UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
PROGRAMA DE POSGRADO EN FILOSOFÍA DE LA CIENCIA
FILOSOFÍA DE LA CIENCIA

MÁQUINAS QUE PIENSAN CAUSAS

TESIS
QUE PARA OPTAR POR EL GRADO DE:
MAESTRÍA EN FILOSOFÍA DE LA CIENCIA

PRESENTA:
ALBERTO DOMÍNGUEZ HORNER

DR. FRANCISCO HERNÁNDEZ QUIROZ
FACULTAD DE CIENCIAS

DRA. KAREN GONZÁLEZ FERNÁNDEZ
UNIVERSIDAD PANAMERICANA, FACULTAD DE FILOSOFÍA

DR. ALFONSO ARROYO SANTOS
CENTRO DE INVESTIGACIÓN GEOPROSPECTIVA

DRA. BEGOÑA FERNÁNDEZ FERNÁNDEZ
FACULTAD DE CIENCIAS

DR. CARLOS ÁLVAREZ JIMÉNEZ
FACULTAD DE CIENCIAS

CIUDAD UNIVERSITARIA, CIUDAD DE MÉXICO, MAYO,
2024

Universidad Nacional Autónoma de México

UNAM POSGRADO Filosofía de la Ciencia

PROPUESTA UNIVERSITARIA DE INTEGRIDAD Y HONESTIDAD ACADÉMICA Y PROFESIONAL
(Graduación con trabajo escrito)

De conformidad con lo dispuesto en los artículos 87, fracción V, del Estatuto General, 68, primer párrafo, del Reglamento General de Estudios Universitarios y 26, fracción 1, y 35 del Reglamento General de Exámenes, me comprometo en todo tiempo a honrar a la Institución y a cumplir con los principios establecidos en el Código de Ética de la Universidad Nacional Autónoma de México, especialmente con los de integridad y honestidad académica.

De acuerdo con lo anterior, manifiesto que el trabajo escrito titulado **Máquinas que piensan causas** que presenté para obtener el grado de **Maestría** es original, de mi autoría y lo realicé con el rigor metodológico exigido por mi programa de posgrado, citando las fuentes de ideas, textos, imágenes, gráficos u otro tipo de obras empleadas para su desarrollo.

En consecuencia, acepto que la falta de cumplimiento de las disposiciones reglamentarias y normativas de la Universidad, en particular las ya referidas en el Código de Ética, llevará a la nulidad de los actos de carácter académico administrativo del proceso de graduación.

Atentamente

Alberto Domínguez Horner, 522000957

Máquinas que piensan causas

ÍNDICE

Agradecimientos

Agradezco enormemente al Dr. Francisco Hernández Quiroz por la dirección de esta tesis y el acompañamiento a lo largo de todo el trabajo. Me acompañó y orientó desde antes de que definiera el tema de la tesis. Me ayudó a corregir cada detalle incluso cuando se trataba de temas algo alejados de las áreas de investigación a las que él se dedica. No pude tener mejor director para esta tesis.

Gracias a la Dra. Karen González Fernández. El hecho de que sea revisora de mi tesis, además de, en la licenciatura, haberme entrenado dos veranos enteros para la Olimpiada de Lógica y, al terminar la licenciatura, haberme orientado y aconsejado para hacer una buena elección de maestría, para mí resulta invaluable.

El Dr. Alfonso fue quien me introdujo al trabajo de Pearl. Nos sugirió leerlo cuando terminó el curso de Lógica II. También me orientó en mis primeros pasos hacia el *Machine Learning*. Le estoy muy agradecido. Esta tesis hubiera sido otra muy distinta de no haber sido por él.

En el último semestre de la maestría tuve el honor de presentar mi trabajo ante el seminario de Historia y Filosofía de las matemáticas que organiza la Dra. Carmen Martínez Adame. En esa sesión del seminario, la Dra. Begoña Fernández Fernández me compartió muchas observaciones y algunas referencias bibliográficas sobre las que después conversamos ella y yo. Me dio la oportunidad de hacer relevantes las ideas de esta tesis no sólo para las discusiones de filosofía de la ciencia, sino para el diálogo con los mismos científicos que aplican los métodos que aquí se discuten.

Llegué a ese seminario por invitación del Dr. Carlos Álvarez Jiménez. En los dos cursos que cursé con él, nos enseñó a apreciar —a esforzarnos por reconstruir— la complejidad y genialidad de los objetos matemáticos en el momento en el que fueron pensados por primera vez. Nos enseñó a disfrutar esa genialidad. Así que, motivado por ello, no pude evitar trasladar tales reflexiones a los esfuerzos que en las últimas décadas se han realizado para cuantificar la causalidad. Me honra mucho que sea revisor de mi tesis.

Quiero agradecer también a mi profesor y amigo el Dr. Fernando Galindo Cruz. Me dio los mejores consejos en momentos difíciles y de mucha incertidumbre.

Gracias a mi mamá, a mis hermanos Marce y Daniel por su apoyo incondicional y por creer en mí. Gracias a mis amigos Checo, David, Jorge, Gustavo, Gabi, Binnui, Dani, Jime, Carla, Vanessa, Octavio, Daco, Lautaro. A Gustavo por prestarme su bici. A Vanessa por tantas dudas de matemáticas, *Machine Learning* y tan buenas conversaciones. A Carla, por sacarme de la biblioteca. A Lautaro, por sacarme de fiesta. A Daco, por su clase de Cálculo Vectorial y el seminario de Deep Learning. A la Dra. Ruth, por enseñarme estadística.

Y gracias a mis abuelitos, por su casa y porque tanto a ellos como a mi mamá y a mis hermanos, los quiero con todo mi cerebro.

Introducción

Entender las causas de un fenómeno implica un mejor conocimiento que sólo resumir los datos sobre su comportamiento. Esta premisa fundamenta la tesis del presente trabajo. El lector encontrará una comparación entre los métodos de aprendizaje automático predominantes actualmente con los métodos de inferencia causal. Particularmente (pero sin reducirse a ello), este trabajo compara la teoría del aprendizaje estadístico en la que se basa el aprendizaje automático, considerada tal como la expone Vladimir Vapnik, con la teoría de la causalidad probabilista estructural (y de la inferencia de relaciones causales) considerada tal como la exponen Judea Pearl, sus colaboradores y sus continuadores.

No se trata sólo de la comparación de dos métodos de inferencia. Ambos métodos tomados de manera aislada no son incompatibles entre sí. *El propósito es comparar las ideas que los guían y sus presupuestos sobre lo que es escible y lo que es la investigación científica.* Las ideas y presupuestos que acompañan a esos métodos sí son incompatibles. Por lo tanto, hay que rechazar ambas o escoger sólo una propuesta. El que sean incompatibles en sus presupuestos, pero no se contradigan en términos estrictamente formales, ha permitido que algunos de los científicos que los usan no los consideren incompatibles en absoluto.

Escoger la propuesta epistemológica del aprendizaje automático implica rechazar el uso de los métodos de inferencia causal. Sin embargo, escoger la teoría de la causalidad estructural probabilista, no implica rechazar los métodos de aprendizaje automático (sólo algunas ideas que han guiado su desarrollo).

A lo largo de la tesis se verá en qué consisten esos métodos y esas propuestas. La razón por la que la disyunción ocurre de este modo particular es que la teoría causal propone tres niveles de conocimiento: asociaciones, intervenciones y contrafácticos. Los razonamientos en cada nivel posterior presuponen la validez de los niveles anteriores; ello permite que estos tres niveles se interpreten como fases en una investigación, aunque no es estrictamente necesario proceder en ese orden. Puesto que los métodos de aprendizaje automático son especialmente eficaces en tareas de

asociación (incluidas aprendizaje supervisado, no-supervisado, por refuerzo, etc.), se pueden usar en la primera fase y usar métodos causales en las otras dos.

Pero los presupuestos que acompañan al aprendizaje automático preeminente, que hacen de la causalidad una afirmación demasiado ambiciosa como para ser considerada científica, establecen que no hay conocimiento más allá del nivel asociativo. Así que estos presupuestos desplazan tanto a la teoría como a los métodos causales. Valga reconocer que hay investigaciones en las que las preguntas y las afirmaciones causales no resultan pertinentes (gran parte del conocimiento en física cuántica, por ejemplo, se expresa con ecuaciones en las que postular alguna dirección causal resulta inadecuado). Todos los casos que aquí se tratarán son casos en los que resulta razonable pensar en afirmaciones causales tal como se definirán en la siguiente sección (informalmente) y en la sección V (formalmente).

En este trabajo se defiende como tesis que (i) tanto los métodos basados en la teoría del aprendizaje estadístico como los basados en modelos causales tienen mucho que aportar, y en *un nivel meramente técnico* conviene encontrar maneras de fusionarlos, pero que (ii) *la epistemología hacia la que tienden* los métodos actualmente prevalentes de aprendizaje automático ha de ser rechazada en favor de la epistemología a la que están asociados los métodos de razonamiento causal. (iii) El propósito de la actividad científica (perfeccionar el conocimiento colectivamente) es una razón fuerte para preferir la epistemología de los métodos de razonamiento causal.

La tesis se defiende con tres argumentos que se exponen en la Sección VI, y que se corresponden con tres perspectivas desde las cuales se puede pensar en el propósito de la actividad científica: la epistemológica, la del proyecto científico concreto de Inteligencia Artificial, y la de la cognición humana a un nivel fisiológico —a fin de cuentas, es con esa fisiología con la que los humanos practicamos la actividad científica—. Dichos argumentos no son los únicos que se pueden aducir para defender la tesis, y tampoco son argumentos compulsivos (es decir, la tesis no se sigue deductivamente de la verdad de los argumentos).

El primer argumento parte de tres cualidades epistémicas que, desde el punto de vista de la perfección del conocimiento, es mejor tenerlas que no tenerlas; tales cualidades están presentes en los modelos causales, y ausentes en la teoría del aprendizaje estadístico. Las tres cualidades son la comprensión en las representaciones, la identificación de las relaciones entre diversas distribuciones de probabilidad, que se corresponden con los diferentes estados de un mismo proceso, y las relaciones de relevancia o irrelevancia que forman parte de un mismo modelo. Llamémoslas *comprensión*, *transportabilidad* y *relevancia*.

El segundo argumento parte de los propósitos explícitos del proyecto científico de Inteligencia Artificial y de la manera en que éste se ha desarrollado históricamente. Los métodos causales son más oportunos a tales propósitos y a tal desarrollo que los métodos de aprendizaje profundo.

El tercer argumento toma como punto de partida los resultados de (Jeong el al., 2022), sobre el rol de la dopamina en el aprendizaje de asociaciones causales. Es probable que la manera en que lo humanos procesamos la información fisiológicamente sea más cercana a los algoritmos de aprendizaje causal que a los algoritmos de aprendizaje profundo.

Las cinco secciones previas a la presentación de esos tres argumentos tienen como objetivo establecer el contexto adecuado en el que han de ser defendidos los argumentos.

Puede resultar llamativo que, si esta tesis se enfoca en los métodos causales, tales métodos se presenten sólo preliminarmente en la Sección I y formalmente hasta la Sección V. La principal razón por la cual se estructuró así la tesis es que se busca presentar a los métodos de razonamiento causal como una respuesta a los métodos de aprendizaje estadístico. Se ha procurado que las ideas principales sean asequibles a lectoras y lectores que no estén familiarizados con los métodos de razonamiento causal; sin embargo, no se pretende que esta tesis sea una introducción a tales métodos. La mejor introducción que el autor conoce es el libro *The Book of Why* (Pearl, 2018). Para tener una idea más clara sobre qué se considerará como inferencia causal al presentar los tres argumentos, es recomendable hojear la Sección V antes de leer la tesis. Quien esté familiarizado o

familiarizada con los métodos de aprendizaje profundo, su historia, y con los métodos de inferencia causal, siéntase libre de leer directamente la sección VI, y después, si le parecen valiosos los argumentos, leer lo demás para encontrar el relieve que se les pretende dar a los tres argumentos.

Algunos comentarios sobre los métodos causales aparecen antes de la Sección V, tales comentarios están pensados para los lectores que están familiarizados con el tema, y se espera que no sean demasiado enfadosos para quienes no lo están. Para abonar a la comprensión se ha añadido un pequeño glosario al final.

El orden obedece a la siguiente lógica: la primera sección presenta de manera informal la idea de causalidad con la que se trabajará en esta tesis; el propósito es que el lector se familiarice con la idea y tenga una referencia con la cual comparar lo que se plantea en la secciones II, III, y IV. Las secciones II y III presentan las características relevantes para los argumentos de la sección VI del proyecto de IA y de la teoría del aprendizaje estadístico, respectivamente. El objetivo principal es presentar los principios que guían el aprendizaje en las redes neuronales profundas y los aspectos de la trayectoria del proyecto de IA que son relevantes para la argumentación que aquí se propone. Históricamente, el razonamiento con redes bayesianas fue propuesto después de que ya se habían probado y presentado las redes neuronales. Epistemológicamente, si tomamos en cuenta lo que Pearl llama la *escalera de la causalidad* (los tres niveles de conocimiento causal), las redes neuronales se encuentran en un nivel epistemológico más básico que los modelos causales. El hecho de que sean precedentes en ambos ámbitos hacen razonable presentar primero a las redes neuronales, y proponer a los modelos causales como una manera de alcanzar cualidades epistémicas que no son alcanzables sólo con las redes neuronales y el aprendizaje estadístico. La Sección IV busca explicar y aclarar ese salto del razonamiento meramente asociativo al razonamiento causal. Por ello es hasta la Sección V que se presenta el cálculo causal. Y la sección que el autor de esta tesis considera la más importante es la Sección VI, donde se presentan los argumentos con los que se defiende la tesis. Los argumentos asumen la idea sembrada en las secciones anteriores: ver al razonamiento causal como un campo de estudio que propiamente

pertenece al proyecto de IA, y como una mejora frente a los métodos de aprendizaje estadístico.

Hay dos temas que no trata esta tesis y conviene anticipar por qué. El primero es la posibilidad de la conciencia artificial. Esta tesis se concentra en los razonamientos (artificiales o naturales), no en la manera en que se experimentan tales razonamientos, y tampoco en las experiencias que los provocan. Por lo tanto, el problema de la conciencia no será tratado aquí.

El segundo tema son los LLMs (*Large Language Models*). Al momento en que se escribe esta tesis, el *producto* que ha recibido mayor atención relacionado con la Inteligencia Artificial más a es GPT-4, un producto de *OpenAI*. Los LLMs sí están relacionados con el tema de esta tesis, pero no se abordarán directamente.

La mayoría de los usuarios en el mundo interactúan con los LLMs en el momento en que ya están entrenados; los LLMs como GPT han sido liberados al público como productos. Sin embargo, responder *prompts* es la parte menos inteligente del proceso. Las genialidades de los LLMs se encuentran en su proceso de entrenamiento —aprendizaje— , en su arquitectura y en la manera de orquestar tantos recursos computacionales e informáticos (memoria, GPUs, redes, la información de internet, etc.).

Los LLMs son redes neuronales. Comparten con las redes neuronales profundas (*Deep Learning*) los mismos principios generales de aprendizaje.

Esta tesis no se enfoca en detalles propios de los *productos* que ha traído consigo el desarrollo del proyecto de inteligencia artificial, sino en los principios generales de aprendizaje que comparten los métodos de *Machine Learning* más utilizados actualmente. Tales principios están capturados en la *teoría del aprendizaje estadístico*¸ que aquí se toma tal como la expone Vladimir Vapnik (2000), y en el algoritmo de *retropropagación*, que apareció por primera vez publicado en *Nature* (Rumelhart et al., 1986a).

Se asume que las *tendencias epistemológicas* que aquí se describen respecto de la teoría del aprendizaje estadístico y el algoritmo de retropropagación aplican también a los productos de mercado como GPT-4. Por lo tanto, preferir la epistemología de los modelos causales implica también

cierta actitud con relación a tales productos. No es que los productos que ahora hemos visto en procesamiento de lenguaje y procesamiento de imágenes sean 'malos'. De hecho, albergan una utilidad inmensa. El punto es que resultan inútiles para cierto tipo de problemas, y podemos esperar mejores resultados cuando se trata de la Inteligencia Artificial Causal.

I. ¿De qué causalidad estamos hablando?

No es una causalidad estilo Newton; no se asumirá aquí el supuesto de que la causalidad es determinista y sólo aplica a sistemas deterministas. A lo largo de la argumentación en las siguientes secciones, asumo una idea probabilista estructural de la causalidad. No es 'probabilista' en el mismo sentido que el proyecto de causalidad probabilista asociado con Patrick Suppes y Wesley Salmon (Suppes 1970, Salmon 1988). Por eso se añade el adjetivo 'estructural'.

La conferencia inaugural de 1971, en Cambridge, impartida por G. E. M. Anscombe permite introducir esta idea de causalidad. Anscombe explica que, durante muchos siglos, la idea de causalidad estuvo engarzada a la idea de necesidad. Incluso David Hume, el gran crítico de las afirmaciones causales, presupone ese vínculo en sus argumentos (Hume, 1772, §48-61). De acuerdo con Anscombe, todas las teorías de la causalidad que presuponen la necesidad como constitutiva de ésta, comparten el siguiente presupuesto:

> Si un efecto ocurre en un caso y un efecto similar ocurre en otro caso aparentemente similar, debe haber una diferencia relevante por descubrirse. (Anscombe, 1971, p. 1)[1]

Contraria a esta idea, al negar que la causalidad esté engarzada con la necesidad, tenemos la siguiente afirmación: aunque un efecto ocurra en un caso, puede no darse un efecto similar en otro casi similar sin que haya alguna diferencia relevante.

Permítaseme usar un ejemplo diferente a los de Anscombe. (Ella usa ejemplos de física que, dados los propósitos de esta investigación, nos obligarían a detenernos innecesariamente.)

Soy una persona que se enferma fácilmente de las vías respiratorias. Antes de dormir suelo entrecerrar la ventana, de modo que aún entre oxígeno, pero que no pueda cambiar abruptamente la temperatura en la

[1] Las traducciones son del autor.

habitación. Por simplicidad, supongamos que sólo hay dos escenarios: o cerré la ventana o la dejé abierta. Me ha pasado varias veces que, exactamente el día en que olvido cerrar la ventana, amanezco resfriado. Es razonable suponer que haber dejado la ventana abierta es la causa de mi resfriado, pero nótese que (1) pude haber dejado la ventana completamente abierta y no resfriarme y (2) también es posible resfriarme sin que haya dejado la ventana completamente abierta. Con suficientes datos podríamos asignar una probabilidad a cada caso.

Hay otra consecuencia más de esta noción no necesaria de la causalidad. Anscombe señala que «es más fácil retrotraer los efectos hacia las causas con certeza, que predecir efectos a partir de las causas». Y añade: «con frecuencia conocemos una causa sin saber si hay una generalización sin excepciones de ese tipo» (1971, p. 6). Se refiere a que las afirmaciones causales pertenecen más al razonamiento retrospectivo —entendamos lo que ocurrió una vez que ocurrió— que al razonamiento prospectivo —predigamos lo que ocurrirá a partir de los indicios de los que disponemos—.

Vemos entonces tres facetas de la negación de que la causalidad sea un asunto de necesidad: la causalidad (i) no siempre es una conexión necesaria, (ii) no siempre es una generalización sin excepciones, y (iii) no siempre ofrece utilidad predictiva.

Tales tres maneras de hablar refieren a una misma idea: afirmar algún tipo de causalidad no nos compromete con la presuposición de que el fenómeno que estamos tratando esté *pre-determinado*. Como advierte Anscombe más adelante, es diferente que un suceso esté determinado a que un suceso esté predeterminado (Anscombe, 1971, p. 17). En el ejemplo del resfriado, mis síntomas y la apertura de la ventana están ya determinados; pero sería una afirmación muy fuerte decir que, una vez que me dormí sin haber cerrado la ventana, estaba ya (pre)determinado que me resfriara. Es decir, sería muy atrevido pensar que, antes de que sucediera, de todas las posibilidades que podemos imaginar (resfriado, no-resfriado, alergia, coronavirus, enfermedad bacteriana, pulmonía, etc.), el resfriado junto con todas sus características peculiares era el único resultado factible. Podemos aceptar que algunos sucesos fueron determinados por otros sin comprometernos con que hubieran estado predeterminados.

Hay una razón más por la que se prefirió usar este ejemplo en lugar de los de Anscombe: La apertura de la ventana no es una causa directa del resfriado. Lo que sucede es que ésta *permite* el intercambio de temperatura entre el exterior y el interior, lo cual *altera* el sistema inmunológico. No se trata de una relación exclusiva entre dos variables (causa y efecto), sino de una estructura causal y de las relaciones entre varias variables que, a su vez, pueden ser causas y efectos, directas o mediadas.

Podemos preguntarnos ¿por qué dejar la ventana abierta es causa del resfriado? Y responder: porque afecta al sistema inmunológico. Notar que hay algo *en medio* del enfriarse y el resfriarse permite señalar que la causalidad es una propiedad de estructuras de variables, y no tanto de dos variables aisladas.

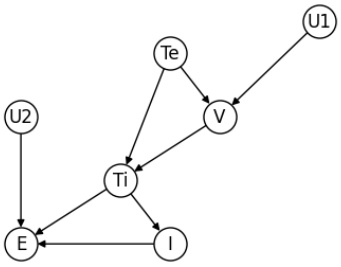
En las siguientes páginas, cuando hablo de causalidad, me refiero a una relación que no se compromete con los supuestos que el texto de Anscombe nos permitió hacer explícitos. Se trata de una causalidad sin necesidad, que, sin embargo, puede entenderse en términos de cierta regularidad, de modo estructural y en algunos casos como una indagación sobre la naturaleza. Para algunos podría parecer una noción poco atractiva para la investigación científica, sobre todo por obedecer un razonamiento retrospectivo más que uno prospectivo. No obstante, más adelante (al hablar del valor empírico de las probabilidades contrafácticas) mostraré que después de realizar el análisis retrospectivo, las afirmaciones causales nos permiten obtener una habilidad predictiva que la mera asociación de hechos no nos puede ofrecer.

La noción de causalidad en la que se centra este trabajo es la de Judea Pearl. De acuerdo con el trabajo de Pearl, la inferencia causal es un paso posterior a la inferencia estadística. Esto es, las relaciones causales se infieren a partir de distribuciones de probabilidad sobre un conjunto de variables. Primero se estima una distribución de probabilidad $\hat{P}$; después se hallan las dependencias entre las variables (que definen una red bayesiana, o un

conjunto de redes bayesianas);[2] por último, se infiere cuáles de estas dependencias son relaciones causales o productos de relaciones causales; de ello se obtiene la estructura causal, que se expresa con un grafo acíclico dirigido (un DAG).

Los DAGs (*Directed Acyclic Graph*) son la herramienta básica de la inferencia causal para la representación del conocimiento disponible y de las hipótesis propuestas. Un DAG está compuesto por nodos y enlaces. Los nodos representan las variables y los enlaces las relaciones entre éstas. Son dirigidos porque las relaciones tienen explícitamente un origen y un destino. Son acíclicos porque no admiten que un nodo sea al mismo tiempo el inicio y el final de la cadena de relaciones del diagrama ni de un subconjunto de ella: no admiten ciclos.

Podemos construir un DAG sobre el ejemplo del resfriado:



V: ventana,
Te: temperatura exterior,
Ti: temperatura interior,
I: sistema inmunológico,
E: enfermedad,
U1: factores desconocidos,
U2: factores desconocidos.

[2] Algunos autores como Embrechts, McNeil & Straumann (2002) han advertido sobre malas prácticas en contextos financieros en los que erróneamente se infiere dependencia a partir de correlaciones. Sus advertencias son válidas en general, no sólo en los contextos financieros. Uno de sus ejemplos es la comparación entre una distribución normal y una distribución Gumbel. Muestran un caso con la misma correlación $\rho = 0.7$ pero con diferentes estructuras de dependencia. El autor de la presente tesis es consciente de tales problemas probabilistas, pero los argumentos que aquí se discutirán se centran en el segundo momento de la inferencia, cuando las relaciones de dependencia están dadas y se busca entender la estructura causal que las origina.

De este modo, se aprecia la información causal con mayor transparencia. El grafo hace explícitos ciertos supuestos tanto por la presencia como por la ausencia de enlaces. Hay una flecha directa de *V* a *Ti*, lo cual significa que uno puede afectar la temperatura de la habitación cerrando o abriendo la ventana. Pero no hay una flecha directa de *V* a *I*, lo cual deja en claro que la ventana no influye por sí misma directamente en el sistema inmunológico; sólo lo hace a través de su efecto en la temperatura de la habitación.

Este primer ejemplo de un grafo también nos permite introducir de manera informal la *condición de Markov*. Se suelen interpretar las relaciones entre las variables con una metáfora genealógica: en cualquier cadena de relaciones, las variables que preceden son padres, abuelos, ancestros, etc., y las variables que suceden son hijas, nietas, descendientes, etc. La *condición de Markov* establece que una variable es independiente de todas las demás, dados sus padres. Por ejemplo, los padres de *E* son {*Ti, U2, I*}. Eso significa que, si se conoce el estado de esas tres variables, se puede calcular la probabilidad de que me enferme, y en nada afectará al resultado del cálculo saber o no si la ventana está abierta. En específico, como la ventana sólo afecta a la enfermedad a través de su efecto en la temperatura interior, el diagrama implica que, dada la temperatura interior, la enfermedad es independiente de la ventana (se escribe $E \perp\!\!\!\perp V \mid Ti$).

