido común de los seres humanos. La esperanza era que el proyecto CYC culminara en una base de conocimientos que sirviera de apoyo para futuras generaciones de sistemas expertos. Tras seis años de desarrollo se introdujeron más de un millón de afirmaciones enciclopédicas en la base de conocimientos de CYC. Los creadores de CYC también introdujeron varias afirmaciones sobre hechos que consideraríamos evidentes. Sin embargo, incluso con una base de conocimientos tan enorme, se demostró finalmente que CYC carecía de la capacidad de entender ciertos conceptos. Por ejemplo, CYC no entendió una historia sobre una persona llamada Fred que se afeitaba la cara [Linde, 1992]. Su motor de inferencia detectó una incoherencia en la historia: sabía que las personas no tienen partes eléctricas, pero como Fred sostenía una maquinilla de afeitar eléctrica, creía que la entidad "Fred" contenía partes eléctricas.

Las dificultades a las que se enfrentan programas como CYC sugieren que los sistemas de IA necesitan la capacidad de entender conceptos. Crear un sistema con capacidad de comprensión es, por supuesto, todo un reto, sobre todo porque no sabemos qué tipo de reglas deben implementarse en la base de conocimientos de un programa para que podamos decir que es capaz de entender los conceptos que se le suministran. Hay otros problemas que entran en esta categoría: no sabemos cómo escribir las reglas que necesitará un programa para reconocer objetos o comprender el habla. Esto llevó a algunos informáticos a proponer que, en lugar de crear programas que utilicen reglas y motores de inferencia, deberíamos intentar desarrollar programas con la capacidad de adquirir su propio conocimiento. Esta capacidad se conoce como aprendizaje de máquinas —*machine learning*—. Este es el segundo concepto del que hablaré en este capítulo.

CONCEPTO 2: APRENDIZAJE DE MÁQUINAS.

Al igual que "inteligencia artificial", el término "aprendizaje de máquinas" es ambiguo y tiene más de un significado. En este escrito voy a definir el aprendizaje de máquinas como la parte de la inteligencia artificial en la que los investigadores desarrollan algoritmos que las computadoras pueden utilizar para analizar grandes cantidades de datos y luego utilizar esa experiencia para resolver problemas. En esencia, el objetivo del aprendizaje de máquinas es enseñar a las computadoras a resolver problemas mediante instancias.

Uno de los grandes retos de los modelos tradicionales de aprendizaje automático es un proceso llamado *extracción de características*. En concreto, el programador tiene que decirle a la computadora qué tipo de cosas debe buscar en un conjunto de datos para tomar una decisión. Proveer al algoritmo con datos sin especificar las características relevantes rara vez funciona, por lo que la extracción de características es una parte crítica del aprendizaje de máquinas tradicional. Esto supone una enorme carga para el programador. El *aprendizaje profundo* es una forma de aprendizaje de máquinas con la que

podemos sortear los retos de la extracción de características. Los algoritmos de aprendizaje profundo son capaces de aprender a centrarse en las características correctas por sí mismos, requiriendo poca orientación por parte del programador.

Los modelos utilizados en el aprendizaje profundo están muy inspirados en la estructura neuronal humana: los científicos crean una red de neuronas artificiales organizadas en capas, con cada capa interconectada a la siguiente. A la red neuronal se le entrena proporcionándole datos como, por ejemplo, imágenes de objetos. La computadora toma los datos e intenta adivinar qué es cada objeto. Las conjeturas iniciales tienden a fallar, pero, a medida que los programadores proporcionan información, la red ajusta la forma en que sus neuronas están conectadas hasta que produce conjeturas muy precisas.

Merece la pena destacar las diferencias entre programas como MYCIN y los algoritmos de aprendizaje profundo contemporáneos porque, al igual que otras formas de IA que se han descrito en este texto, el aprendizaje profundo no está exento de problemas. Hay dos rasgos notorios que caracterizaban a MYCIN: 1) era capaz de explicar al programador por qué y cómo llegaba a una determinada conclusión; y 2) interactuaba con el médico utilizando un lenguaje natural. Los algoritmos de aprendizaje profundo carecen de estos dos rasgos. Debido a la estructura de sus redes neuronales, los algoritmos de aprendizaje profundo suelen ser descritos como cajas negras: podemos observar los *inputs* y los *outputs*, pero no sabemos con claridad qué ocurre dentro de la red neuronal. En otras palabras, conocemos los datos que recibe el algoritmo, pero no sabemos por qué o cómo llegó a una determinada conclusión. Al igual que las neuronas no artificiales de nuestros cerebros, las redes neuronales artificiales funcionan de formas que no podemos comprender del todo. Esta cualidad suele conocerse como opacidad. Cuando un algoritmo es opaco, su comportamiento no puede ser descifrado ni por sus usuarios ni por los programadores que lo desarrollaron.

Hay varias razones por las que la opacidad en el aprendizaje profundo es problemática. Si no entendemos por qué un algoritmo genera ciertos resultados, puede ser difícil saber si la conclusión a la que está llegando es errónea o no [Tutt, 2016]. Además, si no sabemos cómo funciona cierto algoritmo, no podemos regularlo para garantizar que no codifique comportamientos discriminatorios. Un artículo publicado en 2019 describe el sesgo racial en un algoritmo utilizado por el sistema sanitario de los Estados Unidos [Obermeyer, et al.]. El algoritmo fue diseñado para decidir qué pacientes deben participar en programas de "atención de alto riesgo". Los programas de atención de alto riesgo buscan mejorar la atención de los pacientes con necesidades sanitarias complejas. El artículo presenta pruebas de cómo el algoritmo reproducia un sesgo racial: le daba preferencia a los pacientes blancos sobre los negros como participantes en los programas de atención de alto riesgo, incluso cuando los pacientes negros que intentaban recibir el tratamiento estaban considerablemente más enfermos que los blancos. El algoritmo, por supuesto, no fue creado con esta intención y, como los procesos internos del algoritmo eran opacos, los programadores que lo desarrollaron no se dieron cuenta —al menos no inmediatamente— de que reproducía un sesgo.

Estos problemas prácticos no son los únicos derivados de la opacidad del aprendizaje profundo. Algunos filósofos han argumentado que debemos ser críticos con el uso de ciertas inteligencias artificiales en la investigación científica. Una de las muchas razones que motivan esta actitud crítica es precisamente la opacidad que caracteriza a algunos algoritmos: si la opacidad no nos permite identificar las relaciones causales que existen entre las entradas y las salidas, entonces los algoritmos opacos no nos sirven para producir explicaciones causales sobre del mundo [Ratti y López Rubio, 2018].

El problema de la opacidad cobra especial importancia dentro de áreas como la medicina porque los médicos tienen la responsabilidad ética y epistémica de justificar las decisiones que toman y de fundamentar los diagnósticos que les presentan a sus pacientes. En el contexto clínico, apelar

a la autoridad o a la experiencia no es suficiente. Si no entendemos cómo funcionan los algoritmos de diagnóstico, no podemos evaluar sus decisiones ni afirmar que están justificadas. Thomas Ploug y Søren Holm señalan que los pacientes deberían poder ejercer el derecho a rechazar los diagnósticos realizados por inteligencias artificiales opacas [Ploug y Holm, 2019]. Asimismo, Christian Bjerring & Jacob Busch sostienen que la "medicina de cajas negras" o el uso de sistemas de diagnóstico opacos impide que el profesional y sus pacientes compartan información y tomen decisiones juntos [Bjerring & Busch, 2020].

