Mauricio Molina-Delgado y Eugenia Gallardo-Allen

El principio de composicionalidad y los algoritmos de aprendizaje de máquina

Resumen: *El presente artículo explora las consecuencias que las críticas de Fodor y Pylyshyn en 1988 hacia las representaciones conexionistas podrían tener con respecto a los desarrollos recientes del Aprendizaje de Máquina. El cuestionamiento que estos autores realizan se suele denominar el reto de la sistematicidad. En particular, se analizará aquí el llamado principio de composicionalidad, el cual establece que el significado de una oración se determina por el significado de sus partes, principio del que según los citados autores no se puede dar cuenta mediante representaciones distribuidas (subsimbólicas). Para abordar el punto, se realiza una radiografía sobre el estado de la Ciencia Cognitiva y la Inteligencia Artificial (IA) a finales del siglo XX, en donde se emplea la distinción entre enfoques simbólicos y subsimbólicos; y se analiza el desarrollo de ambas disciplinas durante las primeras décadas del siglo XXI. Se concluye que las herramientas estadísticas adoptadas por el campo del Aprendizaje de Máquina mantienen características que permiten dividir estas técnicas mediante la distinción simbólico/subsimbólico, de modo que los argumentos de Fodor y Pylyshyn pueden en principio aplicarse a dichas técnicas. Las consecuencias de esto estarían emparentadas con el llamado problema de opacidad epistémica.*

Palabras clave: *Ciencia Cognitiva, Inteligencia Artificial, Redes Neuronales, principio de composicionalidad, aprendizaje de máquina.*

Abstract: *This article explores the consequences that Fodor and Pylyshyn's 1988 criticism of connectionist representations could have with respect to recent developments in Machine Learning. The questioning that these authors raise is usually called the challenge of systematicity. In particular, the so-called compositionality principle will be analyzed here, which establishes that the meaning of a sentence is determined by the meaning of its parts, a principle that according to the aforementioned authors cannot be accounted for through distributed (subsymbolic) representations. To address the point, we present an overview the state of Cognitive Science and Artificial Intelligence (AI) at the end of the 20th century, where the distinction between symbolic and subsymbolic approaches is used; and the development of both disciplines during the first decades of the 21st century is analyzed. It is concluded that the statistical tools adopted by the field of Machine Learning maintain characteristics that allow these techniques to be divided through the symbolic/subsymbolic distinction, so that the arguments of Fodor and Pylyshyn can in principle be applied to these techniques. The*

consequences of this would be related to the so-called problem of epistemic opacity.

Keywords: *Cognitive Science, Artificial Intelligence, Neural Networks, compositionality principle, Machine Learning.*

La revolución cognitiva

Para entender el panorama de la IA y la Ciencia Cognitiva del siglo XX resulta tentador intentar una aproximación kuhniana que establezca el desarrollo de posibles paradigmas de dicha disciplina. Con el objetivo de dibujar un panorama de la evolución de estas disciplinas asumimos aquí dicha tarea desde una perspectiva instrumentalista, sin mayores compromisos con el planteamiento de Kuhn (1962).

Un buen punto de partida para esta tarea es el trabajo de Claudio Gutiérrez, quien en diversos artículos plantea una serie de interpretaciones kuhnianas del desarrollo de la Ciencia Cognitiva y de las disciplinas que la conforman. En su conferencia de 1985 sugiere que el trabajo de Allan Turing (1937), donde se describe la llamada máquina universal, inaugura un nuevo paradigma para las ciencias del conocimiento y en particular para la psicología, disciplina que inicialmente se habría desarrollado dentro del paradigma introspeccionista[1] de Wundt (Wunt y Pintner 2016), moviéndose luego hacia el conductismo de raíz pavloviana y finalmente, bajo la influencia de Turing al computacionalismo. La misma posición sería presentada en un artículo de 1987, donde se refiere a la interpretación de Boden (1984) sobre el artículo de Turing como surgimiento de la IA.; aquí Gutiérrez adjudica nuevamente a la IA el haber provocado un cambio paradigmático en la psicología frente al anterior paradigma conductista, pero en este artículo no hace referencia al pasado introspeccionista de esta ciencia. Siguiendo esta línea de desarrollo, la psicología habría logrado un salto metodológico importante al pasar de la introspección de los estados internos al estudio experimental de la conducta observable, sin embargo, este paso implicaba una pérdida importante: el renunciar al uso de términos mentales (intensionales)[2] en las explicaciones psicológicas. La revolución computacional, permitió recuperar el vocabulario de las actitudes proposicionales y los estados intensionales, bajo el entendido de que dichos estados podían ser entendidos en términos de algoritmos y estructuras. En otra parte, Gutiérrez (1993) plantea la posibilidad de considerar a la informática como una ciencia empírica y sugiere que la hipótesis de los sistemas de símbolos físicos de Newell y Simon (1976) sería el trabajo que logró dotar a esta ciencia de un paradigma, enfoque al que denomina paradigma funcionalista, palabra que se refiere a una posición filosófica respecto de la mente de la cual hablaremos más adelante. En el mismo texto, Gutiérrez afirma que la Ciencia Cognitiva antes de contar con el paradigma computacional, empleó un paradigma basado en la investigación animal (propio del conductismo). La aparente confusión que encontramos en los distintos trabajos de Gutiérrez en realidad responde a una gran coherencia en su planteamiento. Además del hecho de que la terminología de Kuhn resulta a menudo difusa, hay que recalcar que estamos en un momento en que la conformación de las disciplinas consideradas y sus límites están en constante cambio. De hecho, esta situación es aún la norma ante la proliferación de campos interdisciplinarios como la Ciencia de Datos. Lo que se desprende del análisis de Gutiérrez es que para fines del siglo XX, diversos campos tienden a agruparse bajo la etiqueta común de Ciencia Cognitiva y que esta disciplina cuenta en general con un marco de entendimiento común que parte de los dos artículos seminales de Turing, se alimenta de la reacción anti-conductista en psicología (y quizás de los mismos desarrollos tardíos del conductismo como es el caso de Tolman, 1932), de la evolución de la informática bajo la hipótesis de los sistemas de símbolos físicos, así como de los planteamientos funcionalistas en filosofía de la mente. En el siguiente apartado intentaremos analizar la naturaleza de dicho programa de investigación y la presencia a lo largo del siglo XX de propuestas alternativas para enfocar el estudio de la cognición.