Otro elemento relevante son las variables *U*. La letra refiere al inglés *unknown* (desconocido). Se utilizan para tomar en cuenta los factores externos al modelo causal. *U1* puede significar el estrés, nivel de cansancio, o cualquier condición que implique estar más o menos alerta para cerrar la ventana. *U2* puede significar la presencia aleatoria de patógenos que afectarán al organismo.

Por último, el autor de esta tesis no es médico. La conjetura de la ventana bien podría resultar falsa. Sin embargo, definirla a través de un DAG permite mostrar con transparencia qué es lo que está afirmando esa conjetura, qué es lo que supone y qué es incorrecto en ella en caso de que resultara ser incorrecta.

Ahora bien, el propósito de este trabajo no es defender ni justificar la idea de causalidad que respalda los métodos de Judea Pearl. En la presente argumentación, presupongo que la idea es correcta. De la misma manera, aunque no defiendo que las herramientas de inferencia causal desarrolladas por Pearl sean perfectas, asumo que están orientadas en la dirección correcta. Estas dos afirmaciones funcionan como premisas en este trabajo. En trabajos posteriores, presentaré la justificación y defensa de ambas premisas.

II. Ideas que han guiado el desarrollo de la IA

A lo largo de la historia de las redes neuronales computacionales, desde el artículo de McCulloch y Pitts (1943) pero en especial con el teorema de convergencia de Frank Rosenblatt (1957), se desarrolló una discusión sobre la dirección hacia la que deberían encaminarse los esfuerzos científicos. Una opción era concentrarse en que las máquinas pudieran aprender por sí mismas o, por decirlo de otra manera, que pudieran ir más allá de lo que el programador explícitamente les había ordenado, y proporcionar resultados 'inteligentes' a los problemas para los que fueron diseñadas, *aún si no podíamos saber o controlar exactamente cómo llegaron a esos resultados*; siguiendo la costumbre de muchos de estos mismos científicos, llamemos a este enfoque *programación neuronal*.

La otra opción proponía concentrarse en desarrollar maneras en que las máquinas pudieran *representarse adecuada y cabalmente la información del problema en cuestión*, aún si no la aprendían por ellas mismas; se le conoce como *programación simbólica*,[3] porque usaron lenguajes de programación simbólicos (LISP, por ejemplo) para implementar las representaciones en los sistemas de IA.[4] De hecho, las representaciones adecuadas eran consideradas un requisito previo a la posibilidad del aprendizaje, pues ¿cómo iban las máquinas a aprender algo que no podían representarse?

En principio, las dos opciones no eran irreconciliables. Lo que estaba en juego era saber cuál de ambas traería consigo un mayor progreso en la solución de los principales problemas de la inteligencia artificial. Había que decidir en cuál enfoque invertir financiamiento y esfuerzos intelectuales. Muchas revistas de investigación, libros, laboratorios, conferencias y programas de estudio e investigación manifestaban claramente en

[3] Para una caracterización más completa de ambos enfoques, véase Minsky & Papert (1988, pp. 274-275), Minsky (1968), Sejnowski (2018, pp. 27-42) y Darwiche (2018).

[4] Conviene tener presente que representación y simbolización no son lo mismo. Esta tesis se enfoca en el aspecto representacional del proyecto de investigación de *programación simbólica*. Para entender la parte simbólica de este proyecto McCarthy (1960) es una buena referencia. En ese artículo, McCarthy presentó el sistema de programación LISP y, con éste, las expresiones simbólicas y la manera en que las computa una máquina.

sus títulos y nombres hacia cuál lado se inclinaban. Del primer caso, vemos títulos de revistas como *Neural Computation*, y *Biological Cybernetics*; especialmente importante fue el programa *Neural Computation and Adaptive Perception (NCAP)* y los congresos *Neural Information Processing Systems* (*NIPS)*, cuyas historias relata Sejnowski (2018) en pasajes de varios capítulos de su libro (v.g., pp. 127-129, 161-167). El proyecto original de inteligencia artificial cuyo inicio se suele situar en el taller de Dartmouth, organizado por John McCarthy en 1956, estuvo asociado más a la programación simbólica; además, un congreso importante de este enfoque ha sido la *International Conference on Principles on Knowledge Representation and Reasoning*; una publicación que recaba varios trabajos realizados en esa dirección es Minsky (1968).

Este contraste entre dos maneras de perseguir los objetivos del proyecto científico de inteligencia artificial se ha transformado con el tiempo, pero no ha desaparecido. Hace cinco años (poco, si consideramos que esta discusión inició en los años cuarenta), Adnan Darwiche, antes director del departamento de Ciencias de la computación en la UCLA y ahora director del grupo *Automated Reasoning* de la misma universidad, escribió:

> Está emergiendo una tendencia en la que las investigaciones de aprendizaje automático se encasillan en investigaciones sobre las redes neuronales, bajo la recién adquirida etiqueta de 'aprendizaje profundo'. Esta percepción ha provocado que algunos se pregunten sobre la sensatez de continuar invirtiendo en otros enfoques de aprendizaje automático o, incluso, en otras áreas centrales de IA (como la representación del conocimiento, el razonamiento simbólico y la planeación). (Darwiche, 2018)

Actualmente, el vínculo entre el aprendizaje automático y el aprendizaje profundo como método preeminente ya no es una tendencia. Está establecido.

La confrontación entre programación simbólica y programación neuronal no resulta del todo adecuada para el contexto actual. De cualquier modo, hay dos tradiciones relativamente separadas: el aprendizaje automático que se basa en modelos y el aprendizaje automático que busca aproximar funciones. En las páginas siguientes, recupero las ideas

relevantes desde el taller de Dartmouth, que convencionalmente se toma como el inicio del proyecto científico de inteligencia artificial.

Marvin Misnky define la inteligencia artificial (IA) con las siguientes palabras: «la ciencia de hacer que las máquinas realicen cosas que requerirían inteligencia si las realizara un ser humano» (1968, p. v). Él fue uno de los redactores de la propuesta del taller de verano en Dartmouth, donde se considera que comenzó el proyecto científico de IA. Aun si es discutible, la idea de que la inteligencia humana es el referente claro para evaluar la IA guio inicialmente el proyecto.

En la propuesta del taller no hay una definición como tal; encontramos más bien la conjetura que estructuró al programa:

> El estudio procederá basado en la conjetura de que cada aspecto del aprendizaje o de cualquier otra característica de la inteligencia puede en principio ser descrita de un modo tan preciso que pueda construirse una máquina que lo simule (McCarthy *et al.*, 1956).

En el documento añaden diez aspectos de esta conjetura en los que proponen concentrarse. La conjetura nos muestra una segunda idea: el aprendizaje es la principal característica de la inteligencia (es la única que mencionan explícitamente). Para algunas tareas parecía imposible encontrar un programa con las instrucciones para que una máquina las realizara. En cambio, encontrar un programa que habilitara a la máquina para aprender por sí misma cómo realizar la tarea en cuestión parecía más factible.

Uno de los asistentes al taller de Dartmouth, Arthur Samuel, diseñó en 1952 una serie de programas que jugaban a las damas. Estos programas, tras un entrenamiento, aprendieron a jugar damas al nivel de torneo. Russell y Norvig escriben que Arthur Samuel «desacreditó la idea de que las computadoras sólo pueden hacer lo que se les ordena, pues su programa rápidamente aprendió a jugar mejor que su creador» (2021, p. 19). La máquina de Samuel aprendió a realizar una tarea para la que no fue explícitamente programada. En el código no había instrucciones como “si el oponente mueve la ficha a la casilla 20, entonces mueve la ficha a la casilla 13”.

El hecho de que la máquina jugara mejor que su creador se consideró una prueba de que ésta había realizado un aprendizaje.

Al año siguiente del taller de Dartmouth, Frank Rosenblatt presentó el perceptrón, junto con el teorema de la convergencia del perceptrón (Rosenblatt, 1957). El perceptrón es el primer autómata con una arquitectura neuronal. Su propósito es determinar si un input cumple o no con cierto predicado; por ejemplo, si en una fotografía cualquiera hay o no un semáforo. Sejnowski (2018) presenta una explicación resumida y adecuada de los perceptrones. Una exposición formal y detallada la encontramos en Minsky & Papert (1988).

El input es un vector de los valores $[x_1, x_2, \dots, x_p]$ de las variables $[X_1, X_2, \dots, X_p]$; usualmente se considera una constante $x_0 = 1$ para ajustar el eje del modelo (cuando se trata de imágenes, cada característica X_i es un pixel). A cada una de estas características le está asignado un peso que pondera su relevancia; de este modo se tiene un vector de pesos $[w_0, w_1, w_2, \dots w_p]$. Para calcular el resultado del perceptrón, se toma el producto del vector de los valores de las características y el vector de los pesos $(x^T w)$, y dado un umbral θ se determina si el ejemplo es positivo, o si es negativo, es decir, si el objeto del input presenta o no el predicado en cuestión. Uno o cero.

$$f(x) = \begin{cases} 1, & x^T w > \theta \\ 0, & x^T w < \theta \end{cases}$$

Este diseño plantea el problema de cómo hacer que el perceptrón escoja los pesos (w_i) que solucionan el problema en cuestión (que aprenda a partir de los datos). Si inicia con un valor aleatorio para cada peso, probablemente dará muchos resultados erróneos. Pero para aprender los pesos correctos no basta con saber si está equivocado o no, sino qué tanto contribuye al error cada uno de los pesos. Éste es un caso del problema más general conocido como el problema de la *asignación de crédito*.

Dada la simplicidad del perceptrón, vemos que la proporción entre el resultado xw^T y el peso w_i determina el crédito respecto del error. Esta es la idea clave del teorema de la convergencia del perceptrón; el teorema

establece que, con un número suficiente de ejemplos, si existe un conjunto de valores óptimos para los pesos, uno puede hacer que los pesos converjan en tales valores óptimos, y que el perceptrón funcione para casos nuevos.

El procedimiento consiste en modificar proporcionalmente los pesos según el éxito o el fracaso del algoritmo en la clasificación de cada ejemplo. Así, el cambio en los pesos (Δw_i) se determina por el índice de error $\delta \in \{0, 1, -1\}$ multiplicado por una constante de aprendizaje α:

$$\Delta w_i = \alpha \delta x_i$$

donde $\delta = predicción(x) - valor\ real\ (x)$.

Rosenblatt creyó que con su algoritmo había logrado construir un perceptrón que generalizara una tarea compleja: un autómata que clasificaba imágenes según contenían o no tanques de guerra. Cuando presentó su teorema, tuvo la sutileza de especificar que el algoritmo convergería en los pesos óptimos "si tal conjunto de pesos existe". Por desgracia, después se descubrió que dicho conjunto, para el perceptrón, existe en muy pocos casos. Resultó que su perceptrón para tanques, en realidad, no estaba clasificando los tanques, sino la hora del día (Sejnowski, 2018, p. 47). En concreto, dicho conjunto de pesos sólo existe para el perceptrón en casos linealmente separables.

Marvin Minsky y Seymour Papert escribieron un libro en 1969 para entender teóricamente las limitaciones y los alcances de los perceptrones, que se titulaba justo así, *Perceptrons* (aquí utilizo la segunda edición, que conserva la distinción entre lo que se escribió originalmente y lo que fue añadido o corregido: 1988). Uno de los casos más escandalosos fue la imposibilidad de que el perceptrón realizara la función XOR. Muchos han interpretado el libro de Minsky y Papert como un traspié para el proyecto neuronal. Sejnowski habla de Minsky como 'el diablo del enfoque neuronal' (2018, p. 258). Pero el propósito del libro era alcanzar una comprensión teórica clara sobre lo que es el perceptrón, explicar matemáticamente por qué puede resolver los problemas que resuelve —sobre todo, entender con cuánta eficiencia— y por qué no puede resolver los que no resuelve. En ese

sentido, el libro alentó y orientó para resolver los problemas que habían quedado sin solución (Minsky & Papert, 1988, p. xii).

El siguiente hito en la evolución de las redes neuronales fue *retropropagación.* Este algoritmo se puede entender como la versión multicapa del algoritmo de aprendizaje del perceptrón, siempre y cuando se advierta que éste no está garantizado para converger en el óptimo global del espacio de hipótesis. De cualquier modo, la alta dimensionalidad del espacio de hipótesis, junto con algunos métodos que imitan el momento de fuerza (según se entiende en física), hacen que en la práctica esa diferencia no constituya una preocupación real.

La llegada de este algoritmo, junto con la posibilidad tecnológica de implementarlo para resolver problemas reales (más que sólo teóricos, si nos permitimos esta distinción operativa entre 'práctico' y 'teórico'), habilitó a los científicos a utilizar redes neuronales con una gran capacidad y con varias capas, al grado de resolver los problemas que no pudo el perceptrón, y muchos más. Se desarrollaron capas convolucionales, con las que se resolvió el reconocimiento de imágenes (Hinton, 2012, y Goodfellow, 2016, cap. 9, en especial pp. 321-330), y también redes recurrentes que sirven para problemas relacionados con series de tiempo (cf. Chollet 2021, cap. 10), y muchos otros diseños neuronales. Todos estos desarrollos implementan retropropagación.

El siguiente es un recordatorio de la estructura matemática del algoritmo, basado en el artículo que presentó el algoritmo a la comunidad científica (Rumelhart *et al.*, 1986a).

Sea $i = 1, 2, \dots, n$ el índice de los ejemplos en la base de datos de entrenamiento. Y sea $k = 1, 2, \dots, m$ el índice de las capas de una red neuronal, cada capa con dimensión $r = 1, 2, \dots, p$. El índice $k = 0$ representa al vector de entrada. Tómese $x_i(k)$ como el estado del vector de datos en la capa k. Entonces, para cualquier ejemplo i, nos podemos referir a la información en la capa k como el vector $x_i(k) = \left(x_i^1(k), \dots, x_i^p(k)\right)$. Nótese que $x_i(0) = \left(x_i^1(0), \dots, x_i^p(0)\right)$ es el vector input, y que $x_i(m) = \left(x_i^1(m), \dots, x_i^p(m)\right)$ es el vector output. Normalmente, se tiene un nodo

$x_i^0(k) = 1$ en cada capa $k \neq m$ para introducir los sesgos, pero por razones de simplicidad no considero aquí esos nodos.

Ahora, sea $w(k)$ una matriz de coeficientes con dimensiones $p_k \times p_{k-1}$ para cada capa $k = 1, 2, \dots, m$, que conecta las capas $k - 1$ con sus sucesores k.

El primer paso en el algoritmo es conocido como *propagación*. En éste, cada vector $x_i(k) = \left(x_i^1(k), \dots, x_i^p(k)\right)$ se calcula (normalmente) como una función sigmoidal del producto $w(k)x_i(k-1)$, para $k = 1, 2, \dots, m$. De modo que

$$x_i(k) = S\{w(k)x_i(k-1)\}.$$

Estrictamente hablando, *retrorpopagación* es el segundo paso. Primero se calcula el error E de la siguiente manera:

$$E = \frac{1}{2}\sum_i \sum_{r=1}^{p_m} (x_i^r(m) - y_i^r)^2$$

donde $Y_i = \left[y_i^1, \dots, y_i^p\right]$ es el vector del output deseado, con tamaño $p = p_m$.

Después, para cada instancia I, tomamos la derivada de E respecto de cada unidad del output, y obtenemos:

$$\frac{\partial E}{\partial x_i^r(m)} = x_i^r(m) - y_i^r.$$

Por último, para cada capa $k = 1, \dots, m-1$, obtenemos la derivada $\frac{\partial E}{\partial x_i^r(k)}$ aplicando la regla de la cadena. Puesto que cada capa es una función linear de su predecesora, siguiendo este razonamiento podemos saber en qué dirección (creciente o decreciente) cada peso contribuirá a disminuir el error total E. Este segundo paso, las derivadas de E respecto de cada peso, resuelve el problema de la asignación de crédito para las redes neuronales multicapa.

El tercer paso consiste en actualizar los pesos en la dirección que minimice el error $b_i(k) = \frac{\partial E}{\partial x_i(k)}$ para cada instancia i y cada capa k, con una constante de aprendizaje α:

$$w(k) \leftarrow w(k) - \alpha\,(\cdot) \sum_{i=1}^{n} b_i(k)\,\nabla S\{w(k)x_i(k-1)\}\,w(k)x_i^T(k-1).$$

Ésta es sólo la estructura del algoritmo. En las implementaciones del algoritmo se requiere considerar muchos detalles más. Empero, es una sana práctica en la filosofía de la computación tener presentes las ideas matemáticas de las que estamos hablado.

Al tercer paso también se le conoce como *descenso gradiente*. Misnky y Papert advierten que, como una mera extensión de la idea de Rosenblatt mediante la regla de la cadena, el algoritmo de *retropropagación* no es tan asombroso como se piensa (1988, p. 260). El hecho es que, asombroso o no, permitió a los científicos de la programación neuronal resolver problemas que los perceptrones y la programación simbólica no habían podido. También es importante tener en cuenta que no fue nada trivial el crecimiento del poder computacional disponible.

El título del artículo de Rumelhart et al. (1986a) resulta sugerente: la traducción sería “Aprender representaciones retro-propagando los errores”. En una red neuronal, la matriz W de los pesos puede tomarse como una representación. Rosenblatt, Minsky y Papert lo consideran así. La representación que anula o reduce el error a un nivel aceptable se considera una representación adecuada. Pero ninguna de las representaciones de este tipo se corresponde con la manera en que el ser humano representaría conscientemente un fenómeno. Tampoco se corresponde con la manera en que el cerebro humano opera a un nivel fisiológico; es totalmente

implausible que el córtex esté implementando retropropagación (cf. Lillicrap *et al.* (2020), Hinton (2022), and Jeong *et al.* (2022)).

No tendría mucho sentido ubicar la separación de la inteligencia humana y la inteligencia artificial en algún suceso o descubrimiento específico. Evidentemente hay una gran brecha entre ambas. El punto es que, a pesar de ello, las investigaciones entorno a uno y otro tipo de inteligencia se han enriquecido mutuamente. Cuando los científicos de la computación diseñan un nuevo algoritmo de aprendizaje general, los neurocientíficos se preguntan si el cerebro humano podría estar implementando ese algoritmo. Y viceversa, cuando los neurocientíficos realizan algún descubrimiento, los científicos de la computación se preguntan qué pueden aprender de ello para diseñar mejores algoritmos. Algunos científicos, como Geoffrey Hinton, han hecho enormes aportes en ambos campos. Uno de los más grandes desarrollos del aprendizaje profundo —por mencionar un caso— son las redes convolucionales, cuya arquitectura imita el funcionamiento de la corteza visual humana (Hinton *et al.*¸ 2012).

A pesar de que sabemos aún muy poco sobre el cerebro humano y el sistema nervioso (hemos explicado muy pocas de las funciones que este sistema desempeña y probablemente no hemos identificado todas las funciones que desempeña) y de que el proyecto de IA para nada es un proyecto terminado, se puede usar el conocimiento que hemos adquirido en ambos campos de estudio. Los descubrimientos en sobre el sistema nervioso humano arrojan pistas sobre cómo nuestro organismo resuelve los problemas cognitivos que quisiéramos que los sistemas artificiales también resolvieran. Basarnos en esas pistas es, en cierto sentido, basarnos en la inteligencia humana para mejorar la inteligencia artificial. Aún si nuestro conocimiento sobre la inteligencia humana es limitado. Además, el hecho de que tales pistas sean útiles no implica que se deban implementar exactamente de la misma manera en los sistemas artificiales. De hecho, sería muy sorprendente que fuera posible. Un ejemplo claro de esta brecha es la discrepancia entre la manera en la que los humanos nos representamos el conocimiento y las representaciones con las que operan los sistemas de inteligencia artificial.

Una característica deseable de una representación humana es que podamos entender exactamente qué función desempeña cada uno de los elementos en los procesos de inferencia que realizamos con dicha representación. Para un humano no es posible observar la matriz de los pesos de una red neuronal suficientemente grande y entender qué función específica desempeña cada uno.

Nótese a dónde conduce la trayectoria que se ha delineado en esta sección. Primero, los científicos del proyecto de inteligencia artificial buscaron emular artificialmente la inteligencia humana. Concluyeron que emular el aprendizaje permitiría resolver todos o gran parte de la gama de problemas que se propusieron resolver artificialmente. En un segundo momento, se dividieron; la programación neuronal propuso concentrar la evaluación del aprendizaje en los resultados (outputs) de la máquina, mientras que la programación simbólica propuso concentrarse en que las representaciones y la capacidad representacional fueran adecuadas, porque resultaba difícil aceptar que una máquina pudiera resolver un problema sin poder esta misma representarse adecuadamente el problema.

Esta trayectoria resalta una pregunta íntimamente ligada al proyecto de inteligencia artificial: ¿cuál es el papel de las representaciones en la valoración de un comportamiento o un procedimiento inteligente? Otra manera de decirlo es la siguiente: ¿qué nos puede ofrecer una mejor representación?, la cual incluye la cuestión sobre la cognición humana. ¿Qué tipo de criterio es la cognición humana (lo que sabemos de ella) en la evaluación de las representaciones y los procesamientos inteligentes?

La presente tesis se enmarca dentro de dicha cuestión, pues compara dos tipos de representación: las que se adquieren mediante el *aprendizaje estadístico* y los *modelos causales*. La discusión gira en torno al tipo de cosas y relaciones que podemos representarnos, que podemos aprender, y también en torno al riesgo de incorporar información inadecuada en los modelos o describir cosas que no existen.

En secciones posteriores se aborda esta discusión. Será provechoso tener en mente la pregunta por la relación entre las representaciones y el conocimiento. Desde la perspectiva de los modelos causales estructurales probabilistas, se le puede objetar a la teoría del aprendizaje estadístico que

es incapaz de adquirir representaciones de relaciones causales. No obstante, bien se podría responder a dicha objeción que, del hecho de presentar un modelo cuantitativo preciso, no se sigue que el fenómeno representado exista ni que de hecho se comporte así.

III. Epistemología de la teoría del aprendizaje estadístico

Algunos científicos han sustentado matemáticamente la inferencia que se realiza en los algoritmos de aprendizaje automático. Esto es, han desarrollado formalmente la *teoría del aprendizaje estadístico*.

Quien ha sentado las bases para la *teoría del aprendizaje estadístico* es principalmente Vladimir Vapnik (2000). Recientemente, él ha trabajado en lo que llama la *teoría completa del aprendizaje estadístico*. Nótese que estamos hablado de dos teorías distintas; en la segunda, Vapnik ha añadido el adjetivo 'completa' para distinguirlas. La primera usa fuerza bruta (computacional y metodológicamente hablando) al momento de optimizar las hipótesis a partir de una gran cantidad de datos, mientras que la segunda utiliza estrategias para detectar invarianzas estadísticas y busca ser efectiva incluso cuando la cantidad de datos es menor. Puesto que la teoría en la que aún se basan las prácticas de aprendizaje automático en la industria y en la investigación científica es la *teoría del aprendizaje estadístico* (la primera teoría), esta es la que se toma en cuenta en este trabajo. Las referencias para hablar de las prácticas actuales de aprendizaje automático son Alpaydin (2021), Goodfellow *et. al* (2016), Lee (2019) y Chollet (2021). Quedará abierta la pregunta sobre qué tanto los argumentos que se aducen en esta tesis atañen también a la (segunda) *teoría completa del aprendizaje estadístico*.

Vapnik fue el mismo que desarrolló el algoritmo SVM (*Support Vector Machine*). La idea principal de este algoritmo es proyectar los datos a un espacio de mayores dimensiones y encontrar el hiperplano que mejor los separa. El desarrollo de SVM está estrechamente vinculado con el de la teoría del aprendizaje. Una característica especial de SVM es lo bien que generaliza: garantiza la mejor generalización posible dentro de la capacidad elegida, dados los datos disponibles.

En su libro, Vapnik describe también la manera en la que las habilidades de ajuste de las redes neuronales se pueden coordinar con las habilidades de generalización de las máquinas SVM. Las primeras fijan el intervalo de confianza y minimizan el riesgo empírico; las segundas fijan el riesgo empírico y minimizan el intervalo de confianza (Vapnik, 2000, p.

124). Dedica también unas páginas al reconocimiento de dígitos manuscritos a partir de la base de datos del Servicio Postal de E. U. A., y la base NIST. En ambos casos, el error de la máquina SVM fue igual al de las redes neuronales LeNet: 4% y 1.1%, respectivamente (cf. 2000, 147-154 y 172-174).

El pilar fundamental de la *Teoría del aprendizaje estadístico* es el principio ERM: *Empírical Risk Minimization Inductive Principle*, esto es, el principio inductivo de la minimización del riesgo empírico. Este principio surge de la distinción entre el *riesgo funcional*, $R(\alpha)$, y el *riesgo empírico*, $R_{emp}(\alpha)$. El primero mide qué tanto falla nuestra hipótesis si la examinamos a partir de la distribución de probabilidad real del fenómeno que nos interesa (en inglés, '*the actual probability distribution*'). Se define así:

$$R(\alpha) = \int L(y,\ f(x,\alpha))\ dF(x,y)$$

donde 'L' es la función de pérdida, que mide la discrepancia entre la hipótesis $f(x,\alpha)$ y el caso real y. La función f relaciona las características x con los parámetros α, y F es la medida de probabilidad real del fenómeno (cf. Vapnik, 2000, p. 18).