En respuesta a los problemas producidos por la opacidad ha surgido una línea de investigación conocida como *inteligencia artificial explicable*. Este es el tercer y último concepto que abordaré en este capítulo.

CONCEPTO 3: INTELIGENCIA ARTIFICIAL EXPLICABLE.

En 2020, una revisión sistemática de la literatura sobre aprendizaje profundo mostró que la investigación acerca de la opacidad ha experimentado un crecimiento significativo en los últimos años [Vilone & Longo, 2020]. En concreto, muchos trabajos sobre aprendizaje profundo se han centrado en crear algoritmos que puedan producir explicaciones de su propio comportamiento. El trabajo que surge de cualquiera de esta línea de investigación se conoce como inteligencia artificial explicable o XAI. La inteligencia artificial explicable se ha investigado en muchos sectores, como los vehículos autónomos [Shen, et. al., 2020], la visión artificial [Jair Escalante, et. al., 2018] y, por supuesto, el diagnóstico médico [Weng, et. al., 2017; Holzinger, et. al., 2017].

Al examinar el trabajo de los científicos que tratan de hacer explicable el aprendizaje profundo, uno no puede evitar preguntarse: ¿qué tipo de explicaciones podemos esperar de la XAI? Después de hacer una revisión exhaustiva de la literatura disponible en XAI, Tim Miller, un informático de la Universidad de Melbourne, afirma que es justo decir que la mayoría de los trabajos en inteligencia artificial explicable utilizan solo la intuición de los

investigadores sobre lo que constituye una "buena" explicación [Miller, et. al., 2017]. El tipo de explicaciones que puede ofrecer la XAI son descripciones altamente técnicas de los procesos que sigue un algoritmo para pasar de un *input* a un *output*. Aunque sin duda pueden ser interpretadas por programadores e informáticos, estas explicaciones no son fáciles de entender para el usuario promedio.

Hay, al menos, cuatro grupos de personas que se ven afectadas por la opacidad del aprendizaje profundo:

1. Los desarrolladores de algoritmos de aprendizaje profundo.
2. Organismos y entidades encargados de regular los programas de IA.
3. Usuarios de los algoritmos —agentes de seguros, médicos, etc.
4. Individuos afectados por la decisión tomada por los algoritmos —pacientes, personas que quieren adquirir un seguro, etc.

Las explicaciones altamente técnicas que existen actualmente en la XAI pueden ser comprendidas por los miembros de los primeros dos grupos. Este no es el caso de los usuarios y las personas afectadas por la decisión de los algoritmos. Los médicos, los pacientes y otras personas relacionadas con los programas de aprendizaje profundo no tienen —necesariamente— la formación técnica requerida para entender las explicaciones que puede ofrecer la XAI. Esto es lo que denomino como **el problema comunicativo de la XAI**.

Alguien podría argumentar que no es necesario que estas personas entiendan las explicaciones de la inteligencia artificial explicable porque, después de todo, quienes corrigen los sesgos y errores de los algoritmos de aprendizaje profundo son los programadores, no los usuarios. Sin embargo, este argumento ignora que si bien los usuarios no van a reparar las fallas del algoritmo, aún así necesitan saber cómo es que el programa funciona para decidir si confían en su capacidad para tomar decisiones. Si las personas que se dedican a desarrollar inteligencias artificiales transparentes de verdad quieren atender a toda su audiencia, entonces deberían hacer un esfuerzo por

crear sistemas que ofrezcan explicaciones que puedan ser interpretadas sin la necesidad de una formación especializada.

En lo que queda de estas tesis presentaré un modelo sobre cómo los humanos comunicamos las explicaciones que hacemos acerca de nuestro comportamiento o el comportamiento de otros agentes. Argumentaré que si fuese el caso que la inteligencia artificial explicable comenzase a basarse en este modelo, entonces podrían producirse algoritmos capaces de comunicarse con una audiencia que no se componga únicamente de programadores y otros expertos en el área. Dicho de otro modo, argumentaré que adoptar el modelo que tengo en mente servirá para resolver el problema comunicativo de la XAI.

2. UN MODELO CONVERSACIONAL DE LA EXPLICACIÓN

Como mencioné en el capítulo anterior, esta parte de mi tesis está dedicada a exponer una teoría que pretende capturar la manera en la que los humanos comunicamos las explicaciones que hacemos acerca de nuestro comportamiento o acerca del comportamiento de otros agentes. La teoría que tengo en mente es conocida como el "modelo conversacional de la explicación" y es desarrollada por Denis Hilton [1990]. De nuevo, mi hipótesis es que incorporar tal modelo en la XAI permitiría desarrollar sistemas que sean capaces de comunicarse con una audiencia que no se componga únicamente de programadores y otros expertos en el área.

En lo que sigue voy a exponer las características principales del modelo de Hilton. Esta exposición está dividida en dos apartados:

1. La propuesta de Hilton se basa fuertemente en la teoría de las máximas conversacionales que Paul Grice propuso en 1975. La primera sección de este capítulo se titula "Grice y las Máximas Conversacionales", y en ella expongo algunos rasgos principales de la teoría de las máximas griceanas, así como algunos estudios empíricos que sirven para respaldarla.
2. La segunda sección se llama "El Modelo Conversacional de la Explicación" y está dividida en tres partes. Cada parte está enfocada en algún aspecto importante del modelo de Hilton: 1) el concepto de explicación que adopta Hilton; 2) su supuesto de que las explicaciones son actos del habla; y 3) su tesis de que el intercambio de explicaciones se adhiere a las máximas griceanas.

1. GRICE Y LAS MÁXIMAS CONVERSACIONALES

En *Logic and Conversation* [1975], Paul Grice presenta una teoría que describe cómo las personas logran una comunicación conversacional exitosa. Grice observa que los enunciados de una conversación no se producen de forma

aislada, sino que están orientados hacía algún objetivo: compartir información, crear un vínculo entre los hablantes, etc. Grice señala que para que el intercambio de enunciados conversacionales sea efectivo (es decir, para que se puedan alcanzar los objetivos de los agentes que están conversando) los participantes de la conversación se adhieren a un principio denominado ***el principio de cooperación***: "Haga usted su contribución a la conversación tal y como lo exige, en el estadio en que tenga lugar, el propósito o la dirección del intercambio que usted sostenga." [1975, p. 45]. Los hablantes obedecen una serie de máximas y reglas para hacer cumplir este principio de cooperación. De acuerdo a Grice, estas máximas son:

1. **Máxima de cantidad**. De acuerdo a esta máxima, al tener una conversación, el hablante debe de proporcionarle a su interlocutor tanta información como sea suficiente y necesaria para cumplir con los propósitos del intercambio. La máxima de cantidad se resume en las siguientes reglas: (a) haga su contribución tan informativa como se requiera; y (b) no la haga más informativa de lo que se requiere.
2. **Máxima de calidad**. Los hablantes deben asegurarse de que la información que comuniquen es verdadera. Esta máxima contiene dos reglas: (a) no diga cosas que crea que son falsas; y (b) no diga cosas para las que no tenga suficientes pruebas.
3. **Máxima de relación**[1]. La regla descrita por Grice en relación a esta máxima es: (a) sea relevante. Lo que esto quiere decir es que, para lograr sus objetivos, los hablantes deben de abstenerse de proporcionar información que no esté relacionada con la conversación.

4. **Máxima de manera**. Quienes participan en una conversación deben de comunicarse con claridad. Esto se condensa en tres reglas: (a) evite usar términos oscuros o ambiguos; (b) sea breve (evitar la prolijidad innecesaria); y (c) exprese sus ideas en orden.