Afinidades y disidencias en la Ciencia Cognitiva del siglo XX

Dependiendo de lo que consideremos el nacimiento de una disciplina podríamos decir que las Ciencias Cognoscitivas surgen entre las décadas de los años 50 y 70 del siglo XX (Miller 2003). Siguiendo la discusión de la sección anterior podemos aceptar la existencia de un enfoque dominante en esta disciplina marcado por la hipótesis de los sistemas de símbolos físicos. Sin embargo, la noción de una representación simbólica de los fenómenos cognitivos puede rastrearse hasta el artículo seminal de Turing (1937) donde se define la llamada máquina universal. Es interesante que pocos años después de la publicación de Turing aparece una propuesta alternativa por parte de McCulloch y Pitts (1943) la cual busca modelar el cálculo proposicional mediante un modelo computacional inspirado en los conocimientos de la estructura neuronal del cerebro. Para ello, se emplea una red neuronal artificial cuyos nodos adquieren valores de 0 o 1 para representar su activación. El modelo emula además la sinapsis mediante la aplicación de una función sobre la suma ponderada de los valores de activación de otros nodos. Nótese que la red descrita no requiere de la existencia física de los citados nodos, de modo que se trata simplemente de variables que ocupan espacio en la memoria de una computadora. Esto implica que una red neuronal artificial como la que tempranamente definen McCulloch y Pitts puede implementarse en una máquina de Turing. A pesar de esto, el enfoque computacional de esta propuesta difiere del simbolismo planteado por Newell y Simon (1976) en términos de la ontología con la que ambas posiciones se comprometen y en el modo en que las entidades postuladas asumen significado. Así, la semántica de una red neuronal no necesariamente se concentra en los nodos sino en la totalidad de la red, de modo que no requiere de entidades como los símbolos físicos. Si pensamos en las denominadas redes semánticas (Collins y Quilliam 1969), modelo que Gutiérrez considera como una formulación fundamentalmente equivalente a la de Newel & Simon, aquí cada nodo se refiere a una entidad particular, mientras que los vínculos entre nodos corresponden propiamente a los objetos. Así, un nodo de una red semántica podría corresponder a la representación del concepto de «perro» y otro al de «mamífero» mientras un vínculo ES-UN (IS-A) podría indicar que «el perro es un mamífero». En contraste, las representaciones en una red neuronal estarían distribuidas en la red, de modo que se definirían no por nodos específicos sino por los patrones de activación. En realidad, esta diferencia no podría quedar tan clara en el proyecto esbozado por McCulloch & Pitts (1943), pero sí en los modelos conexionistas desarrollados por el llamado grupo PDP (Procesamiento distribuido en paralelo) en los años 80 del siglo XX (por ejemplo, Rumelhart, Hinton y Williams 1986). Entre ambos momentos, habría que señalar el decisivo papel que juega la publicación de Minsky & Papert (1969) la cual condena el incipiente proyecto conexionista a una época oscura. El problema al que apuntan estos autores es el hecho de que el perceptrón, término acuñado por Rosemblat (1958) que se refiere a una red neuronal con varios nodos organizados en 2 capas: una capa de entrada y otra de salida (del cual la neurona de McCulloch & Pitts resulta ser un caso particular) es incapaz de modelar el XOR (o exclusivo). Es importante señalar que desde los años 30 las redes neuronas artificiales se muestran como una excelente opción para explicar el aprendizaje frente a los modelos simbólicos. Dichas redes cuentan con mecanismos para ajustar sus parámetros a partir de los datos, lo que puede ser interpretado como un modelo de aprendizaje. Sin embargo, el problema del XOR parece plantear un límite definitivo al proyecto conexionista, ya que una red como el perceptrón simple no es capaz de ajustar sus parámetros para aprender esta función lógica. El problema puede resolverse incluyendo una capa denominada oculta, intermedia entre las entradas y salidas, pero en el momento en que Minsky y Papert publican su crítica no existe ningún algoritmo de aprendizaje para el perceptrón multicapa, lo que provoca que la investigación en redes neuronales se detenga por varias décadas. La trayectoria de la perspectivas simbolista y conexionista puede analizarse desde la perspectiva de Lakatos (1978), en el sentido que las teorías conexionistas