De este modo, si uno consigue una función f y unos parámetros α, tales que $R(\alpha) = 0$, *se puede decir que f y α están capturando cabalmente el fenómeno*.[5] Sin embargo, en la mayoría de los casos, no conocemos la medida de probabilidad *F(x, y)*, puesto que nuestros datos sólo son una muestra del fenómeno que nos interesa. Vapnik describe el problema de reconocimiento de patrones del siguiente modo:

> El problema, por lo tanto, consiste en encontrar una función que minimiza la probabilidad de clasificar erróneamente cuando no conocemos la medida de probabilidad *F(x, y),* pero se nos han dado los datos. (2000, p. 19).

Esto implica que, al no poder reducir el riesgo funcional (por falta de información sobre la medida de probabilidad), lo mejor que podemos hacer es reducir el riesgo a partir de lo que sabemos gracias a nuestros

[5] ¿Se puede? Esta afirmación es parte central de la discusión epistemológica.

datos. Ya que los datos se nos han dado como información del fenómeno en cuestión, al riesgo acotado por la información de estos datos se le conoce como *riesgo empírico*, y se define de la siguiente manera:

$$R_{emp}(\alpha) = \frac{1}{l}\sum_{i=1}^{l}(y_i - f(x_i, \alpha))^2$$

Aquí l es el número de ejemplos de la base de datos disponible.

El principio ERM es precisamente la idea de que podemos aproximarnos a la función que minimiza el riesgo funcional, mediante la función que minimiza el riesgo empírico (cf. 2000, p. 21).

Ahora bien, para que el principio sea consistente es necesario encontrar una manera de garantizar que la minimización del riesgo empírico en verdad minimizará el riesgo funcional. Al teorema que muestra esa consistencia Vapnik lo llama el *Teorema clave de la teoría del aprendizaje*:

> **Teorema 2.1**. Sea $Q(z, \alpha), \alpha \in \Lambda$, un conjunto de funciones que satisface la condición
>
> $$A \leq \int Q(z, \alpha)dF(z) \leq B \qquad (A \leq R(\alpha) \leq B).$$
>
> Entonces, para que el principio ERM sea consistente, es necesario y suficiente que el riesgo empírico $R_{emp}(\alpha)$ converja uniformemente al riesgo actual $R(\alpha)$ sobre el conjunto $Q(z, \alpha), \alpha \in \Lambda$, de la siguiente manera:
>
> $$\lim_{\ell \to \infty} P\{\sup_{\alpha \in \Lambda}\left(R(\alpha) - R_{emp}(\alpha)\right) > \varepsilon\} = 0, \quad \forall \varepsilon > 0.$$
>
> (Vapnik, 2000, p. 38).

$z = \{z_1, z_2, \ldots, z_\ell\}$ es un conjunto de observaciones independientes e idénticamente distribuidas (i. i. d.). Vemos que, en los casos en los que converge, mientras más ejemplos tengamos, más se reducirá el riesgo funcional mediante el riesgo empírico. También vemos que el teorema considera la convergencia de la peor función del conjunto $Q(z, \alpha)$, puesto que la peor función debe cumplir con la convergencia (cf. *ibidem*).

Así que se pueden notar ya los dos aspectos centrales de la teoría del aprendizaje:

1. *Aproximación*: ¿cómo encuentro la función que mejor se aproxima a mis datos? (Reducir el riesgo empírico).
2. *Generalización*: ¿Cómo garantizo que mis resultados o aplicaciones serán pertinentes para la distribución de probabilidad real? (Reducir el intervalo de confianza).

Vapnik explica que la teoría del aprendizaje (las ideas que aquí presenté, junto con las demás de su libro y los métodos que también encontramos allí) presenta un cambio de enfoque respecto de la manera clásica de comprender la estadística:

> El enfoque clásico para estimar dependencias funcionales multidimensionales se basa en la siguiente creencia:
>
> *Los problemas de la vida real son tales que existe un pequeño número de 'características fuertes', cuyas funciones simples (por ejemplo, combinaciones lineales) aproximan bien la función desconocida. Por lo tanto, es necesario escoger cuidadosamente un espacio de bajo nivel dimensional para las características, y luego usar las técnicas de estadística comunes para construir una aproximación.*
>
> [...] La nueva técnica se basa en una creencia distinta:
>
> *Los problemas de la vida real son tales que existe un gran número de 'características débiles' cuya 'perspicaz' combinación lineal aproxima bien la dependencia desconocida. Por lo tanto, no es tan importante qué tipo de 'característica débil' uno use; es más importante formar combinaciones lineales 'perspicaces'.*
>
> (Vapnik, 2000, p.177).

La intención que muestra este cambio de enfoque consiste en hacer que la parte formal del procedimiento sea la parte decisiva, y que la parte informal sea relativamente negligible. Las redes neuronales pertenecen a la 'nueva técnica', y los modelos causales están más cercanos al 'enfoque clásico', a pesar de que no se corresponden del todo con éste.

Una de las motivaciones para desarrollar algoritmos de razonamiento causal es poder tratar con un gran número de variables y de

relaciones, pero éstas tienden a ser menos que la cantidad de nodos y pesos de las redes neuronales. El razonamiento causal descarta las variables irrelevantes.

Quizás la 'nueva técnica' ofrece un mejor ajuste a los datos (reduce el riesgo empírico en mayor medida). Pero, desde un punto de vista epistemológico, nos hace perder comprensión. Se ajusta mejor a los datos porque utiliza funciones sumamente complejas para capturar la estructura *de los datos*. La razón por la que nos hace perder comprensión es que, al incluir un gran número de 'características débiles', nos impide distinguir la relevancia de cada característica y su relación con el resultado (tanto con el resultado predicho como con el resultado observado). La estructura de los datos no siempre se corresponde con la estructura del fenómeno o del proceso que se está analizando.

Los seguidores del nuevo enfoque se alejan del clásico con el propósito de conseguir más exactitud y precisión. Es difícil no querer dar este giro hacia el nuevo enfoque, al hacer que los resultados dependan de métodos formales, estamos consiguiendo una manera de responder formalmente las preguntas estadísticas que nos ocupan.

En los ejemplos que ofrece Vapnik sobre el reconocimiento de dígitos manuscritos podemos ver que, aunque los nuevos métodos se apoyen principalmente en los procedimientos formales, la parte informal también puede mejorar los resultados. Por ejemplo, vemos cómo los científicos de las redes neuronales utilizaron la base de imágenes para generar imágenes nuevas, distorsionando las originales sin que éstas dejaran de representar un número. Esta manera de generar datos nuevos es una técnica conocida en el enfoque neuronal. Pero pertenece a la parte 'informal', pues estamos asumiendo conocimientos previos a la base de datos. Ya que se trata de que la máquina aprenda qué es un número y cuál número es cuál, estamos dándole *información extra* al asumir que podemos girar un poco un número sin que éste deje de ser lo que es.

Vapnik incluso resalta el hecho de que la máquina SVM resolvió el problema sin presuponer nada sobre la geometría del problema, e invita a encontrar una manera de conseguir los mismos resultados que las redes neuronales que se entrenaron con la base de datos aumentada, pero

recurriendo a procedimientos más generalizables (cf. Vapnik, 2000, pp. 173-174). De hecho, lo que proponía cuando escribió ese libro era buscar la manera de que la máquina detecte por sí misma las invarianzas que permitieron a los del enfoque neuronal ensanchar la base de datos, y que realice el aprendizaje contemplando tales invarianzas. Como se mencionó párrafos atrás, esa es justo la idea que guio su más reciente *teoría completa del aprendizaje estadístico* (cf. Vapnik e Izmailov, 2020).

Del mismo modo en que las herramientas no son neutras, las teorías que guían las aplicaciones de ingeniería tampoco lo son. Uno puede usar una escopeta para algo bueno; por ejemplo, para trabar una puerta que, de no hacerlo, machucaría a alguien. También se puede usar un cepillo de dientes para destruir los ojos de una persona. La moralidad de las herramientas no se debe a sus posibilidades, sino a sus *tendencias*. Una vez que se construye una bomba nuclear, difícilmente podemos pensar en otro propósito que no se derive de la matanza masiva de personas. Si funcionan como contrapeso en la política internacional, es porque se construyeron para matar gente.

Similarmente, no quisiera discutir la 'verdad' de la teoría del aprendizaje estadístico. Sabemos que funciona. Lo que debe ocupar a la filosofía de la ciencia —en este caso, entre otras cosas— es la tendencia hacia ciertos supuestos epistémicos que instauran las diversas teorías del aprendizaje.

La teoría del aprendizaje estadístico, dado que el riesgo empírico converge en el riesgo funcional conforme crece el número de ejemplos, inclina a pensar que siempre es una buena idea conseguir más datos. A pesar de que algunos algoritmos, como SVM, implícitamente distinguen entre datos cruciales (los vectores de soporte) y datos triviales, éstos no garantizan procedimientos para concentrar la búsqueda sólo en los datos cruciales.

El hecho de que la aproximación a la función que reduce el riesgo funcional (una integral) se realiza mediante la función que reduce el riesgo empírico (una sumatoria) inclina hacia un segundo presupuesto: no hay diferencia entre conocer la totalidad de los casos y la estructura del

fenómeno. Dicho con otras palabras, el grado máximo de adecuación empírica se consigue encontrando la función que, dado un valor de las variables explicativas, se obtiene el valor de la o las variables dependientes (por supuesto, se suele aceptar un error de tamaño e que se distribuye según una densidad determinada $E \sim f(x; \theta)$).

De nuevo, el propósito no es poner en duda el TEOREMA 2.1. A fin de cuentas, esa es una manera de definir a las integrales: son sumatorias infinitas cuyos intervalos tienden a cero. El propósito es advertir las creencias (que no forman parte estrictamente de la teoría) a las que la teoría y sus aplicaciones nos inclinan. Si se interpreta la reducción del riesgo empírico como una adquisición de conocimiento, se está presuponiendo que el conocimiento se reduce a la adecuación empírica.

En resolución, se han mostrado dos presupuestos de la teoría del aprendizaje estadístico que, incluso, podrían considerarse como ventajas epistémicas. El primero nos dice que tenemos una opción a la que siempre podemos recurrir para optimizar la solución de cualquier problema definido dentro de la teoría del aprendizaje estadístico: recolectar más datos. El segundo nos sugiere que no es necesario buscar algún tipo de conocimiento más 'profundo'; una vez que minimizamos el riesgo funcional con un intervalo de confianza aceptable, sabemos todo lo que hay que saber.

Las secciones siguientes, ofrecen una refutación para ambos supuestos: hay mejores maneras de perfeccionar nuestro conocimiento, y hay más cosas por saber.

IV. Motivaciones de los métodos de inferencia causal

En esta sección se consideran las motivaciones por las que se ha buscado desarrollar métodos de inferencia y razonamiento causal; estas consideraciones aclaran dudas sobre el alcance, la pertinencia y las pretensiones de estos métodos. Se pueden identificar al menos cuatro ámbitos en los que los métodos causales se proponen para responder a necesidades o para satisfacer ciertos *desiderata*: computacional, filosófico, bayesiano y científico.

La motivación computacional

Históricamente, la programación neuronal surgió con el propósito de resolver problemas que no se podían resolver al modo de la programación simbólica.

Las dificultades aparecen cuando las relaciones del modelo que se busca construir no son tan rígidas o deterministas como se esperaría en un contexto de reglas deductivas. En la conferencia *The Art and Science of Cause and Effect*, impartida en 1996, y recuperada como epílogo en Pearl (2009, pp. 401-428), Judea Pearl habla de "la causalidad como la pesadilla del programador", y utiliza el siguiente ejemplo:

> **Input:** 1. "Si el pasto está mojado, entonces llovió".
> 2. "Si rompemos esta botella, el pasto se mojará".
>
> **Output:** "Si rompemos esta botella, habrá llovido".
>
> (Pearl, 2009, p. 414.)

Con este ejemplo vemos que las proposiciones condicionales, a pesar de capturar cierta asimetría, no guardan una relación intrínseca con la noción de causalidad. Naturalmente, surge la pregunta ¿entonces qué tipo de herramienta formal, si existe alguna, resulta adecuada para tratar la causalidad?

Para plantear el problema, Pearl propone imaginarnos a un robot tratando de interactuar en una cocina o en un laboratorio.

«Conceptualmente», nos dice, «los problemas del robot son los mismos que aquellos que enfrenta el economista tratando de modelar la deuda nacional, o los de un epidemiólogo tratando de entender la diseminación de una infección» (Pearl, 2009, p. 415). Los tres necesitan capturar relaciones causa-efecto en un entorno «usando acciones limitadas y observaciones ruidosas» (*ibidem*).

Una de las principales fuentes de reflexión para Pearl es David Hume. Sabemos que Hume puso en duda que la contigüidad implicara causalidad. De acuerdo con la lectura que hace Pearl de este filósofo, Hume plantea dos enigmas de la causalidad, y al programar robots uno se enfrenta a ambos enigmas.

El primero: «¿cómo es que las personas adquieren, si es que lo hacen, conocimiento causal?» (Pearl, 2009, p. 406). El canto del gallo no causa el amanecer; en cambio, el contacto con el fuego sí causa quemaduras. Pearl añade que es difícil creer que Hume no se daba cuenta de ello. Este primer enigma se traduce, dentro del ejemplo del robot, del siguiente modo: ¿cómo hacer que, mediante acciones limitadas y observaciones ruidosas, un robot capte relaciones causa-efecto? ¿Cómo programarlo para que note que, si por una acción suya le cae una chispa de fuego a un frasco con triyoduro de nitrógeno, el frasco explotará *debido* a su propio descuido? (cf. Pearl, 2009, 413).

El segundo enigma se pregunta qué es lo que añade la afirmación de que una relación es causal (cf. *ibidem,* p. 407). Una vez que sabes que dos hechos no sólo coinciden, sino que uno es causa del otro, ¿qué consecuencias conlleva esa diferencia? Porque, si no hubiera consecuencias, no tendría siquiera significado la noción de causalidad (cf. *ibidem*, p.407).

En el caso del robot,

> Supongamos que deseamos tomar un atajo y enseñarle a nuestro robot todo lo que sabemos sobre las causas y efectos en esta habitación. ¿Cómo debería organizar y usar el robot esta información? (Pearl, 2009, p. 413.)

Enseguida, Pearl añade que así los dos enigmas filosóficos de la causalidad se ven traducidos en problemas de aplicación práctica.

Motivaciones filosóficas (marco filosófico general)

Gran parte de las discusiones en torno a la causalidad en la filosofía de la ciencia del siglo XX giran en torno a la posibilidad —y a las maneras posibles— de distinguir entre causas genuinas y meras relaciones de dependencia. Entre las propuestas más notables se encuentra el proyecto de causalidad probabilista. Las relaciones causales genuinas se dan cuando la dependencia permanece incluso si se consideran los factores relevantes (si se condiciona sobre ellos). Suppes (1970) habla de relaciones causales *prima facie*, espurias, directas e indirectas, todas ellas se definen por relaciones que cambian o permanecen cuando se agregan o se eliminan variables condicionantes.

La distinción entre causas espurias y causas genuinas se puede rastrear hasta el pensamiento de Hume. Además de que Hume presupone que la necesidad es una característica intrínseca de la causalidad (como se mencionó en la primera sección), cara a la discusión contemporánea hay otro supuesto de Hume que conviene notar para una mejor comprensión filosófica: Hume presupone una teoría de la mente en la que toda nuestra experiencia es trazable a impresiones simples. William James criticó esta teoría como 'la teoría de la mente polvo' (James, 1912, p. 43);[6] él asocia metafóricamente a las impresiones simples con motas de polvo, y propone que las relaciones también son un constitutivo primario de nuestra experiencia. Más allá del debate sobre la experiencia humana, adviértase que, si todo se descompone en impresiones simples y la causalidad es una relación de conexión necesaria entre impresiones o ideas (originadas a partir de las impresiones), entonces todas las relaciones causales son cualitativamente iguales entre sí.

Cartwright señala, precisamente, que uno de los *desiderata* de una comprensión filosófica adecuada de la causalidad es que dé cuenta de la pluralidad de (tipos de) causas que hay en el mundo: cuando hablamos de

[6] En Hume (1772, §58) leemos «de modo que, en la totalidad, no se presenta, a lo ancho de toda la naturaleza, una sola instancia de conexión que sea concebible por nosotros. Todos los eventos lucen enteramente sueltos y separados. Un evento sucede a otro, pero nunca podemos observar vínculo alguno entre ellos».

causas, unas empujan, otras atraen, o desalientan, nutren, permiten, informan, abren, aceleran, succionan y un largo etcétera (Cartwright, 2007, p. 20). Unas causas preceden a su efecto y se distancian de él, como las bolas de billar. Otras ejercen una influencia continua, como la tierra fértil que envuelve a las raíces de las plantas.

Si queremos un método para inferir causas y razonar con información causal, éste debería adecuarse de manera aceptable a la pluralidad de las causas. En la opinión de Cartwright, ésta es una de las razones por las que el método de Pearl no es un método universal.

De hecho, no lo es, en el sentido de que no lo podemos aplicar ciegamente y confiar en que traerá buenos resultados por sí mismo. Como con cualquier método de inferencia, hemos de contar con premisas verdaderas y con las condiciones de inferencia adecuadas. El método exige de nuestra parte conocimiento previo sobre el fenómeno que queremos estudiar, para saber si es válido aplicarlo y cuál sería la manera propicia de hacerlo.

Al ser un método formal, completo, consistente y transparente en cuanto a la manera de efectuar razonamientos y de incorporar conocimiento previo, permite que experimentalmente advirtamos sus aciertos y sus limitaciones, y que podamos afinarlo o desarrollar otros mejores a partir de lo aprendido. Desde una perspectiva filosófica, esto significa más experiencia para desarrollar mejores teorías y una mejor comprensión de lo que es la causalidad. Aunque el modelo no sea perfecto.

La motivación bayesiana

La genialidad del teorema de Bayes, más que en la igualdad que establece, se encuentra en la manera en que captura formalmente una idea comúnmente aceptada pero, a la vez, difícil de precisar. El teorema nos dice cómo obtener una creencia actualizada a partir de evidencia nueva considerando también el conocimiento o las creencias de las que previamente disponíamos. Gran parte de la genialidad se encuentra en el hecho de que la creencia actualizada, en vez de resumirse a la suposición de algún valor para las

variables consideradas, consiste en una nueva distribución de probabilidad.

Actualmente podemos hablar de la epistemología bayesiana y de la inferencia bayesiana como campos definidos de la filosofía y la estadística. Uno de los problemas que más ha recibido atención dentro de la epistemología bayesiana es la construcción de una medida de coherencia. En términos muy generales, con la coherencia sucede algo similar a lo que pasa con la idea de incorporar conocimiento previo y evidencia nueva en un mismo proceso de razonamiento: la coherencia es una noción fácil de entender pero difícil de medir.

Piénsese que lo mismo sucede con la causalidad: es fácil para un humano adulto entender que el sol calienta la tierra, que el taco golpea la bola de billar y que la tierra nutre las raíces. No obstante, es muy difícil capturarlo formalmente.

Usemos el caso de la coherencia como ejemplo de los resultados que se pueden obtener desde la epistemología bayesiana. La pregunta es, dados dos conjuntos, $S = \{R_1, R_2, \dots, R_n\}$ y $S' = \{R'_1, R'_2, \dots, R'_n\}$, de proposiciones reportadas por testigos con cierto grado de credibilidad, ¿cuál es más coherente? El grado máximo de coherencia ocurre cuando todas las proposiciones se implican una a la otra; el grado mínimo, cuando cualquier subconjunto de dos o más proposiciones tiene probabilidad igual a cero. Los casos interesantes están en medio.

La idea central de Hartmann y Bovens (2003, p. 30) es medir la coherencia considerando centralmente el papel que ésta juega en nuestras creencias: el papel de aumentar nuestra confianza en lo que se nos informa. En términos generales, decidieron medir una propiedad epistémica a partir de la función que ésta desempeña en nuestros razonamientos. La medida a la que llegaron es la siguiente:

$$C_r(\{R_1, \dots, R_n\}) = \frac{b(\{R_1, \dots, R_n\})}{b^{max}(\{R_1, \dots, R_n\})}$$

Donde C_r es la medida de coherencia; las R_i son las proposiciones reportadas, b es la función que mide el incremento (*boost*) de confianza y

b^{max} es el incremento de confianza si el conjunto $\{R_1, \ldots, R_n\}$ fuera máximamente coherente. El subíndice r en C_r sirve para recordar que la medida depende del nivel de credibilidad de los testigos, medido del 0 al 1 ($r \in [0,1]$). No aparece explícitamente en la fórmula, pero el parámetro r es un argumento de la función b.

En su análisis, Hartmann y Bovens llegan a conclusiones nada obvias. Quizás la más importante es su resultado de imposibilidad: no es posible, bajo los compromisos del coherentismo bayesiano, construir una medida de coherencia completa. Esto es, una medida que establezca un orden completo; que para cualesquiera dos conjuntos de proposiciones determine cuál comporta mayor coherencia o si son exactamente igual de coherentes. La que proponen ellos es una cuasimedida. Por razones de simplicidad, no explicaré este punto y usaré como ejemplo otro resultado.

Hartmann y Bovens evalúan la tesis de Duhem-Quine, según la cual, al poner a prueba una hipótesis, cuando los resultados se oponen a la hipótesis, una opción es rechazar la teoría auxiliar con la que fue diseñado el experimento, en lugar de descartar la hipótesis. Ha de considerarse que la teoría auxiliar muchas veces no es independiente de la hipótesis. Hartman y Bovens, tras aplicar su modelo, encuentran que cuando la probabilidad anterior de la hipótesis $p(h)$ es baja, la evidencia que se obtiene con teorías auxiliares dependientes implica un mayor incremento en la probabilidad posterior $p(h|e)$. En cambio, cuando la probabilidad anterior es alta, entonces la evidencia de las teorías auxiliares independientes es la que proporciona un mayor incremento en la probabilidad posterior. (Bovens y Hartmann, 2003, p. 111.)

Más aún, con suficiente información sobre los escenarios (vg. la credibilidad de los instrumentos, la cantidad de variables en el conjunto de la evidencia) es posible determinar con precisión un umbral para la probabilidad anterior que separe los casos en los que las teorías auxiliares independientes aportan más confirmación de los que no (Bovens y Hartmann, p. 109). Hallamos una prueba de que los modelos formales están contribuyendo a nuestra comprensión epistemológica cuando nos conducen a conclusiones no obvias que de otro modo no conseguiríamos, y nos permiten medir dichas conclusiones con precisión.

Ahora veamos cómo todo esto es relevante para el razonamiento causal. Judea Pearl es conocido, además del desarrollo de los métodos causales, por sus aportes al razonamiento probabilista en la IA. Fue él quien acuñó el término 'red bayesiana' para referirse a los grafos que describen las relaciones entre variables.[7]

Una idea central para pensar que las redes bayesianas son un marco adecuado para abordar problemas causales es que hay una relación entre el hecho de que dos variables sean dependientes y la estructura causal en la que están inmersas, a pesar de que hay muchos casos en los que las variables son dependientes, pero no están relacionadas causalmente o —también es posible— casos en los que las variables están relacionadas causalmente pero probabilistamente son independientes. La limitación está en que las redes bayesianas no capturan la asimetría de la dirección causal. Como escribe Pearl:

> Con las redes bayesianas, les habíamos enseñado a las máquinas a pensar con tonos grises, y este es un paso importante hacia un pensamiento parecido al humano. Pero no podíamos enseñarles a las máquinas a entender causas y efectos. No podíamos explicarle a una computadora por qué girar la aguja del barómetro no causa la lluvia. [...] Sin la habilidad de visualizar realidades alternativas y contrastarlas con la realidad actual, una máquina no puede [...] responder la pregunta más básica que nos hace humanos: "¿Por qué?". Consideré que esto era una anomalía porque no anticipaba que tales preguntas naturales e intuitivas residieran más allá del alcance de los sistemas de razonamiento más avanzados de ese tiempo. (Pearl, 2018, p. 349.)

Así que, a pesar de que eran un buen marco, las redes bayesianas estaban lejos de ser suficientes para responder preguntas sobre la interacción entre variables. Pearl advierte que esa anomalía no sólo estaba presente en los sistemas artificiales, sino en los razonamientos de los científicos mismos:

---

[7] Acuñó el término en (Pearl, 1985); en *Causality* Pearl explica que hay tres aspectos que buscaba enfatizar con ese nombre: (1) la naturaleza subjetiva de la información input, (2) que el condicionamiento bayesiano se toma como base para actualizar información, y (3) la distinción entre dos tipos de razonamiento, el razonamiento a partir de la evidencia y el razonamiento causal; Pearl afirma que esta tercera distinción ya estaba presente en el artículo de 1763 de Thomas Bayes.