[1] En algunas traducciones de Lógica y Conversación la máxima de relación se llama "máxima de relevancia" [Miranda Ubilla y Guzmán Munita, 2012; Garachana Camarero, 2014]

Grice [1975] asevera que para sostener una conversación hay que obedecer estas máximas, y que las personas aprenden dichas máximas a lo largo de su vida. Ahora bien, es posible dejar de cumplir algunas máximas sin dejar de ser cooperativos, ya sea para no romper otra de las otras máximas o para lograr algún objetivo específico, como implicar algo sin decirlo. La ironía y las metáforas, por ejemplo, son formas en las que los hablantes rompen algunas de las máximas conversacionales. Considérese el siguiente caso: P y Q han sido amigos cercanos por un largo tiempo. Un día, Q traiciona a P al revelar uno de sus secretos. Posteriormente, Q tiene una conversación donde él y sus interlocutores son conscientes de la traición de P. En cierto punto, Q comenta: "P es una persona muy confiable". En este caso Q está violando la máxima de calidad porque está afirmando algo que cree que es falso (P no es una persona confiable), pero lo hace con el objetivo específico de ser sarcástico, por lo que, según Grice [1975, p. 53], Q no deja de ser cooperativo con respecto a la conversación.

A continuación hay otro ejemplo de cómo uno de los agentes viola las máximas para implicar algo durante un intercambio conversacional:

> En una reunión de buen tono, A dice "la Sra. X es una vieja bruja". Por un momento el silencio puede oírse, y entonces B dice "Ha hecho un magnífico tiempo este verano, ¿verdad?". [Grice, Op. Cit., 54]

B está violando la máxima de relación porque su comentario no es relevante dentro de la conversación ni está conectado con la afirmación de A. Sin embargo, Grice dice que la intervención de B no deja de ser cooperativa porque su intención al romper la máxima es la de implicar que la observación hecha por A no es cosa que haya que discutirse, y quizás, que A ha cometido un desliz social.

Es importante enfatizar que si bien el principio de cooperación y las máximas conversacionales están formuladas de manera prescriptiva, la teoría de Grice pretende ser descriptiva. En otras palabras, las máximas griceanas no fueron formuladas como una guía para que el hablante logre que su

conversación sea exitosa, sino que se diseñaron como un modelo que pretende reflejar la manera en la que funciona el intercambio conversacional exitoso entre agentes. En este sentido, son especialmente relevantes dentro de esta tesis las investigaciones empíricas que apoyan estas máximas griceanas. Una de estas investigaciones es la de Ben R. Slugoski quien, junto a sus coautores [Slugoski, et. al., 1993], realizó el siguiente experimento: un grupo de participantes recibió información a través de un informe policial sobre un individuo llamado George, quien había sido acusado de agresión tras una pelea escolar. Esta información contenía datos sobre el propio George y sobre las circunstancias de la pelea. A continuación se formaron tres equipos de dos personas. Cada equipo estaba formado por: A) un participante que ya había leído el informe policiaco acerca de George; B) un interlocutor. A cada participante se le dijo que el interlocutor con el que acaba de ser reunido tenía cierta información específica:

1. Al participante 1 se le dijo que su interlocutor tenía información sobre George, pero no sobre la pelea.
2. Al participante 2 se le dijo que su interlocutor tenía información sobre la pelea, pero no sobre George.

3. Al participante 3 se le dijo que su interlocutor no tenía información sobre George y tampoco tenía información sobre la pelea.

Luego de compartirles estos datos se les pidió a los participantes que le explicaran a sus compañeros por qué George había iniciado la pelea. Los resultados mostraron que: 1) los participantes proporcionaron explicaciones que se adaptaban a lo que creían que su interlocutor ya sabía; 2) los participantes cambiaban sus explicaciones del mismo evento cuando se presentaban a interlocutores con diferentes conocimientos previos.

Un segundo estudio que sirve para apoyar la teoría de Grice es el de Tetlock y Boettger [1989]. En tal estudio se realizaron una serie de experimentos controlados en donde un grupo de participantes recibió

información sobre un personaje llamado David. Se le pidió a los participantes que hicieran una serie de predicciones acerca del futuro de David; por ejemplo, cuál sería su promedio escolar. Se formaron dos grupos de control y dos grupos de prueba. A los miembros de uno de los grupos de control se les dijo que David pasaba 3 horas estudiando a la semana (grupo C3), mientras al otro grupo de control se les informó que David estudiaba 31 horas semanalmente (grupo C31). A los grupos de prueba (grupos T3 y T31) se les dio esta misma información y además se les proporcionó una serie de datos irrelevantes sobre David. Los resultados mostraron que los miembros del grupo T3 predijeron un promedio escolar más alto que los del grupo C3, mientras que el grupo T31 predijo un promedio más bajo que el grupo C31. Tetlock y Boettger argumentaron que esto se debe a que los participantes asumieron que la información irrelevante que se les dio en realidad sí era relevante para realizar sus predicciones. Los resultados de esta investigación están relacionados con la propuesta de las máximas griceanas porque muestran que violar la máxima de cantidad genera cambios importantes en las inferencias que nuestros interlocutores hacen con la información que les brindamos.

2. EL MODELO CONVERSACIONAL DE LA EXPLICACIÓN

2.1. EL CONCEPTO DE EXPLICACIÓN DE HILTON

Cuando Hilton habla de explicación lo hace desde una concepción que, al menos para la filosofía, podría parecer extraña. Esto sucede porque, a pesar de basarse en la teoría de Grice, el concepto de explicación que usa Hilton no viene de la filosofía, sino de la psicología social. Tanto en filosofía como en psicología se suele afirmar que una explicación es una respuesta a una pregunta de la forma "¿por qué P?". Sin embargo, las preguntas de la forma "¿por qué P?" que le interesan a la filosofía suelen ser distintas a las que le interesan a la psicología. Considérense los siguientes ejemplos:

A. ¿Por qué Alberto se quedó dormido?

B. ¿Por qué los seres humanos necesitamos dormir?

La psicología social suele tratar con las explicaciones que se hacen ante preguntas como (A). En contraste, la filosofía típicamente se enfoca en explicaciones que surgen en respuesta a preguntas como (B). Una diferencia importante entre ambas preguntas es que la segunda es mucho más general que la primera. La pregunta (A) se refiere a un evento aislado, mientras que la pregunta (B) trata con una aparente regularidad que se presenta entre todos los seres humanos. A primera vista parece que el concepto "explicación" en psicología social se refiere a las respuestas que damos a preguntas de la forma "¿por qué P?", donde P es un *evento particular*. Sin embargo, hay preguntas sobre eventos particulares cuya respuesta es una explicación que no es de interés para la psicología, pero sí para la filosofía. Por ejemplo:

> Se colocan dos kilogramos de cobre a 60 °C en tres litros de agua a 20 °C. Después de un tiempo, el agua y el cobre alcanzan un equilibrio térmico de 22.5 °C. **¿Por qué el equilibrio térmico es de 22.5 grados centígrados?** Dado que los calores específicos del agua y el cobre son 1 y 0.1 respectivamente, y dado que la conservación de energía requiere que la cantidad total de calor no aumente ni disminuya, la pérdida de calor del cobre, es decir, 0.1 x 2 x (60-T), debe ser igual a la ganancia de calor del agua, es decir, 1 x 3 x (T-20), donde T es la temperatura de equilibrio final. Y esto produce 22.5 grados centígrados como valor de T. [Ruben, 1992, p. 3]

La explicación anterior es acerca de un suceso individual, y podría ser relevante para un filósofo de la ciencia, pero no para un psicólogo social. Esto es así porque a la psicología le interesan las preguntas que se puedan hacer sobre el *comportamiento* de uno o varios sujetos. Por consiguiente, en el concepto psicológico que usa Hilton, las explicaciones son respuestas a preguntas de la forma "¿por qué P?", donde P designa *las decisiones o comportamientos de un agente*.