aparecieron como un programa de investigación regresivo en los años 60s y 70s del siglo XX, mientras aquellas propias del simbolismo se comportaban de forma progresiva. Con el desarrollo del algoritmo de propagación hacia atrás o retropropagación (Backpropagation) en los años 80, se resuelve el problema de dotar al perceptrón multicapa de un procedimiento de aprendizaje.

El panorama posterior puede describirse como la competencia entre dos tradiciones que muestran fortalezas y debilidades comparativas en diversos frentes. Mientras que el simbolismo presenta importantes avances en el desarrollo de algoritmos para la resolución de problemas seriales y para modelar el conocimiento experto, el conexionismo muestra ventajas al analizar la percepción, la formación de conceptos y el aprendizaje.

Debe recalcarse el hecho de que, a pesar de que los modelos conexionistas se inspiran en las redes neuronales naturales, ambas posiciones hacen uso de la llamada metáfora computacional y dependen fuertemente de la idea de representaciones del conocimiento. Al respecto, Smolensky (1988) hace una importante reflexión sobre los supuestos teóricos de ambas aproximaciones. Para ello, este autor se refiere al conexionismo como un enfoque subsimbólico en contraste con el enfoque que hemos denominado simbólico. Consideramos que esta distinción resulta acertada para entender las diferencias entre ambos programas de investigación, por lo que adoptaré para el resto de la discusión la terminología empleada por Smolensky.

Para seguir su planteamiento es importante introducir el tema de los niveles de explicación. Una de las primeras consecuencias de esto tiene que ver con el problema de los niveles. Diversos autores como Marr (1982) y Daniel Dennett (1981) han diferenciado tres niveles de explicación relevantes dentro de la Ciencia Cognitiva. Gutiérrez (1993) siguiendo a Dennett caracteriza estos tres niveles del siguiente modo:

1. Nivel físico
2. Nivel de diseño
3. Nivel intensional

El nivel físico se refiere a la base material sobre la que se implementa la cognición. Dicha base podría referirse a las neuronas en el caso de un ser humano o a los circuitos de una computadora. En el caso de esta última, el nivel de diseño se relaciona con la programación que permite a la máquina realizar las tareas. Finalmente, el nivel intensional hace referencia a la descripción en psicología popular (folk psychology) que podemos realizar en términos mentales, incluyendo deseos y creencias (actitudes proposicionales). Gutiérrez (1985) utiliza el ejemplo de una máquina que juega ajedrez para explicar los 3 niveles. Al describir una jugada realizada por la computadora, podríamos decir que «la computadora adelantó el caballo porque cree que el flanco del rey está débil y quiere fortalecerlo» (Epistemología e Informática, 226 Completas, IV). La posición funcionalista en filosofía de la mente puede ser descrita por su relación con estos tres niveles. En primer lugar, el funcionalismo asume que las explicaciones a nivel intensional tienen perfecto sentido tanto para la máquina como para el jugador humano. En segundo lugar, considera que las implementaciones a nivel de diseño tienen un papel fundamental en el desarrollo de la Ciencia Cognitiva. Nótese que al describir el nivel de diseño se hizo referencia al caso del computador, pero no del ser humano. La razón de esto es que no contamos con un modo de acceder directamente a estos programas, pero según el funcionalismo, la implementación computacional sirve como un modelo explicativo de la misma cognición humana. Es este precisamente el proyecto del simbolismo, implementar computacionalmente modelos que den cuenta de lo que ocurre a nivel intensional tanto en el ser humano como en la computadora. En otras palabras, los fenómenos intensionales constituyen el «explanandun» mientras que el diseño en términos de programación sería el «explanans». Finalmente, el funcionalismo considera que el papel del nivel físico es relativamente irrelevante para explicar la cognición. Ciertamente, cuando existen fallos en la cognición podríamos recurrir al nivel físico para encontrar una razón, pero normalmente un mismo evento mental podría ser implementado mediante diversas bases materiales, lo que resta fuerza explicativa al nivel físico.