> Sólo después me di cuenta de que la misma anomalía estaba afligiendo más que a sólo el campo de la inteligencia artificial (IA). Las personas que precisamente deberían preocuparse más por las preguntas "¿por qué?" —a saber, los científicos— estaban trabajando bajo una cultura estadística que les negaba el derecho de formular tales preguntas. (Pearl, 2018, pp. 349-340.)

Hay una similitud metodológica entre el proceder de Bovens y Hartmann y el de Pearl y sus colaboradores. Estos segundos, al igual que los primeros, abordan la causalidad centrados en la función epistémica que ésta juega en nuestros razonamientos: la función de predecir y explicar intervenciones y escenarios alternativos.

Podría pensarse que es una coincidencia superficial, pues la coherencia es una noción plausiblemente evaluable en términos (meramente) probabilistas. Contrariamente, la causalidad es una noción que excede al vocabulario probabilista. El fundamento de los métodos causales, sin embargo, es probabilista y su uso conlleva consecuencias probabilistas.

Los métodos de Pearl comparten las que se conocen como las dos normas nucleares de la epistemología bayesiana: *probabilismo* y el *principio de condicionalidad*. La primera establece que nuestros grados de creencia en las diversas posibilidades deben embonar juntos de tal manera que no sean negativos y sumen 1 entre todos; la segunda, que, dada una evidencia E, las creencias incompatibles con E han de caer a cero, y las restantes han de escalarse de tal modo que sumen 1 (cf. Lin, 2022).

Nótese que la familiaridad con la epistemología bayesiana no se reduce a esas dos normas. Las investigaciones relacionadas con el trabajo de Pearl comparten el objetivo de expresar formalmente cualidades epistémicas que son fácil de entender en sentido general, pero difíciles de capturar en sentido formal. Queremos que los modelos que capturan estas nociones epistémicas se adecuen a nuestros razonamientos en los casos en los que tenemos clara cuál es la manera correcta de razonar, y que nos proporcionen respuestas en los casos que no son evidentes para nosotros. No sólo eso: que entendamos por qué en los casos no evidentes su procedimiento es adecuado.

No es que podamos subsumir a las investigaciones del proyecto de causalidad estructural probabilista en la epistemología bayesiana. Pero están presentes motivaciones bayesianas.

Motivaciones científicas

Hay muchas preguntas a las que se enfrentan los científicos que, sin un razonamiento causal correcto, no podrían responder.

Pearl (2018, pp. 53-92) relata las tensiones en torno a la causalidad que han estado presentes en la historia de la estadística. En especial, la manera en que afectó a este campo de estudio el que Pearson rechazara la causalidad como si ésta fuera una noción inválida o trivial (Pearl, 2018, pp. 66-72).

En las ciencias macroscópicas (economía, epidemiología, ciencias ambientales, sociología, psicología, etc.) nos encontramos con preguntas como: ¿cuál es la relación entre la pobreza y la educación, cómo influye una a la otra?; ¿las emisiones de carbono provocan desastres climáticos?; ¿qué pasaría si el gobierno invierte menos en seguridad y más en educación? Las tres son preguntas causales.

En el libro de Pearl (2018), leemos varios casos en los que diversos científicos que buscaban responder preguntas causales realizaron razonamientos tortuosos con tal de evadir inferencias causales, o enmascaraban ideas causales con vocabulario meramente asociativo, para no lucir poco rigurosos ante la causalidad científica.

Los siguientes tres ejemplos permiten dar cuenta de dicha idea.

El primero es el trabajo de Udny Yule (18 de febrero de 1871 – 26 de junio de 1951) en torno al (posible) efecto causal que los programas de asistencia social ejercen sobre el pauperismo. Por simplicidad, llamémosle 'subsidios' a la asistencia social. Yule colaboró con Pearson un tiempo y después, por algunas diferencias entre ellos, se distanciaron; Pearl (2018, pp. 66-72) ahonda en dicha historia.

La pregunta sobre el efecto causal de los subsidios estaba enmarcada en Inglaterra. Los datos provenían de las uniones a las que se suministraba la asistencia (por ejemplo, unas de ellas eran uniones de

agricultores). Utilizo la palabra 'pauperismo', que es una traducción directa del inglés *pauperism*; no es la más precisa, pero a falta de una mejor palabra, nótese que el significado que le da Yule es *el porcentaje de personas que reciben algún tipo de subsidio* (Yule, 1899, p. 252).

Yule clasificó a las uniones según su densidad de población en rurales, mixtas, urbanas y metropolitanas. Primero obtuvo el cambio promedio del pauperismo en los casos en los que la proporción de los subsidios se había mantenido igual a través del tiempo. Después, substrajo este cambio promedio a los casos que presentaban un cambio en la administración de subsidios a través del tiempo. Infirió entonces que el cambio restante se debía a las variaciones en la proporción de los subsidios. En sus palabras:

> Tomando la tabla de frecuencia de la proporción entre pauperismo y subsidios, encontrar el cambio promedio en el pauperismo de las uniones en las que no había un cambio significativo en la proporción de subsidios. Esto dará el cambio en el pauperismo que no se debe a una alteración en la administración y, por sustracción de este cambio promedio en todas las uniones, la porción que se debe al cambio en la administración. (Yule, 1899, p. 270.)

La afirmación de que dicha porción *se debe* al cambio de administración (un incremento o disminución en los subsidios) es una afirmación causal. Consciente de ello, Yule añade una nota al pie: «Estrictamente, en lugar de 'se debe a', léase 'se asocia con'» (*ibidem*). Sabía que transgredía la epistemología permitida por la estadística de su tiempo. Pearl (2018, p.72) usa este mismo ejemplo para explicar sus motivaciones para desarrollar los métodos de inferencia causal.

El segundo ejemplo es la 'fórmula del ajuste en la puerta frontal' (*front-door adjustment formula*). No entraré en detalles sobre la fórmula. En general, es una fórmula que, dadas ciertas condiciones causales, permite estimar el efecto causal de una variable sobre otra a partir de información no-experimental. Lo que quisiera resaltar es lo siguiente: con las reglas del cálculo causal (*do-calculus*), esta fórmula se puede obtener en siete renglones (siete pasos: dos aplicaciones de los axiomas de probabilidad y cinco de las reglas del cálculo causal); en contraste, la prueba más corta que se conoce con métodos de estadística 'tradicionales' requirió ocho páginas (cf. Heckman & Pinto, 2014). Así que, incluso en problemas solubles para

los métodos 'estándares', la eficiencia que se consigue con el cálculo causal es sumamente provechosa.

El tercer ejemplo es un caso en el que se han aplicado de manera efectiva los métodos causales desarrollados por Pearl. En 2016, Hannart *et al.* presentaron una propuesta para atribuir influencia causal a factores relacionados con el clima. En dicho contexto, más que probar una asociación entre las emisiones de carbono y las olas de calor (por mencionar una instancia), se busca una prueba de que las emisiones de carbono influyeron causalmente en las anomalías climáticas (véase Hannart *et al.*, 2016). De ese modo, se puede sustentar la exigencia de acciones legales, sociales y políticas frente a las autoridades.

Sirvan estos ejemplos para defender la afirmación de que el razonamiento causal, más que un problema enmarcado en lo computacional, en lo filosófico o en la teoría estadística, es un razonamiento necesario (ejemplos 1 y 3) y provechoso (ejemplos 2 y 3) para una gran parte de la actividad científica. El razonamiento causal fomenta que la actividad científica no esté basada en una epistemología que se limite a la recolección de observaciones, sino en una que se aboque además a la comprensión de la estructura de los fenómenos.

V. La matematización de la causalidad

Los métodos de inferencia causal que presenta Pearl (2009) son una herramienta que contempla la estadística, pero que va más de lo que estrictamente permitiría la inferencia estadística. Para aclarar este punto, recuérdese brevemente en qué consisten la probabilidad y la estadística como campos de estudio.

La probabilidad «permite modelar ciertos fenómenos que ocurren en la naturaleza, siendo el modelo básico un espacio de probabilidad (Ω, **F**, **P**) y una variable aleatoria X definida en ese espacio» (Fuentes García *et. al.*, 2019, p. 2).[8] Por otra parte, la estadística es la «rama de la matemática que utiliza conjuntos de datos para obtener inferencias basadas en el cálculo de probabilidades» (Fuentes García *et. al.*, 2019, p. 2). La diferencia está en que, cuando se trata de probabilidad, contamos ya con el espacio de probabilidad y la variable aleatoria, esto es, tenemos el modelo; mientras que en estadística no contamos de antemano con el modelo. En ésta, partimos de observaciones parciales o no-exhaustivas de una población, esto es, una muestra, y buscamos describir esos datos y hacer inferencias sobre las características de la población. Por ejemplo, inferir el modelo que con mayor verosimilitud da origen a la muestra (cf. Fuentes García *et. al.*, 2019, pp. 2-4).

Pearl explica esta distinción en términos de los parámetros de cada disciplina, e incluye la distinción entre la estadística y el cálculo causal:

> Un **parámetro probabilista** es cualquier cantidad definida en términos de una función de probabilidad conjunta. [...].
>
> Un **parámetro estadístico** es cualquier cantidad definida en términos de una distribución de probabilidad conjunta de variables observadas, sin asumir nada respecto de la existencia o no-existencia de variables no-observadas. [...].
>
> Un **parámetro causal** es cualquier cantidad definida en términos de un modelo causal, tal que no es un parámetro estadístico.

[8] Ω es un espacio de probabilidad (el conjunto de todos los resultados posibles); **F** es una *σ-álgebra*, un espacio de sucesos (un suceso es un conjunto de cero o más resultados); y **P** es una función de probabilidad que asigna una probabilidad a cada suceso.

(Pearl, 2009, p. 38.)

Hay un modelo causal de por medio, que distingue a los métodos causales de los estadísticos.

La lógica nos enseña que inferencias con mayor contenido exigen compromisos con más presupuestos. Podemos ver en los *Elementos* de Euclides que el quinto postulado aparece hasta la proposición P.I-29, lo cual significa que todas las proposiciones anteriores podían probarse sin necesidad del quinto postulado, pero que muchas de las siguientes, por ejemplo la P.I-35, no podrían haberse probado sin ese postulado.[9] De un modo análogo, para inferir relaciones causales, necesitamos supuestos causales. Pearl distingue los supuestos causales de los supuestos estadísticos:

> Un **supuesto estadístico** es cualquier restricción sobre una distribución conjunta de una variable observada; por ejemplo, que f es una normal multivariada, o que P es markov-relativa a un dado DAG D.
>
> Un **supuesto causal** es cualquier restricción sobre un modelo causal que no se puede especificar mediante supuestos estadísticos; por ejemplo, que f_i es lineal, que U_i y U_j (inobservadas) están correlacionadas o que x_3 no aparece en $f_4(pa_4, U_4)$.
>
> (Pearl, 2009, p. 39.)

Más adelante en este capítulo, en la sección *Inferencia ¿de cuál causalidad?*, se explica qué es un DAG, qué son las f_i en un modelo causal, el papel de las variables U, etcétera. Por lo pronto, el propósito es notar que se reconoce una distinción terminológica y metodológica entre la estadística y el pensamiento causal, y que la distinción está definida formalmente.

La distinción entre tipos de parámetros y supuestos apunta a una distinción semántica. Es decir, a relaciones e ideas que no pueden ser expresadas mediante parámetros y supuestos meramente estadísticos. Además de la distinción entre lo causal y lo estadístico, encontramos que las afirmaciones causales relativas a los contrafácticos no son del todo expresables mediante afirmaciones causales relativas a las intervenciones. Esto

[9] Para un análisis detallado, véase Álvarez (2021, pp. 145-166).

nos deja con tres niveles en los razonamientos causales, a los cuales se les conoce como la jerarquía causal.[10]

Pearl utiliza una metáfora para explicar estas distinciones: la *escalera* de la causalidad, con la que distingue tres niveles o peldaños de conocimiento:

(3) contrafácticos,
(2) intervenciones, y
(1) observaciones.

(Se prefirió invertir la numeración para mantener la imagen de la escalera; Pearl usa una ilustración (cf. Pearl, 2018, p. 28).)

En el primer nivel, que se basa en la asociación entre los datos, se encuentran la mayoría de los métodos estadísticos y probabilísticos tradicionales: regresiones, distribuciones, emparejamientos, y toda clase de funciones que *se ajustan a los datos*. En palabras de Pearl: «El primer peldaño de la escalera se encarga de predicciones basadas en observaciones pasivas. Se caracteriza por la pregunta "¿qué pasa si veo que...?"» (Pearl, 2018, p. 29). Por ejemplo, en un contexto agrónomo, uno puede medir el pH del suelo; los agrónomos saben que con un $pH = 3$, en general, habría poca cosecha. En ese caso, los agrónomos están infiriendo la probabilidad de la cosecha dada su *observación* del pH: $p(cosecha \mid pH = 3)$.

El segundo peldaño introduce el operador $do( )$. El operador significa una intervención. Por ejemplo, ¿qué sucedería si, independientemente del lugar en que se encuentren, los agrónomos usan un químico para que la tierra tenga un nivel de acidez $pH = 6$? Nos estamos preguntando por $p(cosecha \mid do(pH = 6))$. Ésta es una pregunta distinta de $p(cosecha \mid pH = 6)$, porque sabemos que la acidez del suelo está relacionada con otros factores, como la cantidad de manganeso, la temperatura, la altura, etc. Si sólo observamos un $pH = 6$, podemos esperar que la cantidad de manganeso sea baja, que la temperatura sea adecuada y, por tanto, que la cosecha sea abundante. En cambio, si forzamos un suelo aleatorio a tomar tal

[10] El análisis de esta jerarquía causal más reconocido es Barenboim *et al.* (2022).

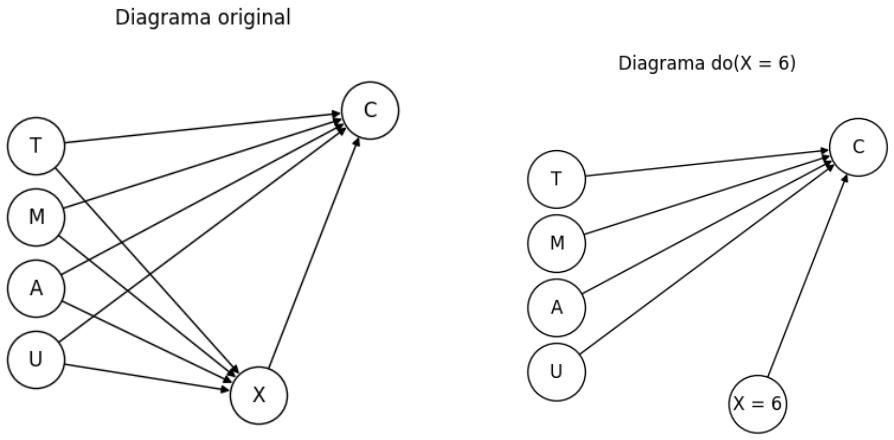
acidez *do(pH=6)*, tenemos menos seguridad de la cantidad de manganeso y de la bondad de la temperatura y los demás factores. Aunque esta tesis no es una investigación en agronomía, no es descabellado decir que

$$p(\textit{cosecha} \mid do(pH = 6)) < p(\textit{cosecha} \mid pH = 6).$$

De cualquier modo, lo importante es que se está midiendo algo distinto, y sólo en condiciones muy particulares se puede presumir que ambas cantidades son iguales. En este segundo caso, los agrónomos *intervienen* la variable.

Como se podrá haber notado, el operador *do( )* realiza una función muy similar a la de las pruebas aleatorias RCT (*Random Controled Trial*). En varias ocasiones se puede obtener el mismo resultado mediante el operador *do( )*, sin necesidad de hacer una prueba RCT. Pearl explica esto en el capítulo 4 de *The Book of Why* (2018), titulado "Confundiendo y desconfundiendo, o campeando la variable furtiva". En efecto, uno de los usos más importantes de este operador es la desconfusión de las variables de confusión.[11]

En los diagramas causales, el operador *do()* implica borrar todos los vínculos que apuntan a la variable intervenida. La razón es que, si fijamos a la variable en un valor determinado, en este caso *do(X = 6)*, no importa



[11] Una variable de confusión es la que afecta tanto a la (presunta) causa como al (presunto) efecto que queremos medir, y suele ser la fuente de correlaciones espurias.

cómo originalmente influyen las demás variables en X; sin importar las demás variables X tomará el valor que le hemos asignado:
donde T: temperatura, M: manganeso, A: altura, y U: otros factores.

Ahora demos un paso hacia el tercer nivel. Supongamos que el suelo de hecho tenía un $pH = 6$, y que en una parcela se recolectaron cincuenta cebollas. A un agrónomo le interesaría saber si un mayor pH (v.g. $pH = 7$) produciría una mejor cosecha. Pero esto ya no es una intervención. El agrónomo no puede regresar el tiempo, intervenir el pH y observar el nuevo resultado. La pregunta es, además de contrafáctica, retrospectiva: ¿cómo hubiera sido la cosecha si el pH hubiera sido mayor? Para expresar esta idea, Pearl propone usar subíndices. Sea que:

C = cosecha, pH = X, a = 6, b = 7.

Entonces, la probabilidad de que la cosecha hubiera sido mayor si se hubiera dado que $pH = X = b$, se escribe así:

$$P(C_{x_b} > 50\ )$$

Se lee "la probabilidad de que la cosecha hubiera sido mayor a cincuenta si X hubiera tenido el valor b". Pero a nuestros agrónomos seguramente les gustaría incorporar la información que ya obtuvieron, es decir, el hecho de que con X = a, se cosecharon cincuenta cebollas. Se expresaría del siguiente modo:

$$P\left(C_{X_b} > 50 \,\middle|\, X = a, C = 50\right)$$

Se lee "La probabilidad de que C hubiera sido mayor a cincuenta si el pH hubiera sido b, dado que el pH fue a y la cosecha fue 50". Este tipo de afirmaciones va más allá de la mera observación y se vincula con el proceso o el mecanismo que subyace a lo observado.

Ya que se han introducido estas nociones, se definirán los diagramas y los modelos causales.

Los diagramas causales son grafos dirigidos. Un grafo está conformado por un conjunto V de vértices (o nodos) y un conjunto E de enlaces (o vínculos). En los diagramas causales, la dirección de los vínculos describe la dirección causal. Dicha dirección moldea el flujo de información entre unos nodos (predecesores) y otros (sucesores).

Las redes bayesianas causales sólo sirven para expresar relaciones de intervención. Si, aunadas a estas relaciones, buscamos expresar también relaciones contrafácticas, necesitamos un modelo causal estructural. La siguiente es la definición de un modelo causal estructural:

> **Definición 7.1.1 (modelo causal estructural)**
> Un modelo causal es una tripla
>
> $$M = \langle U, V, F \rangle$$
>
> donde:
> (i) U es un conjunto de variables circunstanciales [*background variables*] (también llamadas 'exógenas') que están determinadas por factores externos al modelo;
> (ii) V es un conjunto $\{V_1, V_2, \dots, V_n\}$ de variables, llamadas endógenas, que están determinadas por variables del modelo, esto es, por variables en $U \cup V$; y
> (iii) F es un conjunto de funciones $\{f_1, f_2, \dots, f_n\}$ tales que cada f_i es un mapeo de (los respectivos dominios de) $U_i \cup PA_i$ a V_i, donde $U_i \subseteq U$ y $PA_i \subseteq V \backslash V_i$ y el conjunto entero de F forma un mapeo de U a V. En otras palabras, cada f_i en
>
> $$v_i = f_i(pa_i, u_i), \qquad i = 1, \dots, n,$$
>
> asigna un valor a V_i que depende de (los valores de) un conjunto selecto de variables en $U \cup V$, y el conjunto completo de F tiene una solución única $V(u)$.
>
> (Pearl, 2009, p. 203.)[12]

[12] Convencionalmente, PA_i se usa para referirse a los parientes markovianos de la variable V_i, esto es, al subconjunto mínimo de variables que, si condiciona a la variable en cuestión, la vuelve independiente de todos sus predecesores: $P(x_j \mid pa_j) = P(x_j \mid x_1, \dots, x_{j-1})$.

Pearl añade que «todo modelo causal M puede asociarse a un grafo dirigido G(M), en el cual cada nodo se corresponde con una variable y los enlaces dirigidos apuntan desde los miembros de PA_i y U_i hacia V_i. Llamamos a tales grafos 'diagrama causal asociado con M'» (*ibidem*). Un modelo estructural sin una distribución de probabilidad es interpretado como un modelo determinista, y sirve para modelar situaciones que lo son. Aunque, en realidad, los modelos estructurales muestran toda su utilidad especialmente cuando incluimos en ellos una distribución de probabilidad.

Recordemos que Pearl desarrolló este cálculo, en gran medida, motivado por la pregunta sobre cómo enseñarle a un robot o a una máquina a pensar causalmente. Para hacerlo, es necesario primero que el robot pueda representarse conocimientos causales correctamente, muchas veces en situaciones de incertidumbre. Estos modelos son, como se puede notar, computables, y requieren relativamente poco poder computacional (comparados con las redes de aprendizaje profundo). Por todo ello, es especialmente importante para el caso el hecho de que admitan un uso probabilístico:

> **Definición 7.1.6 (modelo causal probabilístico)**
> Un modelo causal probabilístico es una dupla
>
> $$\langle M, P(u)\rangle$$
>
> Donde M es un modelo causal y P(u) es una función de probabilidad definida sobre el dominio de U.
>
> (*Ibidem,* p. 205.)

La probabilidad de un evento *y* se define de manera sencilla en estos modelos:

$$P(y) \triangleq P(Y = y) \; = \sum_{\{u|Y(u) = y\}} P(u)$$

Nótese que $\{X_1, \dots, X_{j-1}\}$ es el conjunto de los predecesores de X_j, y $PA_i \subseteq \{X_1, \dots, X_{j-1}\}$ es el subconjunto mínimo que satisface la igualdad.

$$P(Y_x = y) = \sum_{\{u|Y_x(u) = y\}} P(u)$$

(cf. *ibidem*, p. 205.)

donde $\{u|Y(u) = y\}$ se refiere a las observaciones u de las variables U que implican la observación $Y = y$, y $Y_x(u)$ es la respuesta potencial de Y a X bajo u. Es decir, el valor que Y adquiriría si X fuera x y U se realiza en u. Por lo tanto, la segunda sumatoria se refiere a todas las probabilidades de los valores de U que, bajo la intervención do(X = x) producen Y = y.

Ahora se puede entender cabalmente el significado de 'probabilidad contrafáctica'. $P(Y_x = y)$, 'la probabilidad de que Y adquiriera el valor y si X hubiera sido x', significa la probabilidad de que Y tome el valor y en el submodelo M_x que resulta de la intervención $do(X = x)$.

Dicho así, parecería que no hay diferencia entre $P(Y_x = y)$ y $P(Y = y \mid do(X = x))$. Pero hay algo que las distingue, lo cual las convierte en ideas distintas y permite pensar a las probabilidades contrafácticas como un tipo de conocimiento superior al de las intervenciones: las probabilidades contrafácticas nos habilitan para preguntarnos por las probabilidades de valores distintos de los actuales (los que efectivamente ocurrieron), de un modo que no sea trivial ni contradictorio. Queremos saber cosas del estilo 'la probabilidad de que no me hubiera dado cáncer si no fumara, dado que sí he fumado y sí padezco cáncer'.

Un buen intento de representar una probabilidad tal con intervenciones sería el siguiente. Sean X y Y variables binarias que representan fumar y tener cáncer, respectivamente; entonces intentemos expresar 'la probabilidad de que no me hubiera dado cáncer si no fumara, dado que sí he fumado y sí padezco cáncer' de este modo: $P(\neg y|do(\neg x), x, y)$. No es posible, porque sabemos que $P(\neg y|\, y) = 0$, y que la conjunción $(do(\neg x)\ \&\ x)$ es contradictoria.

En cambio, el lenguaje contrafáctico, que se expresa con subíndices, facilita un tratamiento matemático a tales preguntas sin caer en contradicciones ni en trivialidades: $P(Y_{\neg x} = \neg y \mid x, y)$. La expresión se lee tal como se mencionó dos párrafos atrás, 'la probabilidad de que no me

hubiera dado cáncer si no fumara, dado que sí he fumado y sí padezco cáncer'. El subíndice significa que se considera a la variable Y en el mundo en el que X hubiera tomado el valor $X = \neg x$. Una manera común para calcular tal tipo de probabilidades es construir una *red gemela*, esto es, una red bayesiana en la que las variables del escenario contrafáctico se consideran como variables distintas a las del escenario actual, pero donde ambos conjuntos de variables (el contrafáctico y el actual) comparten las variables circunstanciales; es a través de estas variables que se obtiene la información de un escenario a otro (cf. Pearl, 2009, pp. 213-214). Tal expresión despeja las contradicciones al hacer patente que estamos razonando a partir del contraste entre dos submodelos distintos ($M_{\emptyset}$, M_x).

Así, en el nivel de las observaciones, razonamos considerando una distribución de probabilidad. En el nivel de las intervenciones, consideramos las consecuencias que conlleva un modelo intervenido en la distribución de probabilidad; y en el nivel de los contrafácticos consideramos una familia de submodelos, y los comparamos unos con otros.