Algunos ejemplos de preguntas que hacemos sobre el comportamiento de otros agentes son los siguientes: ¿por qué me sonreía el hombre de la cafetería?; ¿por qué mi amigo de toda la vida ya no me llama?; ¿por qué los policías golpearon al sospechoso? Las respuestas a este tipo de preguntas son muy distintas a las explicaciones que se suelen ver en la ciencia. Cuando queremos dar cuenta del comportamiento de un agente lo primero que hacemos es determinar si sus acciones son intencionales [Malle, 2004, pp. 119]. Luego de decidir si un acto es intencional o no-intencional, procedemos a hacer una ***atribución***. Si un comportamiento no es intencional, le atribuimos una causa. Si un comportamiento es intencional, le atribuimos un motivo o una razón. Entonces, las explicaciones del comportamiento tienen la forma de una atribuciones. Este es el concepto de explicación que se usa en la psicología, que difiere del concepto de explicación usado en filosofía de la ciencia, y que Hilton retoma para su modelo.

2.2. LAS EXPLICACIONES COMO ACTOS COMUNICATIVOS

En 1980, Bas van Fraassen propuso un modelo de explicación científica en donde se enfatiza el carácter pragmático que tienen las explicaciones [van Fraassen, 1980]. Teorías anteriores prestaban poca o nula atención a los rasgos pragmáticos de la explicación. Por ejemplo, en el modelo hempeliano el carácter esencial de la explicación es puramente semántico: una explicación es un argumento válido y la validez de un argumento es independiente de factores pragmáticos. Sin embargo, van Fraassen argumentó que cuando alguien pregunta "¿por qué P?", se está contrastando implícitamente el estado expresado por P con un conjunto alternativo de estados. Tal es el caso de la pregunta "¿por qué el cielo es azul?", que puede entenderse como "¿por qué el cielo es azul y no de otros colores?". Además, la teoría de van Fraassen sugiere que el contexto en el que se emite una pregunta determina los tipos de respuestas que son relevantes para responderla. Si uno se preguntara por qué los primates tienen pulgares oponibles, el contexto determinaría si la respuesta se hará desde la biología evolutiva o desde la biología del desarrollo.

Menciono la teoría de van Fraassen porque, a pesar de que ya argumenté que el concepto de explicación en filosofía es distinto al de la psicología, Bertram Malle [2014, pp. 153-155] observa que la llegada de este énfasis en la pragmática de las explicaciones tuvo un impacto en el trabajo de psicólogos sociales como Denis Hilton. Hay dos elementos importantes que modelos como el de Hilton retoman del giro pragmático iniciado por van Fraassen. El primero de estos elementos es el énfasis en la idea de que las explicaciones del comportamiento son actos de habla. Típicamente construimos explicaciones cuando queremos darle sentido a las acciones de quienes nos rodean. Si bien es verdad que estas atribuciones suelen ocurrir de manera interna (es decir, en nuestra mente, sin que las comuniquemos), las explicaciones de la conducta también pueden ser acciones verbales interpersonales. Cuando se les pregunta por qué hicieron algo, las personas explican verbalmente su comportamiento, ya sea para justificar sus actos o para hacer que su interlocutor conozca los motivos detrás de sus acciones. En consecuencia, las explicaciones de la conducta existen como procesos mentales y como actos comunicativos.

El segundo elemento que propuestas como las de Hilton retoman del nuevo enfoque pragmático es que, al presentar una explicación, un hablante considera el contexto y la situación epistémica de su interlocutor. Esto quiere decir que una persona necesita tomar en cuenta que su audiencia tiene ciertos conocimientos y presunciones. Si la audiencia le hace cierta pregunta, el hablante responderá con una explicación que encaje con aquello que sus interlocutores ya saben o dan por sentado. El siguiente es un ejemplo de cómo los agentes producen explicaciones que encajan con el conocimiento de su interlocutor:

> El **Hablante X** asume que Alberto es pobre y no tiene dinero; el **Hablante Y** sabe cuáles son los supuestos del **Hablante X** respecto a la situación económica de Alberto.
>
> **Hablante X**: ¿Por qué Alberto compró un Mercedes?

Hablante Y: Porque repentinamente heredó una gran cantidad de dinero.

En este caso el **Hablante Y** podría haber respondido que Alberto compró un Mercedes porque piensa que es un buen auto. Esta respuesta bien podría ser verdadera, pero el hablante está tomando en cuenta que su interlocutor está formulando su pregunta desde su suposición de que Alberto no tiene dinero. Por lo tanto, las explicaciones se hacen "a la medida", a partir de cierto contexto y en consideración del trasfondo con el que cuenta una audiencia.

Hasta ahora he adjudicado tres elementos al modelo de Hilton. En primer lugar, Hilton supone que una explicación es atribuirle una causa o una razón a un comportamiento. En segundo lugar, el modelo afirma que las explicaciones existen al menos en dos niveles: como un proceso interno y como un acto del habla. En tercer lugar, la propuesta de Hilton asegura que, en tanto actos de habla, las explicaciones se hacen a partir de una pregunta específica y considerando el conocimiento previo de quien formula tal pregunta.

2.3. LAS MÁXIMAS GRICEANAS EN LA COMUNICACIÓN DE EXPLICACIONES

Como ya lo mencioné, las explicaciones pueden existir como actos del habla. Hilton agrega, además, que el intercambio de explicaciones entre un agente y un interrogador se presenta en la forma de una conversación y, como tal, está constreñido por las máximas conversacionales propuestas por Grice. Al igual que cualquier otro tipo de conversación, el intercambio de explicaciones no ocurre de manera aislada. Las explicaciones surgen en respuesta a una pregunta concreta, lo que significa que quien explica un comportamiento tiene un objetivo: responder a su interrogador. Con el fin de alcanzar este objetivo, el hablante tiene que iniciar considerando qué es lo que su interrogador encuentra desconcertante o anormal. Posteriormente el hablante producirá una explicación que cumpla con algunas de las siguientes características:

1. La explicación proporcionará toda la información que sea necesaria y suficiente para responder a la pregunta (máxima de cantidad).
2. La explicación estará genuinamente relacionada con la pregunta de su interrogador (máxima de relación).
3. La explicación será breve, ordenada, y consciente de aquellos términos y conceptos que el interlocutor maneja y de aquellos que ignora (máxima de manera).

4. La explicación no incluirá información falsa (máxima de calidad).

El siguiente caso reúne todos los elementos que el modelo de Hilton asegura que existen en el intercambio de explicaciones. El hablante es un médico y sus interlocutores son un colega y la esposa de uno de sus pacientes:

> Un médico que diagnostica alcoholismo en un paciente. Cuando un colega que no conoce el historial del paciente le pregunta por qué este paciente es alcohólico, el médico puede responder que el paciente bebe ahora, y no lo hacía antes, porque perdió su trabajo. Sin embargo, cuando la esposa del paciente pregunta por qué su esposo se ha vuelto alcohólico, el médico puede responder que el esposo tiene una predisposición genética al alcoholismo. Esto explicaría por qué se volvió adicto al alcohol, mientras que sus compañeros de trabajo, que también fueron despedidos por la misma fábrica al mismo tiempo, no lo hicieron [Hilton, 1990, pp. 65-66].