El argumento central de la propuesta de Smolensky (1988) es que, a pesar de las apariencias,

la posición subsimbólica no busca explicaciones de la cognición a un nivel de implementación física como sí ocurre, por ejemplo, en las posiciones eliminativistas defendidas por Patricia Churchland-Smith (1986) y Paul Churchland (1984) o en los planteamientos de la teoría de la identidad en filosofía de la mente (Place 1956).

Para explicar la diferencia entre la posición simbolista y la subsimbólica, (1988) introduce un nuevo nivel. Es evidente que los niveles planteados son una abstracción que podría realizarse con mayor o menor granularidad (Bermúdez 2014). Así, el nivel físico en realidad podría subdividirse (para el caso de un animal humano o no humano) en un nivel químico, un nivel de neuronas, un nivel de circuitos neuronales, etc. Similarmente, al pensar en el nivel de diseño podríamos considerar algoritmos de alto nivel o bien en una descripción en lenguaje de máquina. Valiéndose de esto y considerando que las descripciones del nivel de diseño corresponden a los llamados símbolos físicos, Smolensky sostiene que los modelos de explicación de la cognición deberían plantearse a un nivel más bajo, que sin ser propiamente físico emplee una abstracción de las propiedades formales de las redes neuronales naturales. Esta sería la estrategia subsimbólica, que no recurre a neuronas físicas sino a nodos con valores de activación y en lugar de preocuparse por las propiedades químicas de la sinapsis emplea funciones matemáticas que al aplicarse a las salidas de los nodos de una capa determinan la posible activación de los nodos de la siguiente (ver Figura 1).

Esta formulación deja en claro que la propuesta subsimbólica recurre a representaciones y algoritmos (al menos en cuanto a algoritmos de aprendizaje) tanto como las explicaciones generadas por el enfoque simbólico, solo que las representaciones de la primera están distribuidas en los patrones de activación de la red y no localizadas en los nodos individuales. Esto es particularmente notable en los nodos de las capas ocultas, cuyos patrones de activación carecen de una interpretación clara. Por el contrario, cada nodo de un modelo simbolista es, valga la redundancia, un símbolo, y como tal podría representar el concepto de «perro» mientras otro nodo podría representar el concepto de «viejo pastor ovejero inglés». En el enfoque subsimbólico, una misma red podría representar ambos conceptos mediante distintos patrones de activación, siendo el caso que resultaría imposible identificar el valor semántico de nodos específicos.

El auge de los modelos subsimbólicos generó, a pesar de su éxito para explicar el aprendizaje de conceptos, una importante crítica por parte de Fodor y Pylyshyn (1988). A grandes rasgos, estos autores señalan que las explicaciones subsimbólicas carecen de sintáxis y semánticas combinatorias. A pesar de esta crítica planteada hace más de 30 años, las herramientas subsimbólicas siguen teniendo un papel relevante en las aplicaciones actuales de aprendizaje de máquina; sin embargo, el argumento es sólido y será central para valorar los desarrollos recientes de la I.A.

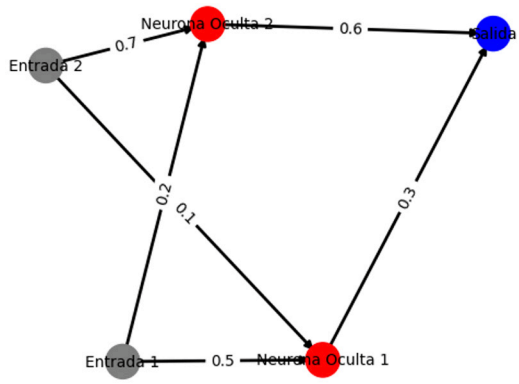

Figura 1. Red neuronal Perceptron con una capa de entrada, una capa oculta y una capa de salida.

La composicionalidad y las redes subsimbólicas

En el anteriormente citado artículo seminal de 1988, Fodor y Philyshyn defienden la tesis de que el nivel de explicación empleado por los enfoques conexionistas (subsimbólicos) no es el correcto para modelos explicativos en Ciencia Cognitiva. Su línea de argumentación contiene los siguientes aspectos:

1. El pensamiento es sistemático, de modo que si entendemos una proposición del tipo A mira a B, podemos también entender otra como B mira a A.
2. Para garantizar la sistematicidad descrita en 1) las representaciones deben estar vinculadas de forma sintáctica y semántica.
3. Las redes conexionistas no logran dar cuenta de estas relaciones sintáctico/semánticas.