Causas potenciales y genuinas

En el apartado anterior, se definió qué es un modelo causal. Sin embargo, hay tres preguntas cruciales sin cuyas respuestas los modelos causales no tendrían sentido, las cuales se responderán en ésta y en las siguientes dos secciones, respectivamente:

¿Qué es una relación causal?
¿Cómo inferir una relación causal?
¿Cómo calcular un efecto causal?

Como se mencionó, la causalidad de la propuesta de Pearl no es una causalidad absoluta ni necesariamente determinista. De afirmar que 'una variable X es causa de otra variable Y' no se sigue que X sea la única causa, ni que sea causa de modo determinista. Por ello, en vez de 'X es causa de Y', resulta más adecuada la frase 'X influye causalmente en Y' para afirmar una relación causal.

La característica propia de las relaciones causales (en el enfoque de Pearl) es que, si una variable X es causa de otra variable Y; entonces, si *sacudes* a X, *sacudes* a Y. Es posible alterar a X alterando a Y. Se dice 'sacudir' por el contexto probabilista. Decir que *manejando* a X *manejas* a Y sería mucho decir; en este contexto, de que X influya causalmente en Y no se sigue que a toda acción en X le siga una reacción en Y en la misma dirección y de manera estrictamente proporcional.

Se ha de considerar también que, en la mayoría de los contextos en los que nos interesa inferir información causal, no observamos todas las variables relevantes. Para contemplar en la inferencia a las variables no observadas, Pearl define las *estructuras latentes*:

> Una *estructura latente* es una dupla $L = \langle D, O \rangle$ en la que D es una estructura causal sobre V, donde $O \subseteq V$ es un conjunto de variables observadas. (Pearl, 2009, p. 45.)

Recuérdese que **V** es el conjunto de variables endógenas (variables cuyo valor es una función que toma como argumentos una o más variables del mismo modelo. Una estructura causal D es el diagrama acíclico dirigido (DAG) que representa un modelo causal. En este caso, distinguimos a las variables O de las variables V porque, si queremos inferir relaciones causales contemplando variables no observadas, se entiende que el conjunto V de las variables endógenas no está completamente definido.

Puesto que tenemos (sólo) observaciones, estimamos una distribución de probabilidad $\hat{P}$ sobre las variables O. Sabemos también que, por lo regular, hay muchos modelos distintos que pudieron haber dado origen a la distribución $\hat{P}$. Haciendo referencia a la navaja de Ockham, Pearl introduce la condición de que las inferencias causales se realicen considerando sólo los modelos consistentes con $\hat{P}$ que presentan una estructura mínima, y no todo el conjunto de modelos consistentes con la distribución.[13] A esta condición se le conoce como minimalidad.

[13] Se sabe que Ockham nunca formuló la frase 'No se multipliquen las entidades más allá de lo que es necesario'. Spade & Panacci (2019) explican con claridad la versión histórica de esta idea de Ockham. Pearl aquí está hablando de Ockham en su versión coloquial.

Una manera de definir la causalidad es definir qué es lo que inferimos cuando inferimos causalidad. Pearl ofrece una definición de causalidad inferida:

> **Definición 2.3.6 (*Causalidad inferida*)**
> Dada $\hat{P}$, una variable C tiene un efecto causal en E si y sólo si existe una ruta dirigida de C a E en todas las estructuras latentes mínimas consistentes con $\hat{P}$. (Pearl, 2009, p. 46.)

El hecho de que haya una ruta en todas las estructuras latentes (mínimas) significa que (1) todos los modelos (mínimos) consistentes con lo observado conllevan una relación de dependencia entre C y E, y (2) la relación entre C y E es asimétrica, pues la ruta se presenta de C hacia E. Lo cual sugiere que puedo alterar E alterando C, pero no viceversa.

Esta definición no es suficiente para entender qué significa propiamente una relación causal. En especial, no es suficiente para entender por qué la relación causal no sería entonces una mera relación de dependencia. Una relación causal es una propiedad estructural de un modelo causal. Para especificar tal relación, Pearl primero define qué es una causa potencial, y después qué es una causa genuina (la segunda presupone la primera):

> **Definición 2.7.1 (*causa potencial*)**
> Una variable X tiene una influencia causal potencial en otra variable Y (inferible a partir de $\hat{P}$) si se cumplen las siguientes condiciones:
>
> 1. X y Y son dependientes en todo contexto.
> 2. Existe una variable Z y un contexto S tales que
>   (i) X y Z son independientes dado S (i.e. $X \perp\!\!\!\perp Z \mid S$) y
>   (ii) Z y Y son dependientes dado S (i.e. $Z \not\perp\!\!\!\perp Y \mid S$)
>
> (Pearl, 2009, p. 55)

Un contexto es un «conjunto de variables con valores específicos asignados» (Pearl, 2009, p. 55). Esta definición asegura dos características de la relación entre X y Y; que (1) X es distinta de Y (X no puede ser causa de sí misma), y que (2) X no depende funcionalmente de Y. Si hay una S y

una Z tales que (X ⫫ Z | S) y (Y ⫫̸ Z | S), eso implica que Y es posterior a X en algún ordenamiento temporal. Luego, la dependencia no podría darse de Y hacia X.[14] Por eso la dependencia entre ambas variables podría deberse a que X es causa de Y.

La causalidad genuina se define del siguiente modo:

> **Definición 2.7.2 (*causa genuina*)**
> Una variable X tiene una influencia causal genuina en otra variable Y si existe una variable Z tal que:
>
> 1. X y Y son dependientes en todo contexto, y existe un contexto S que satisface:
>   (i) Z es una causa potencial de X (por definición 2.7.1),
>   (ii) Z y Y son dependientes dado S (i.e. Z ⫫̸ Y | S), y
>   (iii) Z y Y son independientes dado $S \cup X$ (i.e. Z ⫫ Y | $S \cup X$);
>
> o ya sea que
>
> 2. X y Y están en la clausura transitiva de la relación definida en el criterio 1.
>
> (Pearl, 2009, p. 55.)

El primer criterio define la relación entre dos variables adyacentes; el segundo establece que la relación es transitiva.

Hay dos ideas en la definición de causalidad que, al menos hasta donde el autor de esta tesis ha investigado, no habían sido formuladas en la historia de la filosofía previa al proyecto de causalidad probabilista: (1) se necesita, al menos, una tercer variable para afirmar la relación causal, y (2) la tercer variable ha de preceder (en un ordenamiento temporal estadístico) a la que se está afirmando como causa. (El trabajo de Suppes (1970) fue muy influyente en cuanto a estas dos ideas.)

A partir de lo anterior, lo característico de una relación causal se puede sintetizar del siguiente modo: una relación causal es una relación estructural de dependencia direccionada entre dos variables. Se infiere a partir de una familia de modelos mínimos consistentes con la distribución $\hat{P}$, que se estima a partir de un conjunto de observaciones. Se presume que, para cualquier $\hat{P}'$ procedente de un conjunto distinto de observaciones

[14] Pearl distingue entre 'tiempo estadístico' y 'tiempo físico' (véase 2009, p. 58).

(otro experimento, observaciones en otras circunstancias, etc.) en las que estén involucradas la variable causa C y la variable efecto E, se mantendrá la relación estructural entre C y E. Empero, tal presunción no puede afirmarse categóricamente, pues las distribuciones $\hat{P}'$ se estiman a partir de las observaciones, con métodos estadísticos que involucran incertidumbre o riesgo en sus razonamientos (cf. Cartwright, 2007, p. 30); por ejemplo, en el intervalo de confianza o en el nivel de significancia.

Una relación causal predice intervenciones; esto es, permite decir cómo se alterará E si se altera C. También sintetiza contrafácticos del tipo 'dado que vimos $C = c$ y $E = e$, E sería e' si C fuera c'. Pearl, con cierta complacencia, nos recuerda que la definición contrafáctica de causalidad aparece en la obra de David Hume: «si el primer objeto no hubiera estado, el segundo nunca hubiera existido» (Hume, 1772, sección VII, parte II, §60). La complacencia de Pearl se nota en la manera en que presenta esta definición (Pearl, 2018, p. 265), y es totalmente justificada: Hume, el mismo filósofo que evidenció que la conjunción constante no implica por sí misma una relación causal —idea fácilmente asociable al espíritu de Pearson según el cual o la causalidad es un caso trivial de correlación total o ninguna correlación implica causalidad—, resulta ser el mismo filósofo que planteó la idea fundamental de la propuesta de Pearl: la causalidad es una noción contrafáctica, pues captura una relación estructural que no se puede reducir a una observación.

Inferir efectos causales

Definir modelos sería baladí si, junto con ellos, no presentáramos la manera de inferirlos y de asegurar que son apropiados para las tareas en cuestión. Como parte del método de inferencia causal, Pearl presenta un algoritmo que permite implementar computacionalmente su propuesta. El algoritmo se llama IC: causalidad inductiva (por sus siglas en inglés, *inductive causation*). Es el algoritmo fundamental de la inferencia causal. Su papel es análogo al que el algoritmo de retropropagación desempeña en el enfoque neuronal.

Existen dos versiones del algoritmo: IC e IC*. La versión simple es IC, que supone que todas las variables han sido observadas, y produce la familia de diagramas DAG mínimos compatibles con la distribución $\hat{P}$. Estrictamente hablando, el algoritmo supone que $\hat{P}$ fue generada por un DAG D_0 subyacente, y obtiene la clase equivalente de D_0. En cambio, IC* es una versión realista, pues considera la presencia de variables no-observadas; por lo mismo, es un poco más complejo y su resultado es un modelo parcial; no siempre se puede inferir todas las relaciones causales que generaron los datos pero, de hecho, se puede inferir algunas (cf. Pearl, 2009, pp. 52-53). En este trabajo sólo se presentará la versión simple.

> **Algoritmo IC (*causalidad inductiva*)**
> **Input:** $\hat{P}$, una distribución estable sobre un conjunto V de variables.
> **Output:** un patrón $H(\hat{P})$ compatible con $\hat{P}$.
>
> 1. Para cada par de variables a y b en V, buscar el conjunto S_{ab} tal que $(a \perp\!\!\!\perp b \mid S_{ab})$ se da en $\hat{P}$ —en otras palabras, a y b deben ser independientes en $\hat{P}$, condicionando en S_{ab}—. Construye un grafo no-dirigido G tal que los vértices a y b están conectados con una arista si y sólo si no se puede encontrar un conjunto S_{ab}.
> 2. Para cada par de variables no adyacentes a y b con un vecino en común c, revisar si $c \in S_{ab}$.
>
>   Si pertenece, continuar.
>
>   Si no, entonces agrega puntas a los enlaces, de modo que apunten a c (i.e. $a \rightarrow c \leftarrow b$).
> 3. En el grafo parcialmente dirigido que resulta, orientar tantas aristas no-dirigidas como sea posible obedeciendo dos condiciones: (i) cualquier orientación alternativa produciría una *estructura-v*; o (ii) cualquier orientación alternativa produciría un ciclo dirigido.
>
> (Pearl, 2009, p. 50.)

Hace falta aclarar dos nociones antes de explicar los pasos del algoritmo. Primera, un *patrón* $H(\hat{P})$ es la clase de equivalencia de un diagrama D_0 consistente con $\hat{P}$ (cf. Pearl, 2009, p. 49). Segunda, una *estructura-v* consiste en dos flechas que convergen en una variable, donde las variables de las colas no están conectadas por flecha alguna (cf. Pearl, 2009, p. 19).

Las implementaciones computacionales del algoritmo requieren tiempo polinomial. Sobre la complejidad computacional de estos métodos, en general, véase Pearl (2009, p. 21).

El primer paso del algoritmo IC detecta todas las dependencias estructurales. Si a y b son dependientes y no existe un conjunto S_{ab}, ha de ser que la dependencia es estructural y contigua. Los otros dos pasos obedecen a un *tollendo ponens*.

En el segundo, tenemos tres opciones para direccionar los enlaces: colisión ($a \rightarrow c \leftarrow b$), cadena ($a \rightarrow c \rightarrow b$) y bifurcación ($a \leftarrow c \rightarrow b$). Si $c \notin S_{ab}$, se tienen las siguientes relaciones ($a \perp\!\!\!\perp b$), ($a \not\perp\!\!\!\perp c$) y ($b \not\perp\!\!\!\perp c$). Sólo la estructura de colisión satisface tales relaciones.

En cuanto al tercer paso, hay diversas propuestas para implementarlo que pueden revisarse en Pearl (2009, pp. 50-51).

La propuesta inferencial de Pearl exige comprometerse con tres supuestos: (1) minimalidad, (2) estabilidad y (3) la condición de Markov. Explicaré brevemente a qué se refiere cada uno. Los tres están presentes en el funcionamiento del algoritmo IC, y muchas de las críticas en contra de estos métodos se han dirigido hacia los supuestos de estabilidad y la condición de Markov.

Minimalidad refiere a que, al momento de hacer una inferencia causal, de entre todos los modelos consistentes con la distribución de probabilidad observada $P_{[o]}$, escojamos el mínimo. Ahora bien, el modelo mínimo no se corresponde necesariamente con el modelo que tiene la menor cantidad de enlaces o de nodos. Un modelo mínimo es el que se adecua a la menor cantidad de distribuciones además de $P_{[o]}$ (cf. Pearl, 2009, p. 45).

Estabilidad es una propiedad que presumimos en la distribución de probabilidad a partir de la cual realizaremos la inferencia causal. Una distribución P es estable si las independencias que observamos en ella se mantienen en otras distribuciones P' generadas por el mismo modelo pero con distintos parámetros. Para una definición formal de esta propiedad, véase Pearl (2009, p. 48).

Se explicó antes que la causalidad se infiere a partir de las relaciones observadas de independencia (y de no-independencia). Podría suceder que las independencias que observamos no se deban al mecanismo (la estructura del proceso) que generó los datos, sino al azaroso valor que en tal circunstancia tomaron los parámetros. En ese caso, cometeríamos errores al momento de inferir relaciones causales a partir de esas relaciones

accidentales de independencia. La presunción de estabilidad asume que, al obtener nuestra distribución, ninguna relación de independencia se debe a una coincidencia accidental de los valores de los parámetros.

Existe un debate filosófico y científico en torno a la validez de este supuesto. Una de las ideas centrales de Cartwright al objetar contra este supuesto es que no hay razones *a priori* para suponerlo, esto es, que no deberíamos basarnos en este supuesto en problemas en los que no contamos con el conocimiento experto que lo legitime (Cartwright, 2007, pp. 63-72). Puesto que dicha discusión no es estrictamente relevante para los argumentos que se defienden en esta tesis, me limitaré a decir que asumo el argumento que Pearl ofrece para la validez de este supuesto como verdadero (Pearl, 2009, pp. 62-63). La idea principal es que las ligeras variaciones experimentales de los contextos macroscópicos para los que fueron diseñados sus métodos nos salvaguardan de obtener distribuciones inestables.

La *condición causal de Markov* (CMC, *Causal Markov Condition*) establece una relación entre una variable, sus predecesores y sus sucesores. Hay, al menos, dos teoremas clave para entenderla:

> **Teorema 1.2.6 (*condición parental de Markov*)**
> Una condición suficiente y necesaria para que una distribución de probabilidad P sea Markov-relativa a un DAG G es que toda variable sea independiente de todos sus no-descendientes (en G) condicionada en sus padres. (Excluimos a X_i al hablar de sus 'no-descendientes'.) (Pearl, 2009, p. 19).

Se suelen utilizar metáforas genealógicas para referirse a la posición de las variables (padres, hijos, hermanos, abuelos, esposos, etc.). Los padres de X_i son todas las variables que apuntan directamente a X_i. El siguiente es el segundo teorema clave:

> **Teorema 1.4.1 (*condición causal de Markov*)**
> Todo modelo causal markoviano M induce una distribución $P(x_1, .., x_n)$ que satisface la condición parental de Markov relativa a un diagrama causal G asociado con M; esto es, cada variable X_i es independiente de todos sus no-descendientes, dados sus padres PA_i en G. (Pearl, 2009, p. 30.)

Sobre los parientes PA_i, véase la nota a la definición de modelo causal estructural. El teorema 1.2.6 define la CMC en la relación entre una distribución de probabilidad y un DAG. Lo que añade el teorema 1.4.1 es que esta condición aplica a los diagramas causales asociados con un modelo causal.

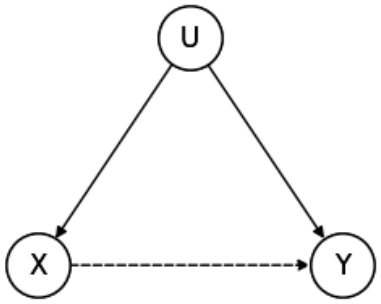
Hay dos afirmaciones en juego concernientes a la CMC: existe un conjunto mínimo y suficiente de padres y, segunda, hay un orden respecto de cómo se define ese conjunto mínimo. De hecho, un teorema anterior a la condición parental es la *condición de orden de Markov* (cf. Pearl, 2009, p. 19).

Calcular efectos causales

No sólo podemos inferir una relación causal, también podemos medirla. En algunos casos, incluso, podemos hacerlo a partir de datos meramente observacionales.

Para concluir esta sección, considérese un ejemplo de cómo calcular un efecto causal. Pearl aplica sus métodos a una discusión histórica, la polémica sobre si fumar es causa de cáncer, en la que participaron estadísticos famosos como R. A. Fisher (el desarrollo detallado de este ejemplo se encuentra en Pearl 2009, pp. 231-234). El problema reside en que, a pesar de que observemos un incremento en los casos de cáncer cuando se trata de personas fumadoras, no sabemos de antemano si el incremento es provocado por el tabaquismo, o si hay otros factores que, ambas cosas, inducen a fumar y te hacen proclive al cáncer, sin que haya una influencia real del tabaquismo hacia el cáncer. Podría existir un gen o una configuración genética que motivara el deseo por la nicotina y, al mismo tiempo, propiciara el cáncer.

Considérese el siguiente diagrama:



U: gen fumador
X: Fumar
Y: cáncer de pulmón

Nos preguntamos si el vínculo $X \rightarrow Y$ es válido. Más precisamente, si existe una ruta dirigida de X a Y.

Sabemos que la probabilidad de cáncer aumenta en casos de tabaquismo, $p(y) < p(y|x)$. Pero queremos averiguar si fumar es causalmente relevante para el cáncer. Esto es, tenemos la hipótesis alternativa:

$$H_a: p(y) < p(Y_x = y)$$

Y queremos contrastarla con la hipótesis nula, según la cual fumar no es causalmente relevante para el cáncer (aún cuando, de hecho, es observacionalmente relevante):

$$H_0: p(y) = p(Y_x = y).$$

Con sólo estas tres variables e información observacional no podríamos averiguarlo, porque no podemos medir la presencia del gen fumador[15] (al menos no se podía cuando se desarrolló históricamente esta polémica),

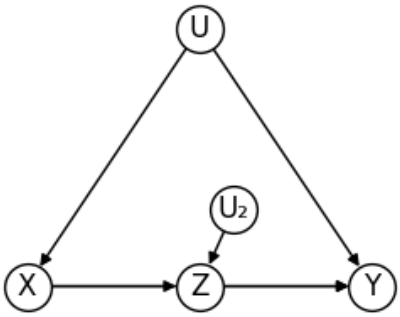
[15] Para mayor claridad sobre a qué se refiere Pearl con el 'gen fumador', véase la siguiente nota al pie.

y, por lo tanto, no podemos determinar con sólo las observaciones si la dependencia entre X y Y es causal o espuria.

Idealmente, necesitaríamos escoger una muestra representativa y asignar aleatoriamente quiénes en esa muestra fumarán y quiénes no fumarán durante toda su vida. Después, comparar cuántos desarrollaron cáncer en cada grupo. Por razones éticas, un experimento así resulta inviable. Incluso insultante.

Lo que sí podemos hacer es encontrar una cuarta variable Z a través de la cual actúe el tabaquismo sobre el cáncer, y que no esté afectada por el factor de confusión (i.e. la U de la genética). La presencia del alquitrán en los pulmones cumple con tales condiciones. Si la incorporamos en nuestros razonamientos, obtenemos el siguiente diagrama:

U: gen fumador



U_2: factores que afectan la cantidad de alquitrán
Z: alquitrán en los pulmones
X: Fumar
Y: cáncer de pulmón

Añadimos U_2 para contemplar otros factores que puedan afectar al alquitrán (v.g. el entorno de trabajo). Hay dos premisas clave en este diagrama: (1) No hay una flecha $U \rightarrow Z$ y (2) tampoco hay una flecha $U \rightarrow U_2$. Lo cual significa que U no confunde a {X, Z}, ni a {Z, Y}; no lo hace directamente y tampoco lo hace a través de U_2.

Pearl muestra, basado en las reglas, axiomas y definiciones de su cálculo, que el efecto causal que fumar ejerce sobre el cáncer $P(Y_x = y)$ es

igual al producto de los efectos de fumar sobre el alquitrán $P(Z_x = z)$ y del alquitrán sobre el cáncer $P(Y_z = y)$:

$$\sum_z P(Z_x = z)P(Y_z = y)$$

(Pearl, 2009, p. 234).

Más aún, prueba que, en este caso particular, la siguiente igualdad se sostiene: $P(Z_x = z) = P(z|x)$. El efecto causal de fumar sobre el alquitrán equivale a la probabilidad de observar una cantidad z de alquitrán cuando se sabe que la persona fuma. Y también prueba que la probabilidad de desarrollar cáncer si se tuviera la cantidad z de alquitrán equivale a la sumatoria, para todos los valores x de X (en este caso, fumar y no fumar), del producto de la probabilidad de x y la probabilidad de tener cáncer dada la cantidad z de alquitrán y la realización de x:

$$P(Y_z = y) = \sum_x P(y|x, z)P(x)$$

(Pearl, 2009, p.233).

Por lo tanto, tenemos que el efecto causal de fumar sobre el cáncer está determinado por las siguientes cantidades:

$$P(Y_x = y) = \sum_z P(z|x) \sum_{x'} P(y|z, x')P(x')$$

(Pearl, 2009, p. 234),

donde *x'* se refiere a todas las realizaciones de *X*.

Si Pearl está en lo correcto (el presente trabajo asume que lo está), halló la manera de calcular efectos causales a partir de información meramente observacional (recabando datos sobre los fumadores, pero sin realizar un experimento diseñado).[16] Nótese que del lado derecho de la

[16] Está en lo correcto en un sentido inferencial; en sentido biomédico podrían hacerse varias precisiones. Pearl (2018, pp. 342-343) provee una descripción del estado de la cuestión en términos biomédicos. Para 2008 ya se sabía que el polimorfismo de nucleótido único (SNP, *Single Nucleotid Polymorphism*) rs16969968 está fuertemente asociado al

igualdad todas las cantidades son observaciones (no hay subíndices contrafácticos ni operadores *do()*). Esto rompe la barrera inferencial entre, por una parte, distribuciones de probabilidad y, por otra parte, afirmaciones y mediciones causales. Algunas dependencias observadas implican causalidad.

En sentido estricto, junto con su equipo de trabajo y la tradición a la que pertenece, Pearl resolvió con su cálculo un problema filosófico que nos espinó durante siglos. Dicho con sus palabras:

> La habilidad de deducir dirección causal a partir de un ensamblaje de mecanismos simétricos (junto con la selección de un conjunto de variables endógenas) significa que detectar relaciones causales no es distinto de detectar (v.g. mediante experimentos) leyes físicas ordinarias, tales como la ley de elasticidad de Hooke o la ley de la aceleración de Newton. Esto no implica que detectar leyes físicas sea una tarea trivial, libre de sutilezas metodológicas y filosóficas. Mas sí implica que el problema de la inducción causal —uno de los más ásperos en la historia de la filosofía— puede reducirse al, más familiar, problema de la inducción científica. (Pearl, 2009, p. 228.)

Se explicó, en el apartado *La motivación computacional*, que Pearl considera dos enigmas de la causalidad. Lo anterior significa que resolvió el primero (¿cómo inferir relaciones causales?).

---

cáncer de pulmón (Hung *et al.*, 2018; Thorgeirsson *et al.*, 2008); a este SNP se le conoce coloquialmente como el gen fumador, pues la asociación con el cáncer se da precisamente en las personas fumadoras. VanderWeele (2014) sustenta tres afirmaciones: el gen no incrementa significativamente el consumo de cigarro, no causa cáncer más que a través del fumar y, tercera, incrementa considerablemente el riesgo de cáncer para las personas que fuman. Pearl (2018, pp. 227-228), además, relata que David Freedman (estadístico de Berkeley) ha criticado el realismo de su planteamiento; de acuerdo con Freeman, si hay un gen fumador, éste podría afectar a la manera en que el cuerpo se deshace de la materia extraña, de modo que las personas con el gen sean más vulnerables a la formación de depósitos de alquitrán en los pulmones. Ello significa que tendría que haber una flecha causal del gen al alquitrán, lo cual invalidaría el cálculo que presenta Pearl. Ante esta crítica, Pearl aprovecha para declarar que él no es un especialista en cáncer y que siempre referirá a un experto para determinar si el diagrama se corresponde o no con la situación real; lo que él quiere notar es que, si se cumplen las condiciones, uno puede separar cuantitativamente el efecto de un factor de confusión sin tener datos sobre el factor de confusión (en el ejemplo del gen fumador, se puede separar cuantitativamente el efecto del gen fumador sobre el cáncer, sin tener datos sobre la presencia del gen fumador en cada caso, es decir, usando los datos de las demás variables).