En el ejemplo anterior el intercambio se inicia a partir de una pregunta ("¿por qué el paciente se volvió alcohólico?"). Contestar a esta pregunta es el objetivo de la conversación. Las explicaciones que el médico ofrece tienen la forma de atribuciones: 1) el paciente se volvió alcohólico porque tiene una predisposición genética; 2) el paciente se volvió alcohólico porque perdió su empleo.

Aunque ambos interlocutores plantearon la misma interrogativa, el médico presenta explicaciones diferentes porque está considerando que la

situación epistémica de la esposa es distinta a la de su colega. El médico no le dice a la esposa que el paciente se volvió alcohólico porque perdió su trabajo ya que está consciente de que ella ya sabe que su esposo fue despedido. Mencionar esa información sería redundante, lo que violaría la máxima de cantidad. Igualmente, el médico entiende que su colega sabe que hay factores de riesgo que propician que un sujeto se vuelva dependiente del alcohol. Decir que un factor genético fue lo que propició el alcoholismo del paciente no sirve para responder a la duda del colega.

Además, el médico asume que el desconcierto de la esposa surge ya que ella no tiene claro por qué los demás compañeros de trabajo de su esposo no se volvieron alcohólicos. En otras palabras, la pregunta de la esposa es "¿por qué mi esposo se volvió alcohólico y sus ex-compañeros de trabajo no?". Responder que el esposo se volvió alcohólico porque lo despidieron no sería suficiente: despidieron a todos los miembros de la fábrica, pero no todos se volvieron alcohólicos. El médico entiende la duda de su interlocutora, y le provee una explicación que encaje con lo que ella quiere saber.

Todo lo anterior muestra, según Hilton, que las máximas conversacionales y el contexto en el que se produce una pregunta constriñen la manera en la que se desarrolla un intercambio explicativo entre dos o más agentes. En la sección 1 presenté evidencia empírica a favor de la teoría de las máximas conversacionales y afirmé que la propuesta de Grice pretende ser descriptiva y no normativa. Lo mismo sucede con el modelo de Hilton: su intención es reflejar la manera real en la que las explicaciones sobre el comportamiento se comunican de un agente a otro. Si tomamos en cuenta que tanto la teoría de Grice como el modelo de Hilton cuentan con un respaldo empírico a su favor, parece plausible sugerir que ambos modelos capturan con éxito elementos de la comunicación humana, y que por lo tanto deberían de ser retomados para crear sistemas de XAI que sean capaces de explicar su comportamiento con una audiencia amplia. El siguiente capítulo de esta tesis argumentaré por qué la XAI debería basarse en el modelo de Hilton.

3. HILTON Y EL PROBLEMA COMUNICATIVO DE LA XAI

A continuación argumento que adoptar el modelo conversacional de la explicación de Denis Hilton permitiría resolver el problema comunicativo de la XAI. El capítulo se divide en dos partes. En la primera parte reviso múltiples reseñas sobre la literatura contemporánea en XAI. Analizar estas reseñas literarias me permite llegar a dos conclusiones: 1) que una inteligencia artificial explicable exitosa debería de tener en cuenta que su audiencia es amplia, y que necesita ofrecer varios tipos de explicación; 2) que muchos teóricos de la inteligencia artificial explicable están de acuerdo en que el intercambio de explicaciones es un acto comunicativo.

En la segunda sección del capítulo parto de estas dos conclusiones y argumento que para resolver el problema comunicativo de la XAI se necesita de un modelo que: 1) tome en consideración el contexto en el que tiene lugar el intercambio explicativo; 2) que proponga máximas que le indiquen al sistema cómo reaccionar ante la situación epistémica del interlocutor; y 3) que entienda que las explicaciones se hacen en respuesta a una pregunta. Concluyo que el modelo conversacional de Hilton cubre todos estos puntos, y que por lo tanto debería de ser adoptado como base para el desarrollo de sistemas explicables.

1. EXPLICACIÓN EN XAI: QUÉ Y POR QUÉ

Hay un problema en torno al significado y el uso de las palabras "explicación" y "explicabilidad" en la XAI. A diferencia de términos como "aprendizaje supervisado" o "aprendizaje por refuerzo", la noción de explicación no existe como un concepto técnico en el área del aprendizaje de máquinas. Con esto quiero que, en el contexto de la inteligencia artificial explicable, no hay una definición única para la palabra "explicación". Al menos tres reseñas de la literatura en XAI [Krishnan, 2020; Lipton, 2016; Miller, 2019] concluyen que el significado de "explicación" varía de trabajo a trabajo y de autor a autor.

Aunado a lo anterior, en más de un artículo se mezclan —de manera no precisamente clara— a las nociones de explicación y explicabilidad con términos como "transparencia", "inspeccionabilidad" o "interpretabilidad". A continuación hay un ejemplo de dos investigadores que hablan de explicación e interpretabilidad como conceptos equivalentes:

> En el contexto de los sistemas de aprendizaje de máquinas, definimos la interpretabilidad como la capacidad de ***explicar*** o presentar algo en términos comprensibles para un humano [Doshi-Velez y Kim 2017].

Al contrario de Doshi-Velez y Kim, en Gilpin et al. [2018] se afirma que la explicabilidad es algo distinto a la interpretabilidad; estos autores sostienen que un sistema es interpretable si es "capaz de resumir las razones de su comportamiento, de ganarse la confianza de sus usuarios, o de dar a conocer las causas de sus decisiones". En contraste, una IA explicable necesita, además, la capacidad de defender sus acciones y de proporcionar respuestas a las preguntas que se le planteen. Por su parte, Rudin [2018] define el aprendizaje de máquinas *interpretable* como aquél en donde se utilizan modelos que *no son cajas negras*, mientras que el aprendizaje de máquinas *explicable* es, para esta autora, cuando se utiliza una caja negra y despues se recurre a alguna técnica computacional para explicar su comportamiento. No tengo la intención de ahondar más en las nociones de "transparencia", "inspeccionabilidad" o "interpretabilidad". Lo que estoy tratando de enfatizar es la manera en la que "explicación", "explicabilidad" y otras palabras afines llegan a tener significados muy diferentes para quienes se dedican a la inteligencia artificial explicable.

En su reseña respecto a la noción de explicabilidad en la XAI, Miller [2019] expone de manera minuciosa las distintas formas en las que el concepto de explicación se usa en las ciencias sociales y las humanidades. Algo que queda manifiesto al leer el texto de Miller es que, al igual que en el mundo de la inteligencia artificial explicable, en áreas cómo la filosofía o la psicología

también hay una multiplicidad de definiciones para el concepto "explicación". Por ejemplo, Miller cita en varias ocasiones a David Lewis, quien afirma que "explicar un acontecimiento es proporcionar alguna información sobre su historia causal… en un acto de explicación, alguien que está en posesión de cierta información sobre la historia causal de algún acontecimiento intenta transmitirla a otra persona" [1986]. Quienes conocen la historia de la filosofía de la ciencia notarán que la definición de explicación que usa Lewis es muy distinta —e incluso contraria— a la propuesta por Hempel [1942], Gardiner [1959] o Nagel [1961]. Así pues, se puede aseverar que tanto en filosofía, como en XAI, como en otras áreas de las ciencias sociales y las humanidades, hay una variedad heterogénea de significados atribuidos a la noción de explicación. Ahora bien, Miller señala dos características que atraviesan a la mayoría de definiciones de explicación que se han ofrecido desde distintas disciplinas:

1. Las explicaciones suelen ser contrastativas. Esto quiere decir que cuando alguien pide que se le explique X, lo que en realidad está pidiendo es que se le explique *por qué X es el caso y no Y*.
2. Explicar algo es una forma de transferir conocimiento y, como tal, las explicaciones se presentan en marcos sociales: una conversación informal, el intercambio de datos entre dos colegas en un laboratorio, etc.