Como señalan Symons y Calvo (2014), el punto 2 puede relacionarse con la llamada hipótesis del Lenguaje del pensamiento (LOT, Language of Thought) enunciada unos años antes por Fodor (1975). Según el planteamiento de Fodor, dado que el valor de verdad de una proposición compuesta depende del valor de verdad de sus partes, y el hecho de que las reglas lingüísticas de transformación pueden dar cuenta de este carácter combinatorio, es posible asumir la existencia de una estructura similar subyacente que dé cuenta de la sistematicidad del pensamiento. Así, el valor de verdad de la proposición compuesta depende de los valores de verdad de *P* y de *Q*, así como de las funciones conectivas de la conjunción y la negación. A este principio se le conoce como *principio de composicionalidad*. Fodor y Pylyshyn (1988), apelan a este principio como una condición fundamental de la que debe dar cuenta cualquier estructura que pretenda explicar la cognición. Además, argumentan que, a diferencia de los planteamientos tradicionales de la I.A. simbolista, las propuestas subsimbólicas carecen de los mecanismos necesarios para dar cuenta de la composicionalidad y en consecuencia no logran pasar el reto de la sistematicidad.

Por el momento, podemos concluir este panorama de la Ciencia Cognoscitiva de fines del siglo XX señalando que los dos programas de investigación señalados como dominantes hacían uso de la IA como una vía fundamental para el desarrollo de modelos de la cognición. Conviene tener este detalle en mente ya que para fines de dicho siglo el funcionalismo recibiría fuertes ataques en especial por su desdén hacia el papel explicativo de las bases biológicas de la cognición, un aspecto del que el enfoque subsimbólico no se apartaba radicalmente. Por otra parte, pronto el énfasis de los desarrollos en IA también variaría de forma determinante.

Aprendizaje estadístico

A más de 30 años de la propuesta de la hipótesis de símbolos físicos, el estado de la Ciencia Cognitiva y la inteligencia artificial ha variado sustancialmente (Simons y Calvo 2014) existiendo actualmente una multiplicidad de enfoques novedosos en ambas disciplinas (cognición encarnada, cognición situada, cognición extendida, sistemas dinámicos, etc.). Dentro de este complejo espectro de posiciones y enfoques, pretendemos analizar la evolución del campo conocido como aprendizaje de máquina (machine learning), el cual ha recogido muchos de los aportes de los modelos conexionistas de fines del siglo XX.

Para entender el papel que juegan los modelos subsimbólico dentro del aprendizaje de máquina es necesario considerar que estos incluyen diversos mecanismos que pretenden dar cuenta de los procesos de aprendizaje (supervisado, no supervisado, por reforzamiento, etc). El éxito que este enfoque obtiene en cuanto a su consideración del aprendizaje parece depender en buena medida de su capacidad para analizar patrones estadísticos en los datos.

Sin embargo, esto no necesariamente es una característica exclusiva de las representaciones distribuidas, de modo que los modelos simbólicos también pueden extraen regularidades estadísticas que permitan emular un proceso de aprendizaje. Es cierto que las redes neuronales de perceptrón multicapa cuentan con la ventaja de ser aproximadores universales (Hornik et al. 1989) es decir, que son capaces de aproximar cualquier función medible con un grado de exactitud establecido. Sin embargo, en general los modelos simbólicos pueden generar aproximaciones suficientemente razonables para ajustar patrones de un conjunto de datos, empleando para ello modelos lineales o no lineales.

El considerar el aprendizaje como una serie de mecanismos para extraer patrones de un conjunto de datos explica el que diversos modelos estadísticos puedan ser interpretados como

algoritmos de aprendizaje. Esto incluye modelos tradicionales como el análisis de regresión múltiple, el cual típicamente utiliza el algoritmo de mínimos cuadrados para aprender patrones lineales minimizando las distancias al cuadrado de la diferencia entre los valores de una variable independiente respecto de una recta de mejor ajuste. Efectivamente, el desarrollo de los algoritmos y técnicas de análisis de datos desarrollados por la estadística clásica durante los siglos XIX y XX fueran apropiados por campos de la IA como el aprendizaje de máquina. Esto incluye a técnicas regresión y clasificación, como los árboles de decisión y el modelo lineal generalizado (por ejemplo, la regresión binomial y logística en tanto formas de aprendizaje supervisado), así como el análisis de conglomerados en sus diversas formas (en tanto aprendizaje no supervisado de categorías). Además, podemos considerar la extracción de dimensiones latentes mediante el análisis de factores y el análisis de componentes principales, técnicas que pueden interpretarse también como una suerte de aprendizaje no supervisado involucrado en la formación de conceptos abstractos. Lo común de estos modelos estadísticos clásicos es que por su naturaleza podemos considerarlos parte del enfoque simbólico, si bien en su gran mayoría no fueron desarrollados originalmente para ser empleados en IA. La razón es que los elementos de este tipo de análisis estadístico son variables con una semántica propia, a diferencia de aquellas que constituyen la capa oculta de una red neuronal. Consideremos un ejemplo típico de aprendizaje de máquina en el que se aplica una regresión logística para predecir si un préstamo acabará en un estado de no pago. Cada una de las variables involucradas en el modelo tiene un carácter simbólico (pago o impago del préstamo, salario del beneficiado, historial crediticio, etc.)

Lo anterior nos permite mantener una división similar a la planteada anteriormente entre los modelos simbólicos, provenientes en su mayoría de la estadística frecuentista clásica desarrollada a partir de los enfoques de Fisher y Neymar Person (Acree 2021), y los modelos subsimbólicos, que actualmente se agrupan bajo la denominación de «Deep learning».