Respecto al segundo enigma (¿qué añade una afirmación causal?), Pearl se dedica toda una sección a explicar exactamente cuál es el contenido empírico de los contrafácticos. Explica, en primer lugar, que los contrafácticos pueden verse como un modo abreviado de hablar de predicciones: lo que sucederá si cambian las condiciones actuales. Pero no tendría sentido desarrollar todo el cálculo de contrafácticos si el contenido empírico de éstos se redujera meramente a paráfrasis o circunlocuciones para hablar de predicciones (cf. Pearl, 2009, p. 218). Los contrafácticos proporcionan dos aspectos empíricos más.

Por una parte, permiten expresar de un modo transparente las cláusulas *Ceteris Paribus*. En experimentos o afirmaciones empíricas en las que 'todo lo demás se mantiene igual' permiten saber con precisión qué es todo eso que se mantiene igual. Con un toque de humor, Pearl lleva esto al extremo diciendo que, obviamente, en un experimento con voltaje hay cosas que no tendríamos que 'mantener igual', por ejemplo, ¡el voltímetro! (cf. Pearl, 2009, p. 218). Lo que se mantiene igual son las variables U, y las funciones f_i de las variables V_i que no hayan sido intervenidas.

Por otra parte, los contrafácticos poseen un valor explicativo, pues capturan información sobre el mecanismo (el proceso) que subyace a los datos. El valor empírico de su aspecto explicativo suele relucir sólo en circunstancias excepcionales, circunstancias en las que se ha alterado el mecanismo subyacente. Ahora bien, el hecho de que el valor empírico añadido se enmarque en condiciones excepcionales, más que restarles importancia, los dota de una relevancia especial, porque es en tales circunstancias cuando más difícil y más urgente suele ser conseguir información empírica.

> Esa es la naturaleza de cualquier explicación causal: su utilidad no se prueba en situaciones estándar, sino más bien en nuevos escenarios que acucian manipulaciones innovadoras de los estándares. La utilidad de entender cómo funciona la televisión surge no de girar las perillas correctamente, sino de la habilidad de reparar el set de TV cuando se descompone. Recuérdese que todo modelo causal provee no uno, sino una hueste de submodelos, cada uno creado al violar ciertas leyes. La autonomía de los mecanismos en el modelo causal, entonces, presenta una invitación abierta a remover o reemplazar tales mecanismos, y es simplemente natural que el valor explicativo de los

> enunciados se juzgue a partir de qué tan bien predicen las ramificaciones de tales reemplazos. (Pearl, 2009, p. 219-220.)

Se trata de una predicción, sí, pero de una que no se puede expresar (menos calcular) con enunciados predictivos convencionales.

En esta sección se presentó la definición formal de causa, el algoritmo para inferir relaciones causales y (un esbozo de) la manera de calcular con precisión la magnitud de un efecto causal. También se habló de las consecuencias filosóficas de poder, en algunas circunstancias, calcular un efecto causal a partir de cantidades meramente observacionales.

Calcular efectos causales con cantidades observacionales implica la reducción del problema de la inducción causal al problema de la inducción científica. Donde 'reducción' no quiere decir que el contenido semántico de una relación causal sea igual que el de una relación observacional, sino que, aún cuando se trata de una afirmación más fuerte (estructural), esta afirmación es (en términos de veracidad) igual de problemática que cualquier otra afirmación de inducción científica. Ya que ambos tipos de afirmaciones son igual de problemáticos, tenemos que decir que son igual de aceptables.[17]

[17] Lo son en los (muchos) contextos en los que los tres supuestos (minimalidad, condición de Markov y estabilidad) son válidos.

VI. La Inteligencia Causal frente al *Deep Learning*

Al comienzo de esta tesis apunté al hecho de que ambos, los métodos causales y los de aprendizaje estadístico no son incompatibles. Pearl (2018, p. 351) dedica una sección a explicar los beneficios que la minería de datos ha traído para la ciencia y, en general, los beneficios de los avances en el enfoque de aprendizaje estadístico. Dos ejemplos que menciona son el proyecto de 1000 genomas (1000 *Genoma Project*) y el *Mikolski Archive for Space Telescopes* de la NASA.

En 2018 se publicó un artículo de Adnan Darwiche, cuyo propósito es reflexionar sobre la situación actual y la dirección del proyecto de Inteligencia Artificial (Darwiche, 2018). Está claro también para Darwiche que no hay una oposición teórica entre, como él los llama, el enfoque basado en modelos (representar y razonar) y el enfoque basado en funciones (ajustar una función a los datos). Lo que él advierte son objetivos y actitudes que difieren desde la perspectiva social, la perspectiva de la ciencia, la de la comunidad científica y la del proyecto de IA (él no las clasifica explícitamente de este modo).

Originalmente, el proyecto de Inteligencia Artificial se proponía emular artificialmente la o las funciones de la inteligencia humana. Puesto que la inteligencia humana es el mejor ejemplo de inteligencia general que conocemos, ésta juega el papel de un récord natural para evaluar los sistemas artificiales.

Darwiche se pregunta cuál es la mejor manera de describir lo que le sucedió recientemente a la IA. Por lo general, se reconoce que hubo un parteaguas entre 2012 y 2018;[18] sin embargo, resulta difícil describir en qué consistió esencialmente. Darwiche propone entenderlo como un conjunto de tres desarrollos.

Al primero lo podemos llamar, propiamente, teórico o científico (en referencia a la estadística y a la ciencia de la computación): el incremento de poder computacional junto con el desarrollo de técnicas estadísticas y

[18] Tomo como referencias el desempeño de las redes convolucionales (Hinton et al., 2012) y el artículo de Darwiche (2018), que ya nota como un hecho dicho parteaguas.

de optimización (descenso de gradiente estocástico, *dropouts* y nuevas funciones de activación).

Un segundo desarrollo es de carácter empírico: «identificamos una clase de aplicaciones prácticas que se corresponden con funciones que, ahora lo sabemos, son suficientemente simples para permitir representaciones compactas que pueden ser evaluadas eficientemente» (Darwiche, 2018, p. 59). Él no usa la palabra 'empírico', pero sólo haciendo la prueba experimental de la adecuación empírica de estas funciones (v.g. redes neuronales) en dicha clase de tareas (muchas de ellas perceptuales, como el reconocimiento de imágenes) se pudo identificar la clase mencionada.[19]

El tercer desarrollo consiste en haber cambiado gradualmente los objetivos y las métricas de evaluación «de maneras en las que se reducen considerablemente los retos técnicos [...], a la vez, manteniendo nuestra capacidad de capitalizar comercialmente los resultados obtenidos» (Darwiche, 2018, p. 60). En otras palabras, el "logro" está en que, a pesar de que disminuyeron las exigencias (lo cual pudo haber provocado un demérito del programa de IA), con ello sacaron un gran provecho comercial.

Un ejemplo de este tercer cambio lo encontramos en las métricas de evaluación de tareas de traducción. «En los primeros días de la IA, el éxito se medía según cuán lejos del 100% estaba la precisión de un sistema, comparado con la inteligencia humana» (*ibidem*, p. 61). Darwiche menciona que se buscaba aplicar estos sistemas en inteligencia gubernamental, donde cualquier pequeño error podría desembocar en una catástrofe política. En cambio ahora, «desde un punto de vista consumista, el éxito se mide se manera efectiva según cuán alejada está del 0% la precisión del sistema» (*ibidem*¸ p. 61). Él advierte lo esencial de este cambio: la capacidad de comprender el texto era un aspecto antes central, ahora ausente, en la evaluación de sistemas de traducción.

El contraste entre lo científico y lo consumista es un punto central de su argumento, que nos lleva a reconocer una tensión entre la labor científica y la industria. Para entender esta tensión es importante notar que el aprendizaje automático basado en funciones no es más que un área de la

[19] Para una explicación más completa sobre por qué este descubrimiento tuvo que ser empírico, véase Wolpert (1996).

amplia gama de enfoques que hay en cuestiones de aprendizaje automático y, en general, en la IA. Darwiche se pregunta si el reciente éxito de estos métodos para resolver problemas especificados principalmente a partir de objetivos comerciales justifica, por una parte, la obsesión con éstos métodos y, por otra parte, el descuido de los demás. Responde que sí lo justifica, si trabajas para una compañía. Pero no lo justifica si te importa la investigación científica (p. 62).

Dentro de la misma actividad científica también hay una tensión (la cual he esbozado en secciones anteriores, pero ahora es momento de caracterizarla cabalmente; caracterizo un aspecto particular aquí y caracterizo la tensión general en el primero de los tres argumentos que presento más adelante). De acuerdo con Darwiche, en el grueso de la comunidad científica se ha establecido la creencia de que «un método que no involucra un modelamiento explícito, o un razonamiento sofisticado, es suficiente para producir niveles humanos de inteligencia» (*ibidem*, p. 58). Tal creencia implica una disociación entre razonamiento e inteligencia.

Tradicionalmente, la actividad científica se ha considerado como una manera —probablemente la mejor— de perfeccionar la inteligencia humana. Tomando en cuenta dicha visión de la ciencia, una disociación entre, por una parte, razonar y elaborar modelos y, por otra parte, ser inteligente tiene consecuencias en la manera de entender la actividad científica en general. Implica que, en sentido estricto, no necesitamos modelos para alcanzar un conocimiento científico; y que ser inteligente se reduce a predecir, estimar y clasificar según las expectativas del contexto.

Por último (la cuarta perspectiva de las que anteriormente enlisté), la tensión también repercute en las prácticas sociales de la comunidad científica. A fin de cuentas, se ha de decidir en qué dirección y en qué proyectos invertir tiempo y dinero. Desde una perspectiva individual (conseguir una beca, publicar un artículo) parece más racional seguir la moda establecida. Pero, como bien advierte Darwiche, desde una perspectiva grupal, descuidar todos los demás campos de la IA no es la elección más sensata. La razón es que, del hecho de que actualmente unos métodos estén siendo exitosos, no se sigue que a largo plazo vayan a ser los más exitosos ni los que más necesitamos (cf. Darwiche, 2018, pp. 62-64).

Concuerdo con el diagnóstico de Darwiche. Pero propongo dos modificaciones a este retrato panorámico de la situación actual del proyecto de la IA. La primera está relacionada con una de las conclusiones de su artículo: propone trabajar en maneras de fusionar ambos enfoques. «La cuestión no se trata de si son funciones o modelos, sino de cómo integrar y fusionar profundamente funciones y modelos» (*ibidem*, p. 67). Tiene sentido hablar de una fusión, si lo pensamos meramente en términos técnicos. Pero pensándolo desde las demás perspectivas (social, proyecto de IA, científica, comunidades de investigación), la fusión no es posible. Tenemos que escoger una dirección o la otra: servir a los intereses (y a las indiferencias) consumistas o buscar una comprensión científica y preocuparnos por los factores involucrados y sus interrelaciones al momento de hacer una inferencia. Es decir, ¿vamos a utilizar a los métodos basados en modelos para perseguir lo objetivos del consumismo? ¿O, viceversa, aprovecharemos los resultados del aprendizaje profundo en favor de los objetivos a los que originalmente han estado orientados los métodos basados en modelos?

Un caso notable de la fusión de ambos enfoques en cuestiones técnicas son las TBNs, desarrolladas por Choi, Wang & Darwiche (2019). Las TBNs (*Testing Bayesian Networks*) son redes bayesianas extendidas con unidades de prueba (*testing units*). La idea que guio el diseño de las TBNs consiste en que, en las redes neuronales, lo que las habilita para ser aproximadoras universales (en oposición a sólo poder aproximar funciones lineales) son las funciones de activación (como la sigmoidal o la ReLU). Las unidades de prueba juegan ese mismo papel en las TBNs, lo cual las convierte en aproximadoras universales.

Si uno se concentra exclusivamente en el diseño de las TBNs, no está claro que un enfoque deba dominar. ¿Están potenciando las redes bayesianas con cualidades de las redes neuronales, o están *neuralizando* las BNs, es decir, aumentado la complejidad con objetivos de aproximación, mera optimización, etc.?

Esta ambigüedad, sin embargo, no es posible al momento de decidir financiamientos, temas de tesis, objetivos científicos y esquemas de razonamiento en la industria y en lo político.

El segundo punto que modificaría del retrato panorámico consiste en añadir dos ámbitos más desde los cuales reflexionar sobre la tensión entre ambos enfoques: el ámbito epistemológico y el ámbito neurocientífico.

En cuanto al ámbito epistemológico, el desacuerdo está en aceptar o negar la existencia de conocimiento causal como un tipo de conocimiento de naturaleza distinta y más perfecta que el meramente asociativo. En cuanto al ámbito neurocientífico, ya se ha mencionado en el capítulo II cómo se enriquecen mutuamente las neurociencias y la IA. El desacuerdo está en si la perspectiva 'neuronal artificial' es adecuada para describir la cognición humana, dada la fisiología del sistema nervioso.

A continuación, presento un argumento relacionado con cada uno de estos dos ámbitos que propongo considerar (argumentos primero y tercero), y otro argumento (el segundo) relacionado con las tareas de control en los sistemas de IA.[20] Darwiche (2018) menciona dicho argumento; lo que añadiré es una estructuración del argumento, a partir de las investigaciones relacionadas con el trabajo de Pearl y apuntando cómo actualizar el proyecto de IA desde una perspectiva científica y sin traicionar a sus objetivos originales.

Primer argumento: epistemología

Juntos, la teoría del aprendizaje estadístico y los métodos de programación neuronal, no implican, pero apuntan, hacia una epistemología que se limita a construir el conocimiento a partir de observaciones pasivas, esto es, apuntan a no considerar como parte de la construcción del conocimiento al proceso que originó las observaciones. Si tomamos como referencia a la jerarquía del conocimiento causal, dicha epistemología niega que sea válido considerar los niveles segundo (intervenciones) y tercero (contrafácticos) en la construcción del conocimiento.

Presentaré tres cualidades epistémicas para defender que dicha epistemología centrada en los datos implica una comprensión limitada, la

[20] El orden de los argumentos responde a motivos de claridad.

cual es indeseable para la labor científica. Estas cualidades parten respectivamente del papel que desempeñan en nuestro conocimiento (1) las representaciones, (2) los cambios en (y las relaciones entre) las distribuciones de probabilidad y (3) las características o variables elegidas para realizar inferencias.

Estas tres cualidades muestran la necesidad de incluir a las relaciones causales como un constitutivo esencial del conocimiento. Es decir, la necesidad de una epistemología orientada según la causalidad.

Primera cualidad epistémica: la comprensión en las representaciones. Sabemos, a grandes rasgos, por qué funcionan los algoritmos que llevan a las redes neuronales a encontrar una solución. Pero no nos es accesible (no es interpretable) la representación mediante la cual operan.[21] Esto significa que la representación no es un resultado ni un objetivo epistémico. Pues, aun si podemos almacenar los pesos de las redes neuronales y después reutilizarlos para otros problemas, lo que soluciona los problemas es la fuerza bruta: un enorme poder computacional para operar extensas redes neuronales (millones o, incluso billones de parámetros) con colosales bases de datos (en algunos LLMs con, prácticamente, la internet entera). No sabemos cuál parámetro es más importante que otro, es decir, cuál nos permitiría prescindir de otros, ni la medida exacta de la importancia que tiene cada uno. No alcanzamos a entender el mecanismo. Más aún, el mecanismo no refleja ni incorpora ningún tipo de comprensión estructural.

El resultado son los datos del *output* que responden a los datos del *input*, y que, aunque nos proporcionen información que no nos era clara o, incluso, que no nos era conocida, están epistémicamente siempre al mismo nivel que los datos del *input*. Si preguntamos '¿por qué obtenemos

[21] En el debate contemporáneo se suele distinguir entre explicabilidad (*explainability*) e interpretabilidad (*interpretability*). Floridi et al. (cf. 2022, p. 16) reconocen que hay varias maneras en que se han definido estos términos; una de ellas nos es útil para este trabajo: explicabilidad refiere tanto a expertos como a no-expertos, en el sentido de que un experto pueda explicarle el mecanismo (o la parte clave del mecanismo) de un algoritmo a un no-experto y éste efectivamente lo entienda, mientras que interpretabilidad refiere específicamente al conocimiento experto, esto es, al hecho de que un experto pueda seguir el razonamiento implementado por el algoritmo y que este razonamiento sea aceptable desde un punto de vista teórico. Valga decir que en Russell y Norvig (2021, pp. 711-712) leemos definiciones de estos términos ligeramente distintas.

ese output?', la única respuesta sería que es el más consistente con el estado de los datos y la arquitectura de la red neuronal en cuestión (la cual, podríamos añadir, se ha mostrado eficiente para este tipo de tareas). No aprendemos nada sobre las particularidades del fenómeno del que tratan esos datos, más allá de la eficacia de estos algoritmos. Vamos de los datos a los datos, y usamos una representación cuya única función es servir tal transferencia.

En los modelos causales, la representación sí es un objetivo epistémico. En ese sentido, vamos más allá de los datos. No se busca una función que imite lo observado. Se busca una representación que capture el proceso subyacente, de modo que ésta sea parte de nuestro conocimiento y que nos sea posible razonar directamente a partir de la representación inferida.

Segunda cualidad epistémica: las relaciones de diversas distribuciones de probabilidad asociadas al mismo proceso. Los métodos basados exclusivamente en la teoría del aprendizaje estadístico son muy precisos para afirmar alguna conclusión a partir de la presencia de nuevas observaciones. Pero no pueden decir nada acerca de situaciones en las que la distribución de probabilidad se altera por una intervención o por algún evento exógeno. No pueden, a partir de la información sobre alguna intervención, inferir directamente los cambios en la distribución de probabilidad. Lo más que podemos hacer con estos algoritmos es reentrenarlos en las nuevas circunstancias. Se esperaría que, si en verdad entendemos un fenómeno, pudiéramos inferir esas consecuencias sin empezar el aprendizaje desde cero.

En términos más precisos, Vapnik habla de dos distribuciones de probabilidad, la empírica y la real. Utiliza la distribución empírica para inferir la distribución real, porque la distribución real es desconocida. Su trabajo consiste en encontrar la mejor manera de aproximarse a la distribución real ajustando alguna función o algoritmo a la distribución empírica. Pero, idealmente, espera que ambas distribuciones sean la misma. No hay razonamientos a partir de la comparación de varias distribuciones, sino razonamientos a partir de *una sola* distribución (sea para inferir la

distribución desconocida o para aplicar a la solución de un problema la distribución inferida).

Pearl, en cambio, trabaja con una familia de 'n' distribuciones de probabilidad,[22] por ello podemos decir que trabaja, más que con datos, con modelos. Su objetivo es encontrar o aproximar el modelo del cual es miembro la distribución empírica. No hay en Pearl algo así como la 'distribución real', porque lo real, en todo caso, sería el modelo subyacente, y las distribuciones generadas por este modelo (entre ellas la distribución empírica) son parcelas de la realidad. Sería más preciso, en todo caso, hablar de la 'distribución actual'.[23] Su método permite realizar inferencias precisas en circunstancias en las que el proceso que genera los datos (el mecanismo) se altera y, por ende, altera la distribución de probabilidad; en especial, estas inferencias se realizan sin necesidad de ejecutar de nuevo un proceso de aproximación a los datos. En lugar de actualizar los parámetros (por ejemplo, de una red neuronal) con una gran cantidad de iteraciones, basta con actualizar el diagrama haciendo unos pocos cambios en los enlaces.

Dicho con otras palabras, los algoritmos neuronales detectan correlaciones, pero sabemos que las correlaciones no siempre implican dependencia. Si queremos razonar a partir de dependencias o de relaciones estructurales (como la relación causa-efecto), necesitamos otro tipo de métodos, los cuales se basan en modelos.

Otra manera de apreciar esta cualidad epistémica está en reconocer que, incluso si pudiéramos tener observaciones sobre todos los estados posibles de un fenómeno, eso no nos daría todo el conocimiento al que podemos aspirar sobre éste. Además de las observaciones, podemos entender por qué se comporta de ese modo y no de otro. Y para ello, es necesario entender la estructura o las relaciones que pautan el comportamiento del fenómeno. Epistémicamente, suele ser más valioso conocer dicha estructura o tales relaciones que conocer los meros datos.

---

[22] Una por cada combinación de $\{x_j, \ldots, x_k, \ldots, x_n\}$ tales que son realizaciones (por intervención) de $\{X_j, \ldots, X_k, \ldots, X_n\}$. Estas combinaciones se representan en los subíndices de las variables de interés. Por ejemplo: $P(Y_{\{x_j,\ldots,x_k,\ldots,x_n\}} = y)$. Pero también se utilizan para representar las distribuciones P originadas por cada intervención; por ejemplo: $P_{\{x_j,\ldots,x_k,\ldots,x_n\}}$.

[23] Puesto que en este caso podría haber confusión, vale aclarar que ésta es una interpretación del autor de la tesis. Pearl no habla exactamente en esos términos.

Tercera cualidad epistémica: identificar características fuertes. Recuérdese la distinción que presenta Vapnik entre el enfoque clásico en estadística y la nueva técnica: uno busca características fuertes y funciones simples, y la otra busca muchas características débiles y funciones audaces. Hay un supuesto implícito en esta descripción; supone o que la selección de características fuertes no puede ser formal y que, por ende, debemos renunciar a tomarla como paso decisivo, o que no ha sido posible desarrollar tal procedimiento. Pero esa selección de características fuertes es precisamente lo que solemos llamar teoría, o pensamiento teórico. Puesto que las características en el proceso que describe Vapnik son muchas y débiles, también en este caso tenemos que describir su propuesta como una manera de ir de la información dada a la información deseada (de lo observado a lo observable) sin pasar por el acrisolamiento de una teoría. Digo 'acrisolamiento' para referirme a la distinción entre las variables causalmente relevantes y las causalmente irrelevantes. El nuevo enfoque no considera dicha distinción.

Mencioné que los métodos de razonamiento causal están más cerca del enfoque clásico que del nuevo enfoque. Nótese que la razón por la que están más cerca no está en que los métodos de razonamiento causal persigan como algo intrínsecamente deseable que las características (fuertes) sean pocas y que las funciones sean simples. Lo que buscan es la *estructura* causal del proceso en cuestión. Y resulta que dicha estructura tiende más a culminar en modelos en los que está claro cuáles son las variables propiamente relevantes.

Exhibiendo estas tres cualidades epistémicas, se concluye que las ideas que guían al aprendizaje computacional actual (tanto las de la teoría del aprendizaje estadístico como las del enfoque neuronal) toman como base y objetivo epistémico a los datos, a las observaciones. La manera de abordar el problema del aprendizaje conlleva supuestos y creencias respecto de lo que es el conocimiento y la inteligencia. Una creencia obvia que acompaña al enfoque basado en los datos es que en el conocimiento y en el aprendizaje la pieza fundamental son los datos. Otra menos obvia es que, cuando el objetivo es inferir lo que desconocemos, lo que buscamos es información del mismo calibre epistemológico que nuestros datos: los casos nuevos. Cualquier

representación o modelo no es más que un utensilio para acceder a la información desconocida en los casos nuevos, y no necesita mostrar ningún tipo de comprensión explicativa respecto de la relación entre los datos muestrales y los casos nuevos. Por lo tanto (se entiende que esto promueve el enfoque de *programación neuronal*), los defectos y confusiones en la representación son negligibles mientras los resultados en los casos nuevos sean aceptables.

Es importante decir que eso no significa que estos métodos no sean eficientes ni deseables en determinadas circunstancias. De hecho, puede ser que en situaciones en las que no sabemos cómo representar el problema, una red neuronal sea mucho más eficiente que cualquier otro método que requiera partir de una representación, y también en las que realmente sólo se busca permanecer al nivel de los datos e incursionar en las representaciones y modelos realmente no traería ninguna ventaja. (Hay que decir que esta última no es la situación de las investigaciones científicas que buscan una comprensión teórica, en las que tener una teoría significa contar con un modelo explicativo.[24])

La renuncia a inferir modelos de las relaciones entre las variables implica permanecer, en lo que respecta al análisis de datos, en las

[24] Hay una tradición de pensamiento en filosofía de la ciencia que, precisamente, afirma que presentar una teoría es, de hecho, presentar un modelo (cf. van Fraassen, 1989, p. 222). Se le conoce como la *concepción semántica de la ciencia* (van Fraassen retoma esa identificación entre teorías y modelos de Patrick Suppes). Se puede notar que la tesis que aquí se defiende discrepa con un aspecto importante de la concepción semántica. De los filósofos relacionados con esta tradición, al menos van Fraassen afirma que los modelos (teorías) son herramientas que ofrecen explicaciones en un sentido pragmático; por ejemplo, afirma que Newton usó su teoría para explicar las mareas del mismo modo en que un carpintero usa un martillo para clavar un clavo (1980, p. 100). De acuerdo con él, el propósito de una explicación es la adecuación empírica; en sus palabras, cuando se trata de una explicación referida pragmáticamente a las personas, «la búsqueda de la explicación es, *ipso facto*, la búsqueda de la adecuación empírica» (*Ibidem*, p. 157). No obstante, en esta tesis se defiende que un modelo correcto es un objetivo epistémico más completo que la mera adecuación empírica; un modelo incluye la adecuación empírica, pero no se limita a ella. Al lector que no esté familiarizado con la obra de Pearl, le sorprenderá saber que Pearl (2009) reconoce a Suppes (1970), Salmon (1984) y Cartwright (1989) como inspiradores e interlocutores suyos. Dedica la sección 7.5 de *Causality* "Structural versus Probabilistic Causality" para dialogar con ellos y, al final, rechazar su propuesta en favor de los modelos causales estructurales. Van Frassen toma una parte central de su tesis de Suppes y refiere a obras previas de Salmon.

limitaciones de los datos. Los métodos dirigidos por los datos son métodos limitados a los datos.