A estas dos características agregaría que muchas definiciones relacionan las explicaciones a preguntas de la forma "¿por qué X?". Además, y más importante, hay un aspecto clave a la hora de intentar definir qué es una explicación: en toda explicación intervienen por lo menos dos sujetos, el que la proporciona y el que la recibe.

Recapitulando: no hay una noción fija de "explicación" en la XAI, así como no la hay en otras áreas del conocimiento. Sin embargo, a pesar de que no hay una concepción concreta, las explicaciones suelen pensarse como: 1) respuestas a una pregunta; 2) formas de compartir información; 3) intercambios entre, mínimo, dos agentes. Todos estos elementos quedan capturados en la teoría de Hilton, pero esta no es una razón suficiente para

adoptar al modelo conversacional como base para el desarrollo de inteligencias artificiales explicables. El modelo conversacional de Hilton posee una definición clara de la palabra "explicación", y acogerlo en la XAI permitiría que los desarrolladores de sistemas explicables trabajen desde una noción homogénea de qué son las explicaciones. Ahora bien, como señala el estudio hecho por Miller, hay muchos autores que, al igual que Hilton, piensan a las explicaciones como intercambios de datos hechos entre dos o más individuos a partir de una pregunta. Es mi obligación, por lo tanto, presentar razones por las cuales el modelo de Hilton, y no alguna otra teoría, debería de ser tomado como cimiento en la XAI. Como mostraré a continuación, las razones para recibir el modelo del Hilton se hacen evidentes cuando nos preguntamos por qué queremos que haya sistemas de AI explicables.

Si bien no podemos dar una sola respuesta a la pregunta de qué es una explicación para los teóricos de la XAI, responder a la cuestión de por qué queremos inteligencias artificiales explicables es una tarea considerablemente más fácil: cuando en la literatura se habla de los propósitos de la XAI, las discrepancias entre un autor y otro son básicamente inexistentes. Según Sameket al. [2017] la búsqueda de sistemas explicables tiene cuatro propósitos:

A. Verificar el sistema: entender las reglas que rigen el proceso de decisión de un algoritmo para detectar y evitar posibles sesgos.

B. Mejorar el sistema: entender cómo opera un modelo para poder compararlo con otros sistemas y así perfeccionar su funcionamiento.

C. Aprender del sistema: extraer conocimiento respecto a cómo se comporta un algoritmo; este conocimiento sería aplicado, posteriormente, para desarrollar otras inteligencias artificiales similares.

D. Atender a las legislaciones y entidades que buscan regular el uso de algoritmos opacos.

Gilpin et al.[2018] coinciden en que estos son los objetivos de la XAI, pero además añaden, con respecto a los puntos **A** y **B**, que queremos sistemas explicables para asegurarnos que no fallarán fatalmente al momento de llevarse a la práctica y para garantizar que algún agente externo no esté interviniendo o modificando su comportamiento. En relación a **C**, Guidotti, et al. [2018] agregan que no solo queremos aprender de los sistemas para crear más programas de IA: el conocimiento que podamos obtener sobre cómo se comporta cierto algoritmo de aprendizaje profundo también podría ser aplicado a otras áreas de la informática y las ciencias de la computación. En cuanto a **D**, Wachter, et al. [26] describen que un programa que pueda explicar sus propias acciones no solo serviría para atender a las preguntas planteadas por órganos regulatorios, sino que también ayudará al usuario a entender por qué un algoritmo ha llegado a cierto resultado y, de ser necesario, le proporcionará motivos para impugnar las decisiones del sistema.

Un quinto objetivo que se le podría adjudicar a la XAI sería el de incentivar la adopción y aplicación de sistemas de aprendizaje profundo. Me referiré a este como el objetivo **E**. Mi idea es que la falta de inteligibilidad de un sistema —en particular si se produce una discordancia entre las expectativas del usuario y el comportamiento del sistema— puede llevar a los usuarios a desconfiar del sistema, a utilizarlo mal o incluso a abandonarlo por completo. Esta podría ser otra razón por la que nos gustaría tener inteligencias artificiales explicables.

En síntesis, hay al menos cinco motivos centrales por los cuales queremos algoritmos explicables. Lo que voy a destacar en primer lugar es que, como ya mencioné, aparentemente no hay ningún desacuerdo entre los teóricos de la XAI respecto a que estos son los propósitos de la inteligencia artificial explicable. En segundo lugar quiero recalcar que los objetivos **A**, **D** y **E** surgen al considerar los intereses de los usuarios y de aquellas personas afectadas por las decisiones del algoritmo; **B** y **C**, por otro lado, aparecen cuando tomamos en cuenta las metas de los desarrolladores y programadores. Hago este hincapié porque aún hay un elemento de discusión que voy a

abordar, a saber, el de cuáles preguntas debería de responder una inteligencia artificial explicable. Hasta el momento solo he mencionado que un sistema explicable debería de responder a preguntas del tipo "¿Por qué **P**?", donde **P** es una decisión tomada por el sistema. Esta caracterización no es incorrecta, y de hecho está presente en casi toda la literatura sobre XAI. Sin embargo, los autores que citaré a continuación creen que hay más preguntas que una inteligencia artificial explicable tendría que ser capaz de responder.

Lim et al. [2009] dicen que una inteligencia artificial debería de ser capaz de responder, como mínimo, a cuatro preguntas: ¿Qué hizo el sistema?; ¿Por qué el sistema hizo **P**?; ¿Por qué el sistema no hizo **Q**?; ¿Cómo puedo lograr que el sistema haga **X**?. Nótese, de nuevo, que algunas de estas preguntas serían planteadas por los usuarios y otras vendrían de los programadores. Ahora bien, algunos autores clasifican las explicaciones que podrían surgir de una XAI en función de si explican cómo funciona el modelo en su totalidad o de si explican una decisión específica [Gilpin et al., 2018; Guidotti et al., 2018,]. En el primer caso, la explicación es mucho más global y puede ayudar a los usuarios y a los desarrolladores a comprender la estructura del sistema en su totalidad . En el segundo caso, la explicación se centra en un *output* determinado y permite a los usuarios comprender mejor las razones por las que se produce tal resultado.

En general, hay múltiples preguntas a las que las XAI deberían dar respuesta. Sin embargo, dentro de los textos citados se puede observar un acuerdo bastante consistente sobre la importancia de las preguntas "¿por qué P?", donde, como ya lo dije, P refiere al *output* de un programa. Otra cosa que se puede notar en los artículos mencionados es que algunos autores y teóricos de la XAI sí están conscientes de que hay una diferencia entre las explicaciones que le interesan a los programadores y aquellas que podrían resultar de utilidad para los usuarios. En el primer capítulo de esta tesis afirmé que la gran mayoría de modelos que actualmente existen en el área ofrecen únicamente explicaciones del primer tipo, es decir, explicaciones que solo son de interés para los programadores. Nombré a este como el ***problema***

comunicativo de la XAI porque las explicaciones en cuestión no solo son de interés para los expertos en el área, sino que además no pueden ser interpretadas por alguien que no tenga una formación técnica en ciencias de la computación. En lo que queda de este capítulo voy a seguir hablando de este problema —y de cómo solucionarlo—, pero primero voy a cerrar esta sección resumiendo mis conclusiones:

> **C1.** En la XAI no existe una definición única de "explicación". Algunas de las definiciones disponibles se contradicen entre sí; otras tantas son muy ambiguas. Sin embargo, muchos autores comparten la idea de que las explicaciones son respuestas a una pregunta, que suelen ser contrastativas, y que se presentan como un intercambio de información entre dos o más agentes.
>
> **C2.** Los teóricos de la inteligencia artificial explicable están de acuerdo en que los sistemas explicables se crean con varios propósitos; tales objetivos pueden responder tanto a los intereses de los usuarios como a los de los científicos que diseñan los programas.
>
> **C3.** Hay diferentes preguntas que se espera que una XAI sea capaz de responder. Estas preguntas provienen de grupos de personas con conocimientos muy heterogéneos.