Por otra parte, existen dos aspectos que deberíamos considerar para valorar los cambios que se han dado tanto en la Ciencia Cognitiva como en la IA durante el siglo XXI. En primer lugar, es notable el hecho de que la Ciencia Cognitiva ha dado una especie de giro biológico que la aleja de los dogmas funcionalistas del siglo anterior.[3] En cuanto a la IA, en los últimos años se han dado grandes avances en diversos campos; tal es el caso del procesamiento de lenguaje natural. Sin embargo, los cuestionamientos al valor de la prueba de Turing (1950) y el declive del funcionalismo, hacen que estos impresionantes éxitos hoy no representen, como habría sido hace 50 años, la llave para develar la naturaleza de la inteligencia. Por otra parte, en el imaginario social, el concepto de Inteligencia artificial ha quedado prácticamente reducido al campo del aprendizaje de máquina.

Los dos aspectos señalados se pueden resumir diciendo que la crítica realizada por Searle (1980) en su famoso «argumento del cuarto chino» parece haber tenido un peso decisivo en el desarrollo de la Ciencia Cognitiva lo mismo que en el de la I.A. Searle plantea un experimento mental donde una persona encerrada en un cuarto recibe del exterior una serie de textos y preguntas sobre esos textos, todo escrito en caracteres chinos. Aunque la persona no conoce en absoluto esta lengua, cuenta con un libro que le permite responder las preguntas mediante una serie de reglas formales; es decir, que ante la presencia de ciertos caracteres en la pregunta y en el texto este manual le señala el modo en que debe responder empleando también caracteres chinos. Esta situación resulta análoga a la de la prueba de Turing. Según Searle, la computadora que se hace pasar exitosamente por un hombre o una mujer, lo mismo la persona que pretende engañar a sus interlocutores haciéndose pasar por alguien que domina el chino, no harían otra cosa que una manipulación sintáctica sin poseer ningún estado propiamente intensional (en el caso de la persona obviamente los tendría, pero no relacionados con la comprensión del texto). Searle plantea dos consecuencias de su argumento: la primera, que la IA fuerte, es decir, aquella que pretende explicar la cognición, corresponde a un proyecto destinado al fracaso; no así la IA débil, la cual

busca simplemente el desarrollo tecnológico de herramientas inteligentes sin pretensiones teóricas. La segunda consecuencia es que la única máquina que tiene y podría tener estados mentales debe ser una máquina biológica, de modo que según Searle la cognición depende de los poderes causales de la base material.

Pasados 23 años del siglo XXI, los planteamientos de Searle resuenan como profecías: hoy la IA débil se desarrolla con fortaleza en campos como el reconocimiento de patrones y la ciencia de datos, en detrimento de la IA fuerte; mientras la ciencia cognitiva reconoce ampliamente el papel causal de la biología en las explicaciones de los fenómenos mentales.

Sobre las consecuencias de un nuevo imperio de la mente

Los planteamientos críticos contra el funcionalismo por parte de autores como Searle y Penrose causaron una fuerte polémica en su momento. Como ejemplo, considérese el texto de Gutiérrez (en Epistemología e informática, 1993) titulado «Del cuarto chino y otros cuentos de hadas». Hoy en día, muchas de esas críticas han sido asumidas al interior de la Ciencia Cognitiva y la IA, disciplinas que han redescubierto el papel de la corporalidad y los afectos en los procesos cognitivos frente a la formulación aséptica y etérea del funcionalismo.

El rumbo tomado por la IA y en particular por el campo del aprendizaje de máquina parece también haber resultado sumamente productivo, dando pie al desarrollo de nuevas disciplinas como la Ciencia de datos.

A pesar de todos estos aspectos positivos en los que ambas disciplinas han evolucionado de forma más o menos independientes, pensamos que la IA debería seguir teniendo un papel fundamental en el desarrollo de la ciencia cognitiva. La dirección opuesta es igualmente trascendental: dejar de lado las implicaciones de los hallazgos de la ciencia cognitiva posiblemente iría en detrimento de los logros en la IA.

Valga simplemente señalar dos posibles elementos que muestran las implicaciones mutuas entre ambas disciplinas. En primer lugar, queremos argumentar aquí que las aplicaciones actuales de Deep learning adolecen del problema de falta de sistematicidad señalado por Fodor y Pylyshyn (1988) con respecto a los modelos subsimbólicos del siglo XX. Este problema estaría relacionado con el de la opacidad epistémica de los modelos (Stacewicz et al. 2021).