Tras estudiar los métodos propuestos por Pearl es fácil ver cómo las representaciones, las relaciones entre distribuciones de probabilidad y la identificación de características relevantes están, las tres, íntimamente relacionadas. Sin embargo, a partir de la investigación realizada para esta tesis sobre la historia de la IA, me atrevo a decir que antes del desarrollo de los métodos propuestos por Pearl y por científicos y filósofos afines, no era obvio que estas tres cualidades estuvieran tan íntimamente engarzadas. Esto puede interpretarse incluso como un resultado acerca de las funciones epistemológicas de los modelos: encontrar un modelo adecuado es encontrar una familia de distribuciones de probabilidad en la que, para cada distribución observada tenemos una explicación de por qué sucedió, y además es encontrar, para distintas preguntas, una distinción precisa entre el conjunto de variables relevantes y el conjunto de variables irrelevantes.

Nótese que la epistemología que admite la legitimidad de conocimientos del segundo y tercer nivel en la jerarquía causal es la más adecuada para las ciencias macroscópicas (en la sección IV se ofrecieron ejemplos de este tipo de interrogantes científicas). Así que, aceptar esta dimensión epistemológica que va más allá de lo observado, que comprende la naturaleza de los procesos, es afín a los intereses científicos.

Segundo argumento: hacia los sistemas de IA general

Este argumento se centra en la Condición de Markov y busca mostrar que ésta, aunque no es adecuada para algunos contextos, en vistas a los sistemas de IA general, puede ser una ventaja.

Considérese de nuevo la distinción de Vapnik (la estadística clásica frente a la nueva técnica). Una señal de que comprendemos algo, según se dijo en el argumento anterior, es que podemos distinguir, dada una pregunta, entre información importante e información irrelevante. Adviértase ahora que en una red neuronal no podemos definir un procedimiento para seleccionar un conjunto pequeño (si no el mínimo) de conexiones

neuronales que tornen negligibles a las demás conexiones (esto quedará más claro en el siguiente párrafo). Con los modelos causales, en cambio, sí existe un procedimiento para seleccionar el conjunto mínimo de conexiones causales que hacen negligibles a las demás. Son las conexiones asociadas a los *padres markovianos* de cada variable. Lo cual muestra que esa comprensión que adquirimos al distinguir los factores relevantes depende de la Condición Causal de Markov (CMC: *Causal Markov Condition*).

Nótese que las redes neuronales son DAGs (grafos acíclicos dirigidos). Si quisiéramos hablar de los padres markovianos de un nodo en la red neuronal, o del output, tendríamos que enlistar todos los nodos de la capa previa. Pero eso no nos daría ningún tipo de comprensión, porque en una red neuronal, estrictamente, los nodos tomados por separado no representan nada. En algunas redes convolucionales, por ejemplo en las que se aplican a reconocimiento de "emociones" a partir de la expresión facial, es posible asociar algunos nodos en ciertas capas con características como 'nariz', 'ojos', etc. Este tipo de rastreos en las redes neuronales nos explican cómo funciona la red neuronal, nos dan una noción de cómo ésta llega al resultado que nos presenta, pero no nos dicen nada sobre el mundo; no nos explican los aspectos del fenómeno que estamos analizando.

Los modelos causales junto con la CMC también están sujetos a críticas. La objeción más fuerte que se ha esgrimido contra la CMC consiste en notar que ésta no es adecuada en varios escenarios. Cartwright presenta cinco escenarios en los que falla esta condición, casos en los que, después de haber condicionado en todos los factores relevantes, aún observamos dependencias que no se deben a relaciones causales.

Los cinco escenarios se definen por involucrar, respectivamente: (*i*) causas comunes, (*ii*) causas que cooperan para producir un mismo efecto, (*iii*) poblaciones mezcladas, (*iv*) cambios en la misma dirección del tiempo y (*v*) subproductos. A partir de estos escenarios concluye que la conexión entre dependencias estructurales y relaciones causales no es estricta. Dicho de otro modo, que no funciona todo el tiempo. (Cartwright, 2007, pp. 76-79).

En palabras de Cartwright, «Las causas *pueden* incrementar la probabilidad de sus efectos, pero no es necesario que lo hagan. Y viceversa:

un incremento en la probabilidad *puede* deberse a una conexión causal, pero muchas otras cosas pueden también ser razón de ello» (*ibidem*, p. 79). Ella rechaza la obsesión con encontrar un 'si y sólo si' que con ciertas condiciones probabilistas determine la existencia de un vínculo causal, y concluye este argumento con un consejo que suele dar en sus cursos de ciencias sociales:

> Si observas una dependencia probabilista y estás inclinada a inferir una conexión causal a partir de ésta, piensa arduamente en todas las otras posibles razones por las que esa dependencia podría ocurrir y elimínalas una por una. Y cuando hallas terminado, recuerda: tu conclusión no es más certera que tu confianza en que realmente has eliminado todas las posibles alternativas. (Cartwright, 2007, p. 79.)

Cartwright no critica la *consistencia* de los modelos causales. Su objeción apunta a la *adecuación* de los modelos causales con los fenómenos a los que se aplican. Debemos ser muy cuidadosos al evaluar las características de la población en el fenómeno al que nos enfrentamos. Eso significa que para razonar adecuadamente con modelos causales necesitamos conocimiento experto sobre el tema que nos interesa. Ésta es quizás una de las razones por las que no es tan fácil vincular a los modelos causales con productos de consumo, como sí lo es con las redes neuronales. Podemos hablar de 'usuarios' de redes neuronales como ChatGPT, pero difícilmente hablaríamos de usuarios de modelos causales en el mismo sentido.

Otro aspecto que puede criticársele a la CMC es que en muchos sistemas la manera de hacer válida esta condición consiste en suponer la existencia de estructuras o variables latentes (se ofrece una definición de estas estructuras en la Sección anterior). En los modelos no-markovianos se puede reestablecer la CMC si se acepta la existencia de variables latentes. ¿Qué tan razonable es asumir tales variables? Los métodos de aprendizaje estadístico no suponen nada sobre la existencia o la no existencia de variables latentes, lo cual suele tomarse como una ventaja.

El agnosticismo respecto de las variables no observadas parece ser una actitud propiamente científica, pero puede resultar perniciosa en circunstancias en las que, de hecho, asumir variables latentes es lo más razonable. Pearl reconoce las limitaciones de los métodos que propone

respecto de los modelos no-markovianos: «Declaramos que estamos dispuestos a perdernos el descubrimiento de modelos causales no-markovianos que no puedan ser descritos con variables latentes» (Pearl, 2009, p. 61). Sin embargo, también apunta que «No considero que ésta sea una pérdida seria, porque tales modelos —si alguno existe en el mundo macroscópico— tendrían una utilidad limitada para guiar decisiones» (*ibidem*). Y explica que con los modelos no-markovianos no estaría claro cómo uno podría predecir los efectos de las intervenciones. Se podrían enlistar todos los efectos explícitamente y de manera anticipada, lo cual simplemente mostraría que el modelo carece de utilidad.

En muchos de los problemas macroscópicos somos conscientes de que nuestros modelos no están agotando el fenómeno. Asumir la existencia de variables latentes es una manera razonable de asumir la consciencia de que nuestro modelo no está agotando el fenómeno. Lo que suele ser difícil aceptar es la independencia entre las variables estocásticas u_i. Asumir la independencia de tales variables es necesario para mantener la CMC.

Resulta que hay un tipo de tareas para las que la CMC es completamente razonable: las tareas de control. Hasta donde el autor de esta tesis ha logrado investigar, no hay mejores métodos para calcular sistemáticamente las intervenciones de un sistema sobre su entorno (en términos causales) de modo que este cálculo resulte en políticas aplicables. No los había para el momento de la publicación de *Causality* (Pearl, 2009). Y las aplicaciones más recientes que involucran modelos para tareas de comportamiento toman precisamente el marco de Pearl (véase Méndez-Molina, Morales & Sucar, 2022).

Los métodos basados en funciones pueden optimizar la obtención de algún tipo de recompensa mediante el aprendizaje por refuerzo. Difícilmente, empero, tendremos con ellos un proceso de razonamiento claro sobre los factores que orientaron la decisión. Ese es el tipo de proceso que quisiéramos tener claro en tareas de control, esto es, tareas en las que no basta con ajustar un comportamiento, sino que es necesario encontrar una explicación para distintos tipos de comportamiento en distintos escenarios.

El desarrollo en tareas de control es imprescindible si partimos de los objetivos originales del proyecto de IA. Era claro, según se ha visto en las secciones anteriores de este trabajo, que el proyecto apuntaba a sistemas que emulen a la inteligencia humana de modo general: sistemas de IA general. En vistas a pensar en maneras de actualizar los objetivos de la IA sin traicionar al proyecto, podríamos tomar la idea de 'general' en un sentido que no sea sinónimo de 'total', siempre y cuando no descuidemos un objetivo crucial del proyecto: comprender los procesos de inteligencia.

Una referencia importante en ello es la idea de Marvin Minsky y Seymour Papert que ellos mismos llamaron 'La sociedad de la mente'. Ambos, por separado, dedicaron libros a dicha teoría: *Mindstorms* (Papert), el homónimo *The Society of Mind* (Minsky) y *The Emotion Machine* (Minsky), publicados en 1982, 1987 y 2006 respectivamente. La idea más relevante de su teoría para el presente argumento está expresada, no obstante, en el epílogo a la segunda edición de *Perceptrons*:

> En muchas situaciones, los humanos muestran habilidades excesivamente lejanas a lo que se puede aprender con redes simples y uniformes. Pero cuando tomamos esas aptitudes por separado, o cuando tratamos de hallar cómo fueron aprendidas, esperamos encontrar que éstas fueron hechas mediante procesos que de algún modo combinan el trabajo (que ya se ha realizado) de varias agencias más pequeñas, ninguna de las cuales, por separado, necesita trabajar en escalas mucho más grandes que las de PDP. ¿Esta hipótesis es consistente con el estilo conexionista de PDP? Sí, en tanto que los cómputos que realiza el sistema nervioso pueden representarse como la operación de sociedades de redes. Pero no, en tanto que el modo de operar de tales sociedades de redes (tal como nos las imaginamos) suscita problemas teóricos de diferente tipo. Tenemos la expectativa de que procedimientos como GD no serán capaces de producir tales sociedades. Se necesita algo más. (Minsky & Papert, 1988, p. 267.)

Analicemos este párrafo. Cuando Minsky y Papert hablan sobre el 'estilo conexionista' y sobre PDP (*Parallel Distributed Processing*) se refieren a lo que aquí he llamado el enfoque de programación neuronal (véase Rumelhart *et al.*, 1986b). Ahora bien, algo que dudaban los autores era la capacidad de extender la idea de los perceptrones a redes neuronales gigantes. Nótese que escribieron esto apenas dos años después del artículo que presentó el algoritmo de retropropagación (Rumelhart *et al.*, 1986a). En ese

momento no había sucedido el parteaguas del aprendizaje profundo que describe Darwiche (2018). Con 'GD', Minsky y Papert se refieren a la técnica de descenso gradiente (*gradient descend*), que, como se dijo en la sección II es una característica esencial del algoritmo de retropropagación.

La idea es que, para coordinar la aplicación de distintas redes a distintos problemas, se necesita un sistema central que las coordine. Es difícil pensar que una red neuronal resulte adecuada para conformar un sistema central. Piénsese que en el sistema de *AlphaZero,* la máquina jugadora de Go, la red neuronal no es más que uno de los componentes; además, el sistema está diseñado con la técnica *minimax*, búsqueda estocástica, generación de datos jugando contra sí mismo (*self-play*), funciones de evaluación para cortar los árboles de búsqueda *minimax,* y técnicas de aprendizaje por refuerzo (Darwiche, 2018). Resulta engañoso llamar a este sistema simplemente una red neuronal de aprendizaje profundo.

Los modelos causales son especialmente aptos para la posición de control en un sistema complejo. En tal posición, la CMC conformaría una ventaja, puesto que los modelos causales (markovianos) determinan los factores suficientes para lograr el resultado buscado, expresan con claridad estos factores y presentan, dada la forma en que están definidos y estructurados, una explicación de por qué estos son los factores correctos (los que tornan negligibles a los demás factores). Imagínese la posibilidad de construir un sistema cuyo componente central sea un modelo causal en el que cada variable $v_i \in V$ esté alimentada por información de una red neuronal que desempeña una función cognitiva determinada, y en la que intervenciones complejas, una vez seleccionadas las variables y los valores a los que se buscará fijarlas, se realicen también mediante redes o sistemas complejos estilo *AlphaZero*.

Evidencia de que el progreso en tal dirección es posible la podemos hallar en el trabajo de (Méndez-Molina *et al.*, 2022). Se trata de un trabajo que fusiona *Q-Learning* (un algoritmo central en aprendizaje por refuerzo) con modelos causales. Méndez-Molina *et al.* hablan de tres maneras que se han buscado para poner a colaborar ambos enfoques en el ámbito del aprendizaje por refuerzo: usar aprendizaje por refuerzo (RL) para inferir modelos causales (CM), usar CM para perfeccionar RL, o usar RL y CM para

inferir y perfeccionar el modelo causal y, al mismo tiempo, potenciar el desempeño en RL. Él se ubica a sí mismo en esta tercera opción. Pero pongamos atención a lo que epistemológicamente está sucediendo: puesto que el modelo delimita y concentra las decisiones del agente, ahora tenemos una explicación (y justificación) de sus acciones, más que sólo un resultado.

Buscar este tipo de sistemas (sistemas en los que las tareas de control se basen en modelos) es una de las posibles vías en las que puede actualizarse el proyecto de IA sin traicionar a sus aspiraciones originales. Y es un caso en el que estructuralmente tendría que preferirse a los métodos basados en modelos sobre los métodos basados en funciones.

Tercer argumento: el sistema nervioso

Se mencionó ya cómo, en ocasiones, los neurocientíficos se preguntan seriamente si el cerebro humano está implementando alguno de los algoritmos descubiertos por el proyecto de IA. En concreto, la pregunta sobre si el cerebro humano está implementando *retropropagación* ha provocado mucha discusión en el ámbito neurocientífico.

Científicos como Hinton (2022), Chollet (2021) y Lillicrap *et al.* (2020) sostienen que las arquitecturas neuronales artificiales más usadas no son un buen modelo para el cerebro. Es sumamente implausible que el cerebro humano esté implementando *retropropagación*, si consideramos cómo está constituido fisiológicamente.

Buena parte de la implausibilidad se debe a la fase de realimentación (*backward pass*) del algoritmo. Como se mencionó en la Sección II de esta tesis, *retropropagación* calcula la contribución al error de cada neurona mediante las derivadas parciales del error con respecto del parámetro o peso de cada una de las neuronas ($\partial E/\partial w$). Esto exige la capacidad de (1) calcular las derivadas explícitamente, (2) almacenar la información doblemente para cada una de las neuronas (las variables de las derivadas y las variables de la actividad) y (3) transmitir esa información a través de la red sin alterar la red.

En las neuronas humanas observamos que, de hecho, se transmite información en ambos sentidos de las conexiones (propagando y retropropagando), pero la manera en que se transmite la información altera el estado de la red: la condición 3 falla (Lillicarp *et al.*, 2020). También hay evidencia de que las condiciones 1 y 2 fallan (*ibidem*). La primera porque ello involucra manejar valores extremos; las señales del error incluyen valores muy grandes y valores muy pequeños (se le conoce como el problema de las gradientes que tienden a explotar o a desvanecerse, *exploding and vanishing gradients*). La segunda falla porque, si la información de las derivadas viaja de manera retrógrada por los axones, la comunicación de esta información sería tan lenta que no podría implementar retropropagación; y, si la información se comunica mediante una red separada dedicada específicamente a este paso retrógrado, es poco probable que la estructura de esta segunda red sea exactamente igual a la de la principal.

Lillicarp *et al.* (2020) presentan la hipótesis NGRAD (*Neural Gradient Representation by Activity Differences*). La hipótesis sostiene que el cerebro humano podría estar implementando algún algoritmo que, más bien, aproxime el comportamiento de retropropagación. De acuerdo con esta hipótesis, las conexiones de realimentación podrían utilizar las diferencias (esto es, las discrepancias respecto del resultado buscado) localmente. De este modo, es posible aproximar los resultados de retropropagación y se reduce la implausibilidad fisiológica.

De cualquier modo, Hinton, quien es coautor de Lillicarp *et al.* (2020), argumenta que la hipótesis NGRAD sigue siendo insuficiente, dadas las restricciones fisiológicas. Él propone investigar un algoritmo al que ha llamado *forward-forward* (FF), que sustituye el paso retrógrado por una segunda fase de propagación. De este modo, FF no tiene las complicaciones del paso retrógrado y también puede operar sin concentrar la información precisa del estado de cada neurona. Puede tener una caja negra en medio de la red y, aún así, funcionar bien en general.

Otra característica de FF es la manera en que se entrena el algoritmo. El entrenamiento consiste en dos fases, una 'positiva' y una 'negativa'. En la fase positiva, se presentan los datos con los que ha de ser entrenada la red; cuando se trata de aprendizaje supervisado, el resultado

correcto es parte del input. En la fase negativa, se presentan datos 'falsos', esto es, datos que no se corresponden con el patrón con el que fue entrenada la red. El algoritmo FF clasifica entonces cuándo el input es un dato real y cuándo no. No supera en eficiencia a retropropagación, pero requiere considerablemente menos poder computacional y es, hasta cierto punto, más plausible biológicamente. Hinton asocia la fase positiva con la vigilia y la fase negativa con el sueño, por ejemplo.

Estas redes, tanto las crasamente artificiales (vg. los modelos GPT), como las que buscan modelar algunas regiones del cerebro humano (vg. NGRAD, FF), suelen proporcionar, en ocasiones, respuestas que lucen como el resultado de un proceso de razonamiento. El siguiente es un ejemplo tomado de una conferencia de Hinton.[25]

En la conferencia, Hinton menciona que los humanos aún somos mejores que los mejores sistemas de IA del momento en tareas de razonamiento. Pero narra que planteó a GPT-4 el siguiente problema.

> **Pregunta:** Quiero que todas las habitaciones de mi casa sean blancas. En este momento, algunas son blancas, otras son azules y otras son amarillas. La pintura amarilla se desluce hasta tornarse blanca en el lapso de un año. Entonces, ¿qué debo hacer si quiero que todas ellas sean blancas en dos años?
>
> **Respuesta:** Deberías pintar de amarillo las habitaciones azules.

"No es la solución más natural, pero funciona ¿no?", agrega Hinton. Y explica que es justo éste el tipo de razonamiento que resulta difícil de lograr usando IA simbólica; GTP-4 tuvo que entender qué significa 'deslucir', tuvo que entender un proceso temporal, etc.

Vale la pena cuestionar si en verdad lo *entendió*. No es lo mismo simular un razonamiento —producir los mismos resultados que se obtendrían si se realizara el razonamiento— que, de hecho, efectuar el razonamiento. En el primer caso se trata de una aproximación sin las garantías formales que comparten los sistemas de razonamiento.

[25] Hinton, G. (2023). "The Future of Intelligence", *MIT Technology Review, EmTech Digital.* URL = <https://www.youtube.com/watch?v=sitHS6UDMJc>

Vemos que, concentrándose sólo en cubrir ciertas demandas fisiológicas, las hipótesis vinculadas a los algoritmos NGRAD y FF suponen una estructura para la cognición humana que carece de comprensión. Todos estos algoritmos pueden aproximar una distribución, pero ninguno puede responder preguntas 'por qué'; ninguno puede computar resultados de lo que hubiera sucedido si las cosas fueran diferentes.

Así que también es sano preguntarse si, a despecho del nombre 'redes neuronales', las neuronas reales y, en general, el sistema cognitivo humano operan dentro del marco de redes neuronales tal como lo entienden los científicos del aprendizaje automático. La evidencia más reciente en neurociencias sugiere que no (Juang *et al.*, 2022).

Sabemos que la dopamina es un neuromodulador crucial para el aprendizaje de asociaciones. La hipótesis más aceptada durante muchos años sobre cómo algunos animales (incluidos los humanos) aprenden asociaciones se conoce como TDRL RPE. Esta hipótesis sostiene que, dado un estímulo, el animal predice las posibles recompensas o resultados que ocurrirán subsecuentes al estímulo. Después, considerando la diferencia temporal (TDRL, *Temporal Difference Reinforcement Learning*),[26] a partir de la recompensa se determina el error de la predicción (RPE, *Reward Prediction Error*). De ahí proviene el nombre TDRL RPE. Este modelo es una extensión de la teoría Rescorla-Wagner y, en un sentido ya algo lejano, está relacionado con las investigaciones de Pávlov. Algo esencial en esta teoría es que propone que el aprendizaje se realiza de manera *prospectiva.* El estímulo detona las predicciones de las recompensas. Va de la causa al efecto.

Los resultados de Jeong *et al.* (2022) sustentan una hipótesis distinta. De acuerdo con éstos, la manera en que se realiza el aprendizaje asociativo es *retrospectiva.* La noticia de la recompensa detona un proceso de revisión en la memoria con el objetivo de encontrar la causa de la recompensa. Va del efecto a la causa. A esta teoría la llaman 'Contingencia de redes ajustada para relaciones causales' (ANCCR, *Adjusted Net Contingency for Causal Relations*). Se pronuncia 'áncor'.

[26] Sobre cómo funciona la diferencia temporal, véase Simen & Matell (2016).

Esta manera de realizar el aprendizaje es más eficiente. Considérese que el aprendizaje prospectivo exige calcular continuamente el resultado esperado de todos los (casi infinitos) estímulos; mientras que el aprendizaje retrospectivo, dado el resultado, se concentra en buscar el estímulo relevante para establecer una relación entre éste y la recompensa.

Jeong *et al.* (2022) explican que la evidencia que sustenta TDRL RPE obtenida de los —muchos— experimentos previos igualmente sustenta el modelo ANCCR. En su artículo (1) muestran esta consistencia, (2) presentan resultados de simulaciones computacionales en las que verifican el funcionamiento del algoritmo de aprendizaje que proponen y, lo que es más importante, (3) presentan los resultados de once experimentos diseñados para distinguir ambas hipótesis, en los que midieron la actividad de la dopamina con un sensor óptico (dLight1.3b) usando la técnica de fotometría con fibra óptica (*Fiber Photometry*). Los resultados de cada experimento, todos, apuntan a rechazar TDRL RPE en favor de ANCCR.

Además, en un suplemento complementario, ofrecen los detalles del algoritmo vinculado a ANCCR. Este algoritmo incorpora la representación de modelos causales; TDRL RPE trabaja con relaciones meramente estadísticas.

Esta evidencia respalda la afirmación de que los contrafácticos son una característica intrínseca de la cognición humana. En el modelo ANCCR, el aprendizaje obedece a la pregunta '¿cuáles son los estímulos sin los que no se hubiera presentado la recompensa?'. Más aún, la evidencia nos dice que este tipo de razonamiento no es nada más una función abstracta de nuestra cognición: está enraizado a nivel sináptico en la manera en que se transmite información en nuestras neuronas.

Las redes neuronales profundas, por lo tanto, además de ser menos profundas que los modelos causales (en sentido epistemológico), también son menos neuronales. (Si se me permite algo de humor: no son profundas, no son neuronales, pero sí son *enredadas*.)

En el entendido de que queremos aprender más sobre la inteligencia viva para mejorar los sistemas de inteligencia artificial, deberíamos dejar a un lado las hipótesis según las cuales nuestra manera de entender

asociaciones está constreñida a los límites de la teoría del aprendizaje estadístico.

VII. Conclusiones

Presenté tres argumentos para defender que los métodos causales han de ser preminentes respecto de los métodos basados en funciones. Los argumentos *no concluyen* que la teoría del aprendizaje estadístico y las redes neuronales no tengan valor epistemológico, tampoco que no tengan un lugar importante en el proyecto de IA, ni que sean irrelevantes en cuestiones cognitivas. Lo que concluyen es que son insuficientes, y que conviene ponerlos al servicio de un razonamiento basado en modelos.

Estos argumentos sustentan la tesis que se enunció en la introducción. Contemplan los beneficios de *fusionar ambos tipos de métodos en un nivel meramente técnico* (como sucede con las TBNs), muestran que las epistemologías a las que tienden ambos métodos son incompatibles y que, dados los objetivos de la actividad científica en diversas áreas (una muy importante para el caso es el proyecto de IA), *es más adecuada una epistemología que distinga una jerarquía en el conocimiento causal*. Además, los resultados en neurociencias apuntan a la compatibilidad de esta epistemología con el sistema nervioso humano, de modo que sustentan dicha epistemología y permiten orientar las investigaciones en torno a la inteligencia artificial.