Voy a afirmar que una inteligencia artificial explicable "exitosa" es aquella que cumple con todos los propósitos y que responde a todas las preguntas que describí anteriormente en esta sección. De las conclusiones **C2** y **C3** se sigue que:

> **C4.** Una inteligencia artificial exitosa debería tener en cuenta que se dirige a una audiencia muy heterogénea.

Por otro lado, de **C1** se sigue que si las explicaciones son intercambios de información entre dos agentes, entonces:

C5. El acto de comunicar una explicación debe de adherirse a los principios cooperativos por los que se rigen las conversaciones humanas.

Las conclusiones **C4** y **C5** servirán como los dos ejes de acción que usaré para argumentar que, al adoptar el modelo de Hilton, se resolvería el problema comunicativo de la XAI.

2. EL PROBLEMA COMUNICATIVO Y EL MODELO CONVERSACIONAL DE HILTON

En la sección anterior mencioné que una inteligencia artificial explicable exitosa debería de ser capaz de proporcionar explicaciones a distintos grupos de personas. Esta es la conclusión a la que me referí como **C4**. A continuación clasifico en cuatro categorías a las personas que idealmente interactuarían con una XAI. Cada categoría está basada en los objetivos, antecedentes y relación que el sujeto podría llegar a tener con el sistema:

Grupo 1. Desarrolladores: investigadores en inteligencia artificial, desarrolladores de software o analistas de datos que crean el sistema.

Grupo 2. Usuarios de primer orden: especialistas en el área de conocimiento en la que se aplica el sistema. Por ejemplo: un médico que se apoya en una inteligencia artificial para hacer un diagnóstico; un agente de seguros asistido por un programa que le ayuda a elegir a los clientes que merecen su servicio.

Grupo 3. Usuarios de segundo orden: los destinatarios finales de las decisiones. Por ejemplo: una persona a la que se le acepta o rechaza un préstamo bancario; el paciente al que se le hace un diagnóstico usando una IA; etc.

Grupo 4. Órganos reguladores: personas que evalúan los sistemas antes de que salgan al mercado; esta categoría incluye a profesionales de la ley pero también a investigadores dedicados a la AI.

Si observamos con detenimiento los objetivos de la XAI enlistados en la sección anterior podemos detectar que no hay un perfil fijo para cada uno de ellos. Con esto quiero decir que puede ser el caso que, por ejemplo, los miembros del grupo 1 y el grupo 2 compartan los mismos objetivos. El objetivo **B** —mejorar el sistema— claramente puede apelar a un desarrollador, a alguien que quiere optimizar el algoritmo que ha creado, sin embargo, una entidad reguladora también estaría interesada en que el objetivo **B** se cumpla. Un segundo ejemplo: para los usuarios de primer orden, el objetivo principal puede ser aprender del sistema, comprender los mecanismos de inferencia que utiliza el algoritmo y luego aplicarlos, por su cuenta, al tomar decisiones. Por su parte, algún programador del Grupo 1 tendrá el interés de aprender del sistema, quizás para crear programas nuevos. Como último ejemplo: el objetivo A —verificar el sistema— está dirigido a los usuarios de segundo orden porque las decisiones de la AI pueden tener implicaciones económicas o personales para ellos, aunque verificar el algoritmo también puede ser relevante para los usuarios de primer orden, quienes podrían tener la responsabilidad legal de las decisiones que tome el programa.

El punto que quiero subrayar es que la audiencia que las inteligencias artificiales explicables tendrían que atender está compuesta por grupos de personas muy distintos pero que llegan a tener los mismos objetivos. Ahora bien, piénsese en el siguiente caso: **P** y **Q** son dos personas que recibirán una explicación de una XAI; **P** pertenece al Grupo 1 y **Q** pertenece al Grupo 3. Tanto **P** como **Q** tienen el mismo objetivo: verificar que el sistema no exhiba un sesgo. El problema aquí es que la situación epistémica de **P** y **Q** es diferente. En otras palabras, **P** y **Q** tiene conocimientos y antecedentes distintos, y una explicación que satisfaga las necesidades de **P** quizás no sea suficiente para atender las interrogantes de **Q**. Lo que describí en el primer capítulo como el problema comunicativo de la XAI consiste, precisamente, en ignorar esta diferencia entre las situaciones epistémicas que hay entre los cuatro grupos que conforman la audiencia total a la que las inteligencias artificiales explicables están dirigidas. El desarrollo de XAI debería basarse,

por lo tanto, en un modelo que no ignore estas diferencias epistémicas. Una inteligencia artificial explicable exitosa sabría qué conceptos usar y qué conceptos no, qué información es la que su interlocutor ya tiene y cuál no, qué datos son relevantes para resolver las dudas del humano con el que está interactuando y que datos resultan irrelevantes, etc.

La propuesta de Denis Hilton presenta, precisamente, un modelo que serviría para crear sistemas con la habilidad de generar una pluralidad de explicaciones y atender al conjunto de personas que conforman a la audiencia total de la XAI. Una de las conclusiones a las que llegué anteriormente — **C5** — fue la de que, si explicar es un acto conversacional, entonces alguien que pretende hacer una explicación debe de adherirse a las máximas propuestas por Paul Grice. El modelo de Hilton encaja perfectamente con **C5**: para Hilton, explicar es un acto comunicativo, orientado hacía una pregunta, y que se apega a las máximas griceanas. La forma en la que Hilton retoma la teoría de Grice es especialmente importante en el contexto de esta tesis. La teoría griceana de la conversación ofrece dos herramientas indispensables para resolver el problema comunicativo de la XAI. Estas dos herramientas son la máxima de cantidad y la máxima de relación:

- Según la máxima de cantidad, al tener una conversación, el hablante debe de proporcionar a su interlocutor tanta información como sea suficiente y necesaria para cumplir con los propósitos del intercambio.
- Según la máxima de relación, para lograr sus objetivos, los hablantes deben de abstenerse de proporcionar información que no esté relacionada con la conversación.

De acuerdo a la interpretación que Hilton hace sobre la máxima de cantidad, cuando **P** y **Q** participan en una conversación y **Q** hace una pregunta, **P** debe proporcionar toda la información que sea necesaria y suficiente para responder a la pregunta de **Q**. Con tal de adherirse a esta máxima, P debe de tomar en consideración las cosas que **Q** ya sabe, los datos que **Q** desconoce, los

conceptos que **Q** domina y los que ignora, etc. La interpretación de Hilton sobre la máxima de relación, por otro lado, le indica a **P** que no debe de compartir información irrelevante para resolver la duda de **Q**. Si implementamos estas dos máximas al desarrollar inteligencias artificiales explicables, los sistemas resultantes producirían explicaciones sin ignorar la situación epistémica de su interlocutor.