Para mostrar el punto, baste decir que en general los modelos de Deep learning emplean variables y parámetros que no cuentan con contenido semántico (tal como se ha dicho respecto de los nodos de las capas ocultas), y que la interacción de estas unidades no respeta el principio de composicionalidad. En contraste, las variables de otros modelos como los árboles de decisión y la regresión sí cuentan con una semántica propia y pueden combinarse siguiendo el principio de composicionalidad. Una prueba de esto es el hecho de que los parámetros de una regresión permiten ser expresados mediante lógica de predicados. Considérese la siguiente ecuación de mejor ajuste:

$$Y_i = \beta_0 + \beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i} + e_i$$

El significado de esta ecuación puede ser presentada mediante una descripción en lenguaje natural o bien con la siguiente expresión:

$$\beta_0 (Y) \wedge \beta_1 (Y, X_1) \wedge \beta_2 (Y, X_2)$$

Donde los predicados representan lo siguiente:

$\beta_1 (Y, X_1)$: La variable Y [sustituir por descripción de la variable Y] aumenta en promedio β_1 unidades [sustituir por el valor asociado a β_1] por cada cambio de una unidad en X_1 [sustituir por descripción de la variable X_1].

$\beta_2 (Y, X_2)$: La variable Y [sustituir por descripción de la variable Y] aumenta en promedio β_2 unidades [sustituir por el valor asociado a β_2] por cada cambio de una unidad en X_2 [sustituir por descripción de la variable X_2].

$\beta_0 (Y, X_1, X_2)$: Los valores de la variable Y [sustituir por descripción de la variable Y] difieren de la suma de las operaciones $\beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i}$

en β_0 unidades [sustituir por valor asociado a β_0] más un valor aleatorio.

A la descripción anterior le hace falta una serie de predicados de segundo orden del tipo Sα(β_1) : El valor asociado a β_1 difiere de 0 con una confianza de α. Otra posibilidad sería emplear algún tipo de lógica difusa para representar la credibilidad que se tiene en los resultados.

En el caso de un árbol de decisión (ver Figura 2), esta forma de traducir los resultados a un lenguaje formal es incluso trivial, pero nada parecido podría realizarse para, por ejemplo, un perceptrón multicapa. Aun si lográramos sustituir todas las variables por conceptos (lo cual no es posible respecto de la capa oculta) no podríamos encontrar operadores lógicos que mantuvieran la semántica del modelo.

Lo que pretendemos con esta discusión no es negar la utilidad de los modelos del aprendizaje profundo por el hecho de que no logren pasar el reto de la sistematicidad. De hecho, ya existen avances significativos en la búsqueda de dotar de interpretación a los resultados en condiciones de opacidad. Lo fundamental es comprender las ventajas y límites de estos modelos con respecto a los de tipo simbólico.

Si algo debemos aprender de las discusiones del siglo XX es que los seres humanos nos manejamos de manera tanto simbólica como subsimbólica, pero que carecemos de acceso consciente a los procesos subsimbólicos. Sin pretender asumir un compromiso con la tesis de un lenguaje del pensamiento o con negación de la cognición en seres no lingüísticos o prelingüísticos, es necesario reconocer que la posibilidad de dirigir nuestros pensamientos de una manera análoga a un lenguaje nos permite facilitar nuestros razonamientos. En esto tiene un papel fundamental el carácter composicional de los lenguajes y es este el tipo de consideraciones las que deberían guiar la investigación para dotar de interpretabilidad a los modelos de ciencia de datos.

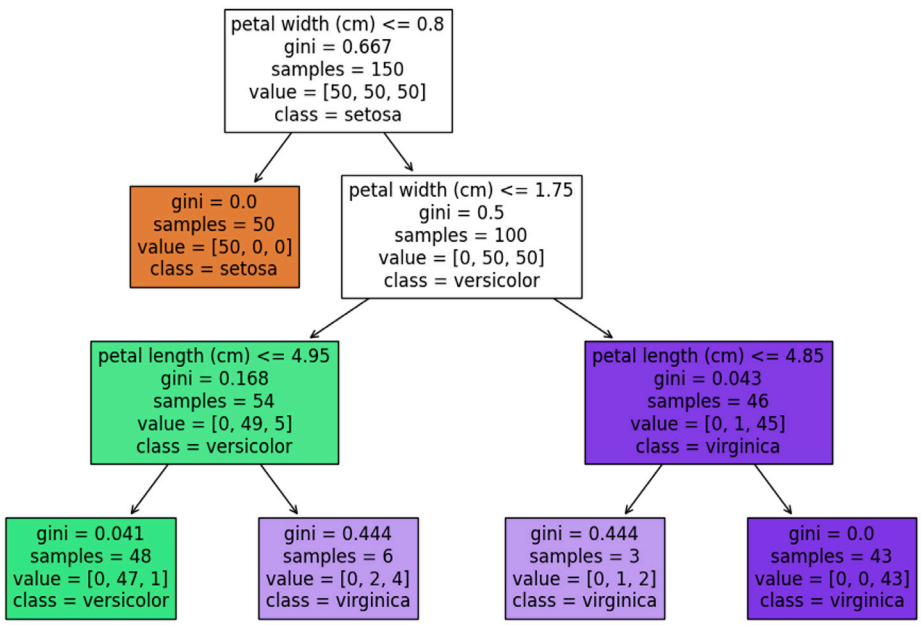

Figura 2. Árbol de decisión para problema de clasificación de un problema clásico de Fisher sobre la flor de iris. Las etiquetas hacen referencia a tres especies de iris (variable dependiente) y el ancho del sépalo y pétalo. El árbol puede interpretarse mediante reglas de producción if-then, de manera que cumple con las exigencias de Fodor y Pylyshyn (1988).