El tercer argumento (sobre el sistema nervioso) admite dos interpretaciones. Si se presupone una epistemología naturalizada, habría de interpretarse como un argumento contundente: *ésta es la epistemología adecuada porque es la que se corresponde con el funcionamiento del sistema nervioso,* y la única manera de negarlo sería probar que tales resultados son falsos o inadecuados. Sin embargo, el autor de esta tesis no se compromete con dicho presupuesto, y quisiera precisar la manera en que propone este argumento: *los recientes resultados en neurociencias sobre la manera causal en que aprende relaciones el sistema nervioso apuntan fuertemente (aunque no definitivamente) a admitir como correcta una epistemología que distinga jerárquicamente el conocimiento causal.* Esto no significa que sea el único ámbito en el que se pueda argumentar en favor de dicha epistemología; el primer y el segundo argumento son ejemplos de ello. Significa que no sería conveniente ignorar la importancia de este ámbito.

Para presentar los tres argumentos fue necesario (1) explicar cómo se puede entender la causalidad sin asumir que la necesidad es una característica indisociable de ésta, una noción de causalidad que parte de una visión probabilista, pero que no puede expresarse en términos meramente probabilistas y, por ello, se le llama causalidad probabilista *estructural*; (2) mostrar cómo la pregunta por la relación entre el aprendizaje, las representaciones y el conocimiento ha estado en el núcleo del proyecto de IA; (3) explicar la idea principal de la teoría del aprendizaje estadístico y apuntar que, sin cuestionar su validez, es posible detectar en ella una tendencia hacia cierto tipo de epistemología que rechaza la jerarquía causal; (4) exponer cómo desde diversas disciplinas era necesario o conveniente definir de manera precisa la causalidad probabilista estructural y desarrollar un método para razonar de manera formal en términos causales; a partir de estas motivaciones es posible entender claramente el propósito de los modelos causales; sin este propósito, los argumentos presentados —al menos el primero y el segundo— no tendrían sentido. (5) También fue necesario presentar formalmente la jerarquía causal, la definición formal de causa, al menos uno de los métodos para inferir relaciones de causalidad y la manera de calcular un efecto causal. Además, fue conveniente (6) presentar la discusión de la que forman parte estos argumentos.

Eso es todo lo que se dijo que se defendería en esta tesis: tres argumentos para proponer la preferencia de una epistemología causal para la actividad científica que estudia fenómenos macroscópicos y para la inteligencia artificial.

Hay algunos temas, sin embargo, que están relacionados con la argumentación que aquí se presentó, pero que no forman parte estrictamente de la conclusión. Tales temas se han decidido tratar aparte. En la siguiente sección el lector encontrará algunas pistas e ideas respecto de las consecuencias y el contexto de la argumentación que aquí se presentó. También encontrará algunas ideas respecto de lo puede hacerse para futuras investigaciones. Hay mucho trabajo por hacer.

Comentarios finales

Hay mucho trabajo por hacer en cuanto a la justificación y la comprensión filosófica de la noción de causalidad con la que aquí se trabaja, y del lugar que la causalidad ocupa en la actividad y las metas científicas. Las explicaciones y los argumentos que aquí se presentaron habrán mostrado ya que, a pesar de que algunos filósofos y científicos han propuesto que la causalidad no es una noción estrictamente científica, de hecho lo es. Sobre tales filósofos y científicos, además de lo mencionado en la sección IV, véase también Russell (1913). El que en algunos campos de investigación científica ésta no sea aplicable no obsta para que en muchos otros sea indispensable.

Una de las razones por las que se consideraba que la causalidad no es una noción estrictamente científica es que ésta no se había logrado medir como tal. Si Pearl está en lo correcto, entonces matematizó la causalidad y ello representa un hito en la historia de la cuantificación. Si no lo está, será porque podemos detectar con precisión fallas en su método, y entonces habremos progresado en claridad respecto de lo que una cuantificación de la causalidad debe cumplir.

Nadie negará que el desarrollo de un método matemático para cuantificar algo que antes no admitía un tratamiento matemático constituye una mejora para el conocimiento científico. Tener en mente algunos ejemplos puede ayudar a dimensionar qué significa una cuantificación así.

El movimiento como objeto de investigación científica es un caso histórico. Podemos trazar su historia —años más, años menos— desde Aristóteles hasta Galileo. Zenón de Elea no hallaba cómo sería viable la cuantificación del movimiento, y Newton trabajó ya a sabiendas de que había una manera de cuantificar el movimiento. Otros ejemplos son la definición de la función seno (Ptolomeo), la cuantificación del calor (Joule), y la cuantificación de la información (Shannon).

Muchos pueden negar que algún método matemático sea el correcto para cuantificar cierto fenómeno, o podría encontrarse que el método depende de una teoría cuya adecuación empírica no resulta convincente. Sin embargo, no se suele negar la posibilidad de cuantificar algo que ya ha sido

cuantificado (aun si lo ha sido erróneamente). En la historia de la cuantificación vemos muchas correcciones, pero pocas, o ninguna, retractaciones.

El autor de esta tesis no ha realizado una investigación histórica exhaustiva como para afirmar categóricamente que no existen tales retractaciones. No sería extraño encontrar algún caso atípico. Pero se puede constatar que las retractaciones en la historia de la cuantificación están lejos de ser comunes. Carlos Álvarez, historiador de la ciencia, afirma a propósito:

> La exactitud y la precisión siempre han sido consideradas como rasgos esenciales de todo conocimiento [...]. Si se entiende así el contraste entre la *verdad* que se busca en el conocimiento y el *error* que nos aleja de ella, nos parecería imposible reconocer algún caso en donde el abandono del camino que conduzca a la exactitud pudiera plantearse como un propósito explícito: En el extremo de la pureza, el conocimiento o es exacto o no es tal» (Álvarez & Martínez, 2004, p. ix).

Así que podemos preguntarnos, al decir que se ha matematizado la causalidad ¿estamos frente a un hecho irretractable, un hito científico? Lo que se ha expuesto en esta tesis apunta a responder que, en cierto sentido, sí lo es. No se ha presentado una argumentación para defender este punto (no era el objetivo), pero lo que se ha presentado puede aportar a la reflexión en torno a estas interrogantes.

Un segundo tema para estos comentarios finales es la teoría de la verdad compatible con los modelos causales estructurales. Los tres argumentos, en especial el primero, se centraron en cuestiones epistemológicas. La epistemología es la reflexión sobre el conocimiento como tal. Las definiciones del conocimiento más plausibles incluyen a la verdad como un aspecto indisociable de éste. Conocimiento y conocimiento verdadero son lo mismo. Esto implica que la teoría de la verdad está estrechamente asociada con la epistemología. Pero los argumentos que se presentaron en la Sección VI no consideran directamente la verdad de los modelos, sino su estructura epistemológica: los tres niveles de causalidad (observaciones, intervenciones y contrafácticos).

A despecho de que no sería parte del argumento, dada la cercanía filosófica de los temas, vale la pena bosquejar un poco la manera en que la

consideración de los modelos causales estructurales como objetos epistemológicos se relaciona con la teoría de la verdad.

Nótese que en el primer argumento se habló se la distribución de probabilidad *real* (al tratar la teoría de Vapnik) y del modelo *real* (al referirnos a Pearl). Podría parecer que no hay mucha diferencia entre hablar de la distribución real y la distribución verdadera, pero es complicado. Con los modelos causales es algo más complicado aún.

La teoría de la verdad es un campo de estudio vastísimo. El autor de esta tesis sabe que no les está haciendo justicia a las filósofas y los filósofos que han dedicado tantas páginas sobre este tema al decir lo siguiente: podemos dividir las teorías de la verdad en teorías *pictóricas* y teorías *pragmatistas*. Tal distinción, aunque tajante, permite mostrar algunas pistas del papel que la verdad desempeña en los dos esquemas epistemológicos que se contrastan en esta tesis.

La teoría pictórica propone que la verdad es una especie de *correspondencia* entre enunciados y hechos, representaciones y situaciones, o ideas y experiencias. La verdad se da cuando un objeto semántico (un enunciado, una representación, una idea, etc.) se corresponde con algún objeto de referencia (un hecho, una situación, una experiencia, etc.).

La teoría pragmatista describe a la verdad, en cambio, como una especie de *oportunidad*: el objeto semántico verdadero es el objeto oportuno.[27] No se evalúa en términos de su adecuación estrictamente, sino también en las consecuencias de tomarlo como verdadero. El pragmatista se pregunta ¿qué consecuencias tiene o qué diferencia provoca el tomar este objeto semántico como verdadero? No es que la correspondencia deje de ser relevante; contar con una representación adecuada es de lo más oportuno. Más bien sucede que no es el único criterio de verdad.

Podemos evaluar en términos pictóricos las funciones obtenidas con *Deep Learning*, o las de otros métodos basados en funciones. La suma de los cuadrados de los errores, el coeficiente de determinación en

[27] En lengua española, la diferencia entre algo oportuno y algo meramente conveniente está en que lo oportuno se realiza a propósito y su conveniencia guarda una relación directa con el tiempo en el que se realiza. Así, es un tanto abstracto hablar de un enunciado verdadero. Hablar de *oportunidad* en este sentido enfatiza, más que los enunciados verdaderos, el afirmar un enunciado en el momento oportuno.

regresiones lineales (conocido como R cuadrada) y el puntaje F1 en *Machine Learning*¸entre otras métricas, miden qué tanto se corresponden las funciones inferidas con los datos.

En cambio, a pesar de que hay medidas similares en la inferencia causal para evaluar la adecuación de los modelos, el hecho de que éstos refieran también a intervenciones (en las que hay un agente involucrado) y a contrafácticos (situaciones o eventos que discrepan de lo que ha sido observado) los muestran difíciles de capturar al estilo de la teoría pictórica de la verdad.

Uno de los supuestos que suele traer consigo la teoría pictórica de la verdad consiste en que hay una sola afirmación verdadera en cada situación, en cada contexto. La mente de Dios, para los filósofos medievales; el sujeto trascendental, para los kantianos. Si dos afirmaciones distintas son verdaderas respecto de lo mismo, al mismo tiempo, entonces son equivalentes, una es consecuencia de la otra o, al menos, son complementarias. No pueden diferir.

La teoría pragmatista, a diferencia, suele estar asociada con una postura pluralista: varios enunciados (ideas, representaciones, etc.) son candidatos a la validez sin que sean consistentes entre ellos. Ello no implica, conviene decirlo, el rechazo al principio de no contradicción, puesto que, a despecho de la inconsistencia entre ellos, sólo uno será afirmado. Lo que se está rechazando es un criterio absoluto como la mente de Dios según la describían los filósofos medievales o el sujeto trascendental.

Ahora bien, es posible que dos modelos causales estructurales se correspondan con las mismas observaciones, pero que impliquen distintas creencias respecto de los resultados de las posibles intervenciones. También es posible que dos modelos distintos se correspondan con las mismas observaciones y con los mismos resultados de intervenciones; pero nunca, en tales casos, a menos de que sean el mismo modelo, implicarán las mismas probabilidades contrafácticas. Que los contrafácticos no puedan ser capturados como una adecuación nos obliga a considerar otras cualidades además de la adecuación, cualidades como la simplicidad o el estado de las hipótesis y el conocimiento de los expertos en determinada área de

estudio. Lo cual hace que, entre un conjunto de modelos con el mismo nivel de adecuación, unos sean más oportunos que otros.

Por lo tanto, la elección de modelos en cuanto a su verdad está más cerca de una visión pragmatista que de la visión pictórica.

La argumentación que aquí se presentó también puede aportar a las reflexiones sobre cuestiones éticas y políticas de la IA. Sabemos que la computación ha estado desde sus orígenes estrechamente vinculada a propósitos militares y comerciales (véase Ceruzzi (2012) y Copeland (2020)). Actualmente contamos con investigaciones, argumentaciones y propuestas para entender, escoger o rechazar diversos rumbos hacia los que puede orientarse la tecnología digital; por ejemplo, en cuestiones de privacidad (Véliz, 2020), ecología y concentración de poder (Crawford, 2021), robots que toman decisiones (Lin *et al.*, 2017), entre otras.

Tal vez éste es el momento para decir que el presente trabajo es el resultado de una trayectoria personal. Luego de titularme de la licenciatura en filosofía, me encontraba un tanto insatisfecho por la poca comprensión técnica que mostraban algunos análisis y ensayos sobre ética y política de la IA y, en general, de la tecnología digital. También estaba deseoso de aprender y perfeccionar nuevos métodos de razonamiento. Ambas disposiciones me condujeron a la decisión de aprender a programar y, después, de aprender estadística y *Machine Learning* —y las matemáticas necesarias para ello, principalmente cálculo, álgebra lineal y probabilidad—. Es obvio que aún me falta mucho por aprender. Desde que comencé la travesía, y aún ahora, he recibido un enorme apoyo de profesores, profesoras, amigos y amigas.

A la mitad de ese camino, fue que descubrí el trabajo de Pearl. Mi intención era aprender más que lo que se necesita para dialogar y tener ideas precisas: aprender lo suficiente como para colaborar con ingenieros y científicos codo a codo. Hoy en día, trabajo como ingeniero de datos en *Ventagium Data Consulting*, una consultora de TI.

El diálogo entre humanistas y científicos históricamente ha sido árido, en parte porque las exigencias son muy altas (lo cual es adecuado, dada la importancia de la cuestión). Para realizar propuestas fructíferas se necesita juntamente un profundo conocimiento en ética y política, y un

profundo conocimiento de la tecnología digital. No es necesario que una sola persona sea experta en todo. El conocimiento se posee colectivamente. Pero compartir de ese modo el conocimiento no es posible si no exploramos, con sudor y dedicación, los demás campos de estudio. Cada vez será más imprescindible que los humanistas aprendamos matemáticas para que nuestros conocimientos sean relevantes, al menos para el área de la Inteligencia Artificial. Aunque no se puede ser experto en todo, sí se requiere cierto nivel de experiencia.

Además de los argumentos relacionados con la actividad y los proyectos científicos y con las interrogantes epistemológicas, espero que esta tesis contribuya a una comprensión conjunta de lo científico y lo ético. Entender los supuestos y las tendencias de los diversos métodos computacionales y estadísticos permitirá juzgar con mayor claridad cómo dirigirnos hacia el tipo de tecnología que buscamos para nuestras sociedades.

Glosario

aprendizaje profundo. En inglés se conoce como *Deep Learning*. Es el tipo de aprendizaje automático que se realiza con redes neuronales de varias capas. De manera informal, podríamos decir que a partir de las cinco capas ya se trata de aprendizaje profundo.

condición de Markov. Un modelo causal cumple con esta condición cuando cualquier variable en dicho modelo es independiente de todas las demás, dados sus padres.

DAG. Las siglas se deben al inglés *Directed Acyclic Graph*, Grafo Acíclico Dirigido. Es un conjunto de nodos y enlaces dirigidos. Se utilizan para representar las relaciones entre las variables de los modelos causales. Los nodos representan las variables, y los enlaces, las relaciones entre ellas. A la secuencia de variables unidas por enlaces en la misma dirección se les puede llamar *rutas*. Estos grafos son acíclicos porque la mima variable nunca aparece dos veces en la misma ruta.

modelo causal estructural. Una tripla <**U, V, F**>, donde **U** es el conjunto de variables circunstanciales (o exógenas), **V** es el conjunto de variables endógenas, y **F** es un conjunto de funciones que conforma un mapeo de **U** a **V**. Todo modelo causal 'M' puede asociarse a un DAG 'G(M)'.

red bayesiana causal. Es un DAG cuyos nodos representan variables y cuyos enlaces representan la dirección causal de las relaciones entre las variables. Una red bayesiana no es un modelo causal, pero puede estar asociada a uno.

retropropagación. Es el principal algoritmo de aprendizaje de las redes de aprendizaje profundo.

variables circunstanciales (U). Cualquier variable de un modelo causal cuyo valor no está determinado por otras variables del modelo.

variables endógenas (V). Cualquier variable cuyo valor esté determinado por el valor de otra variable del mismo modelo causal es una variable endógena.

Referencias

Anscombe, G. E. M. (1971). *Causality and Determination.* Cambridge University Press, London.

Alpaydin, Ethem (2021). *Machine Learning*, MIT Press, Cambridge, Massachusetts.

Álvarez, J., Carlos (2021). *Ensayos sobre Euclides. Volumen I. La geometría de la congruencia*, UNAM, México.

Álvarez J., Carlos, & Rafael Martínez (2004). "Introducción", en *Variar para encontrar*, UNAM, México, pp. ix-xv.

Bareinboim, E., Correa, J. D., Ibeling, D. & Icard, T. F. (2022). "On Pearl's Hierarchy and the Foundations of Causal Inference", *Probabilistic and Causal Inference,* ACM.

Bayes, Thomas (1763). "An Essay towards solving a problem in the doctrine of chances", *Philosophical Transactions* 53, pp. 370-418.

Bovens, Luc, & Stephan Hartmann (2003). *Bayesian Epistemology*, Oxford University Press, New York.

Cartwright, Nancy (2007). *Hunting Causes and Using Them*, Cambridge University Press, New York.

Cartwright, Nancy (1989). *Nature's Capacities and Their Measurement*, Clarendon Press, Oxford.

Ceruzzi, Paul E. (2012). *Breve historia de la computación.* Utilizo la siguiente edición: traducción de Ix-Nic Iruegas, Fondo de Cultura Económica, CDMX, 2018.

Choi, A., Wang, R. & Darwiche, A. (2019). "On the Relative Expressiveness of Bayesian and Neural Networks", *International Journal of Approximate Reasoning (IJAR)* 113, pp. 303-323.

Chollet, François, (2021). *Deep Learning with Python, Second Edition*, Manning Publications, E.U.A.

Copeland, Jack B. (2020). "The Modern History of Computing", en *The Stanford Encyclopedia of Philosophy* (Winter 2020 Edition), Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/win2020/entries/computing-history.

Crawford, Kate, (2021). *Atlas of AI: Power, Politics and the Planetary Costs of Artificial Intelligence*, Yale University Press, New Haven.

Darwiche, Adnan (2018). "Human-Level Intelligence or Animal-Like Abilities?", *Communications of the ACM*, vol. 61, No. 10, pp. 56-67.

Embrechts, P., A. McNeil & D. Straumann (2002). "Correlation and dependence in risk management: properties and pitfalls", en *Risk Management: Value at Risk and Beyond*, ed. M.A.H. Dempster, Cambridge University Press, Cambridge, pp. 176-223.

Floridi, Luciano & Jules Desai, et al. (2022). "The Epistemological Foundations of Data Science: A Critical Review", en *Synthese*, (2022) 200:469, doi: https://doi.org/10.1007/s11229-022-03933-2

Fuentes García, Ruth, et al. (2019). *Inferencia estadística para estudiantes de ciencias*, Facultad de Ciencias, UNAM, CDMX.

Goodfellow, I., Bengio & Courville (2016). *Deep Learning*, The MIT Press, Cambridge, Massachusetts.

Heckman, J., & Pinto, R. (2015). "Causal Analysis after Haavelmo", *Econometric Theory,* Vol. 31, pp. 115-151.

Hinton, G., Sutskever, I., & krizhevske, A. (2012). "ImageNet Classification with Deep Convolutional Neural Networks", *Advances in Neural Information Processing Systems 25 (NIPS 2012)*.

Hinton, G. (2022). "The Forward-Forward Algorithm: some Preliminary Investigations", *Google Brain* (https://www.cs.toronto.edu/~hinton/absps/FFXfinal.pdf). Revisado el 3 de julio de 2023.

Hume, David, (1772), *An Enquiry concerning Human Understanding (edited by Tom L. Beauchamp)*, Oxford University Press, New York, 2000.

Hung, R. J., McKay, J. D., Gaborieau, V., *et al.* (2008) "A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25", *Nature* 452, pp. 633-637.

James, W., (1912). *Essays in Radical Empiricism*, Longmans, Green and Co., New York.

Jeong, H., *et al.* (2022). "Mesolimbic dopamine release conveys causal associations", *Science* 378, eabq6740.

Lee, Wei-Meng (2019). *Python Machine Learning*, Wiley, Indianápolis, Indiana.

Lillicrap T., Santoro, A., Marris, L., Akermann, C., Hinton, G. (2020). "Backpropagation and the Brain", *Nature Reviews Neuroscience* 21, pp. 335-346.

Lin, Hanti (2022). "Bayesian Epistemology", The *Stanford Encyclopedia of Philosophy (Fall 2022 Edition)*, Eduard N. Zalta & Uri Nodelman (eds.), URL = https://plato.stanford.edu/archives/fall2022/entries/epistemology-bayesian/

Lin, P., Jenkins, R., & Abney, K., (eds.) (2017). *Robot Ethics 2.0*, Oxford University Press, New York.

McCarthy, J., Marvin Minsky *et al.* (1955). *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence.* Utilizo el facsimilar que publicó *AI Magazine,* Volumen 7, Número 4, 2006.

McCarthy, J. (1960). "Recursive Functions of Symbolic Expressions and Their Computation by Machine", *Communications of the ACM*. Vol. 3, Issue 4, April 1960, pp. 184-195.

McCulloch, Warren, S., & Walter Pitts (1943). "A Logical Calculus of the Ideas Immanent in Nervous Activity", en *Bulletin of Mathematical Biophysics,* volume 5.

Méndez-Molina, A., Sucar, L. E., & Morales, E. F. (2022). "Causal Discovery and Reinforcement Learning: A Synergistic Integration", *Proceedings of The 11th International Conference on Probabilistic Graphical Models,* PMLR 186, pp. 421-432.

Minsky, Marvin, & Seymour A. Papert (1988). *Perceptrons (Expanded Edition)*, MIT Press, Cambridge, Massachusetts.

Minsky, Marvin (ed.) (1968). *Semantic Information Processing*, MIT Press, Cambridge, Massachusetts.

Morgan, Stephen L. & Christopher Winship (2015), *Counterfactuals and Causal Inference (Methods and Principles for Social Research, Second Edition)*, Oxford University Press.

Pearl, J. (2018). *The Book of Why*, Basic Books, Nueva York.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference (Second edition)*, Cambridge University Press, Cambridge, UK.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems. Networks of Plausible Inference*, Morgan Kaufmann.

Pearl, J. (1985). "Bayesian networks: A model of self-activated memory for evidential reasoning", *Proceedings, Cognitive Science Society*, pp. 329-334, Irvine, CA, 1985.

Rosenblatt, Frank (1957). "The Perceptron—A Perceiving and Recognizing Automaton", Report 85-460-1, Cornell Aeronautical Laboratory.

Rumelhart, David E., Hinton & Williams (1986a). "Learning representations by back-propagating errors", *Nature*, vol. 323, pp. 533 – 536.

Rumelhart, David E, & Grupo de investigación PDP (1986b). *Parallel Distributed Processing. Explorations in the Microstructure of Cognition*, MIT Press, Cambridge, Massachusetts.

Russell, B., (1913). "On the notion of cause", *Proceedings of the Aristotelian Society*, vol. 13, pp. 1-26.

Russell, Stuart & Norvig, Peter (editors) (2021). *Artificial Intelligence. A Modern Approach (Fourth Edition)*, Pearson, E.U.A.

Salmon, W. (1988). *Causality and Explanation*, Oxford University Press, New York.

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

Sejnowski, Terrence J. (2018). *The Deep Learning Revolution*, The MIT Press, Cambridge, Massachusetts.

Simen, P., & Mattel, M. (2016). "Why does time seem to fly when we're having fun?", *Science* 354, pp. 1231-1232.

Spade, Paul Vincent & Claude Panaccio (2019). "William of Ockham", en *Stanford Encyclopedia of Philosophy (Spring 2019 Edition),* Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/spr2019/entries/ockham/.

Suppes, P., (1970). *A Probabilistic Theory of Causality*, North-Holland Publishing Co., Amsterdam.

Thorgeirsson, T. E., Geller, F., Sulem, P., *et al.* (2008). "A variant associated with nicotine dependence, lung cancer and peripheral arterial disease", *Nature* 451, pp. 638-642.

van Fraassen, Bas, (1989). *Laws and Symmetry*, Oxford University Press, Nueva York.

van Fraassen, Bas, (1980). *The Scientific Image*, Oxford University Press, Nueva York, capítulo 5, "The pragmatics of explanation".

Vapnik, Vladimir N. (2000). *The Nature of Statistical Learning Theory*, Springer, New York.

Vapnik, Vladimir N., & Rauf Izmailov (2020). "Complete Statistical Learning Theory (Learning Using Statistical Invariants)", *Proceedings of Machine Learning Research* 128:1-37.

Véliz, Carissa (2020). *Privacidad es poder*, Editorial Debate, primera edición en México, CDMX: 2022.

VanderWeele, T. J. (2014). "A Unification of Mediation and Interaction", *Epidemiology*, vol. 25, No. 5, September, pp. 749-761.

Wolpert, David, H. (1996). "The Lack of a Priori Distinctions Between Learning Algorithms", MIT, *Neural Computation* 8, pp. 1341-1390.

Yule, Udny (1899). “An Investigation into the Causes of Changes in Pauperism in England, Chiefly During the Last Two Intercensal Decades (Part I.)”, *Journal of the Royal Statistical Society*, vol. 62, No. 2 (Jun., 1899), pp. 249-295.