Además, la primera conclusión a la que llegué en la sección anterior — **C1** — fue que si bien no hay una definición única de "explicación" en el campo de la XAI, la literatura indica que para muchos investigadores una explicación tiene que ser una respuesta a una pregunta. Esta, de nuevo, es una idea que está presente en el modelo de Hilton, y, al menos en el contexto de esta tesis, es un rasgo importante de su teoría. Supóngase que **X** es una inteligencia artificial explicable y que **Y** es un agente humano. En este escenario idealizado, **X** es un IA creada a partir del modelo conversacional de Hilton. Como tal, X "entiende" que sus respuestas tienen que resolver la pregunta que su interlocutor le presente. Al implementar en un sistema las máximas de cantidad y relevancia, y la idea de que las explicaciones están orientadas a un objetivo, la inteligencia artificial resultante será capaz de producir explicaciones "hechas a la medida", y esto es precisamente lo que queremos para resolver el problema comunicativo de la XAI.

Una aclaración necesaria: al afirmar que las explicaciones hechas por una XAI basada en el modelo de Hilton estarían "hechas a la medida" no quiero decir que serían *ad hoc*. En filosofía de la ciencia las explicaciones *ad hoc* suelen pensarse como "hechas a la medida" en tanto son explicaciones que pretenden atender las deficiencias particulares de una teoría. La noción de "hecho a la medida" que estoy usando tal vez se traduciría al inglés como "*custom-made*" o "*tailored*". En este sentido, cuando digo que las explicaciones estarían hechas a la medida lo que quiero decir es que serían explicaciones personalizadas, explicaciones que el sistema crearía ajustándose a las situación epistémica, a las preguntas y a los propósitos de su interlocutor.

En síntesis: para resolver el problema comunicativo de la XAI se necesita de un modelo que haga explicaciones de acuerdo al estatus epistémico de su interlocutor, y que además esté al tanto de que todas las explicaciones se hacen con el fin de alcanzar un objetivo. El modelo conversacional de Hilton cubre todos estos puntos: 1) toma en consideración el contexto (y el estado epistémico) en el que tiene lugar el intercambio de explicaciones; 2) propone máximas que indican cómo reaccionar ante la situación epistémica del interlocutor; y 3) entiende que las explicaciones se hacen en respuesta a una pregunta. Por lo tanto, implementar el modelo conversacional de Hilton como base en el desarrollo de sistemas explicables resolvería el problema comunicativo de la XAI.

CONCLUSIONES Y TRABAJO A FUTURO

En el primer capítulo de esta tesis argumenté que los sistemas de inteligencia artificial explicable sufren de una limitación: las explicaciones que ofrecen son demasiado técnicas, posiblemente útiles para un desarrollador o un experto en ciencias de la computación, pero ininteligibles para el usuario final. Denominé a este fenómeno como "el problema comunicativo de la XAI". A lo largo de los capítulos dos y tres argumenté que para resolver este problema se necesitaba de un modelo que cumpliera con ciertas características particulares. Concluí, finalmente, que el modelo conversacional de Denis Hilton es el adecuado para resolver el problema comunicativo de la XAI.

Como mencioné en la introducción, este texto es puramente teórico. En ningún momento tuve la intención de mostrar cómo es que un ingeniero o un informático podría tomar la propuesta de Hilton y llevarla a la práctica. Desconozco cómo es que el modelo conversacional podría insertarse en un sistema de XAI, e incluso —me atrevo a afirmar— tal vez aún no existen los recursos para llevar la teoría de Hilton al mundo de la inteligencia artificial. Reitero que la aportación de este trabajo se encuentra en la manera en la que he presentado un ideal regulativo: una forma ideal en la que todos los sistemas de aprendizaje profundo deberían de desarrollarse, y a la que todos los creadores de XAI deberían de aspirar.

El trabajo a futuro respecto a esta tesis también debería de enfocarse en la posibilidad de que existan otros modelos distintos al de Hilton que también cumplan con las condiciones que establecí como necesarias para resolver el problema comunicativo de la XAI. Desconozco si existe algún otro modelo de este tipo, pero es una posibilidad que no podría descartar. Ahora bien, tengo la certeza de que cualquier modelo que se trate de adoptar para resolver el problema comunicativo debería de apegarse a las máximas griceanas y de capturar el carácter contextual que tienen los intercambios de explicaciones.

En resumen, hay dos líneas de investigación que podrían desprenderse de este trabajo. La primera de ella pertenece a las ciencias de la computación

y consiste en averiguar cómo implementar el modelo de Hilton en la XAI; en caso de que esto no sea posible, la investigación tendría que enfocarse en producir sistemas que se aproximen a lo que el modelo conversacional de la explicación dicta sobre la forma en la que los humanos comunicamos nuestras explicaciones. La segunda línea de investigación que podría surgir de mi tesis es la de si hay algún otro modelo sobre la comunicación de explicaciones que encaje con los requisitos que establecí para resolver el problema comunicativo de la XAI.

REFERENCIAS:

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., & Herrera, F. (2019). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, *58*, 82-115.

Buckner, Cameron Joseph. "Black boxes, or unflattering mirrors? Comparative bias in the science of machine behavior." (2021).

Grice, Herbert P. "Logic and conversation." *Speech acts*. Brill, 1975. 41-58.

Hilton, Denis J. "Conversational processes and causal explanation." *Psychological Bulletin* 107.1 (1990): 65.

Malle, Bertram F. *How the mind explains behavior: Folk explanations, meaning, and social interaction*. Mit Press, 2014.

Ruben, David-Hillel. *Explaining explanation*. Routledge, 2015

Slugoski, Ben R., et al. "Attribution in conversational context: Effect of mutual knowledge on explanation-giving." *European Journal of Social Psychology* 23.3 (1993): 219-238.

Tetlock, Philip E., and Richard Boettger. "Accountability: A social magnifier of the dilution effect." *Journal of personality and social psychology* 57.3 (1989): 388.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Gardiner, P. L., (1959), *The Nature of Historical Explanation*, Oxford: Oxford University Press.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018, October). Explaining explanations: An overview of interpretability of machine

learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)* (pp. 80-89). IEEE.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, *51*(5), 1-42.

Krishnan, M. (2020). Against interpretability: a critical examination of the interpretability problem in machine learning. *Philosophy & Technology*, *33*(3), 487-502.

Lewis, D. K. (1986). Causal Explanation.

Lin, Patrick, Keith Abney, and Ryan Jenkins, eds. Robot ethics 2.0: From autonomous cars to artificial intelligence. Oxford University Press, 2017.

Lipton, Z. C. (2018). The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, *16*(3), 31-57.

Nagel, E. (1961), *The Structure of Science: Problems in the Logic of Scientific Explanation*, New York: Harcourt, Brace and World.

Nyholm, Sven, and Jilles Smids. "The ethics of accident-algorithms for self-driving cars: An applied trolley problem?." *Ethical theory and moral practice* 19.5 (2016): 1275-1289.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, *267*, 1-38.

Rudin, C. (2018). Please stop explaining black box models for high stakes decisions. *stat*, *1050*, 26.

Samek, W., Wiegand, T., & Müller, K. R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*.

Sears, A., & Jacko, J. A. (Eds.). (2009). Human-computer interaction fundamentals. CRC press.

Vilone, G., & Longo, L. (2020). Explainable artificial intelligence: a systematic review. *arXiv preprint: arXiv:2006.00093*.