Notas

1. Quizás sería mejor llamar a esta posición estructuralismo ya que los desarrollos del laboratorio de Leipzig no dependían exclusivamente de la introspección. Sin embargo, el hecho de que el término estructuralismo se emplea en muchos contextos hace que este resulte ambiguo y confuso por lo que en adelante seguiré empleando «introspeccionismo».
2. Seguimos el criterio de Gutiérrez al traducir «intentionality» como «intensionalidad» (con «s»). Intensionalidad se refiere al «aboutness», es decir, la capacidad que tienen ciertos estados mentales para referirse o representar otras cosas, propiedades o estados de cosas (Pierre, 2023). No debe confundirse con «intensionality» que proviene del concepto semántico de «Intension» el cual se refiere a una propiedad o cualidad empleada en una definición, y que según el criterio de Gutiérrez debería traducirse como «intención».
3. Esto es evidente en los últimos trabajos de Gutiérrez, en particular en *El humanismo replanteado: genes y memes en la sociedad globalizada* ([2006] 2011), donde se contemplan diversos desarrollos sobre genética, evolución, artificial life, la teoría de los memes, etc. Otro de los hitos que aún no aparecen en estos textos es la creciente influencia de los estudios de neuroimagen en la Ciencia Cognitiva.

Referencias

Acree, Michael C. 2021. «The Fisher and Neyman-Pearson Theories of Statistical Inference». En: *The Myth of Statistical Inference*. Berlín: Springer, pp.281-334.

Bermudez, José Luis 2014. *Cognitive science: an introduction to the science of the mind*. Cambridge: Cambridge University Press.

Boden, Margaret. 1987. *Artificial Intelligence and Natural Man*. Cambridge: MIT Press.

Churchland, Paul 1984. *Matter and consciousness*. Cambridge: MIT Press.

Churchland, Patricia Smith. 1986. *Neurophilosophy*. Cambridge: MIT Press.

Collins, Allan M., & Quillian, M. Ross. 1969. Retrieval time from semantic memory. *Journal of Verbal Learning & Verbal Behavior 8*: 240-247.

Fodor, Jerry. A. 1975. *The language of thought*. Cambridge: Harvard University Press.

Fodor, Jerry A. y Pylyshyn, Zenon W. 1988. «Connectionism and cognitive architecture: A critical análisis». *Cognition* 28: 3-71.

Gutiérrez, Claudio. 1985. «Un nuevo paradigma para las ciencias del conocimiento». *Revista de Filosofía de la Universidad de Costa Rica 23*: 131-135.

Gutiérrez, Claudio. 1993. *Epistemología e informática*. San José: EUNED.

Gutiérrez. Claudio. 1987. «Perspectivas de las máquinas inteligentes o la psicología de las computadoras». *Revista de Filosofía de la Universidad de Costa Rica 25*: 109-115.

Gutiérrez, Claudio. 2006. *El humanismo replanteado: genes y memes en la sociedad globalizada*. San José: EUNED.

Lakatos, Imre. 1978. *The Methodology of Scientific Research Programmes: Volume 1: Philosophical papers.* Cambridge: Cambridge University Press.

Hornik, Kurt, Stinchcombe, Maxwell y White, Halbert 1989. «Multilayer Feedforward Networks are Universal Approximators». *Neural Networks 2*: 359-366.

Jacob, Pierre, «Intentionality». *The Stanford Encyclopedia of Philosophy* (Spring 2023 Edition). Editado por Edward N. Zalta & Uri Nodelman. URL = <https://plato.stanford.edu/archives/spr2023/entries/intentionality/>.

Kuhn, Thomas S. 1962. *The structure of scientific revolutions*. Chicago: University of Chicago Press.

Miller, George A. «The cognitive revolution: a historical perspective». *Trends in Cognitive Science 7*: 141-144.

Minsky, Martin L. y Papert, Seymour A. 1969. *Perceptrons*. Cambridge: MIT Press.

Marr, David. 1982. *Vision*. San Francisco: W. H. Freeman.

Newell, Allen. y Simon, Herbert. A. 1976. «Computer science as empirical inquiry: Symbols and search». *Communications of the Association for Computing Machinery 19*: 113–26.

Place, Ullin T. 1956. Is Consciousness a Brain Process?, *British Journal of Psychology 47*: 344–50.

Rumelhart, Daniel.E., Hinton, Garrett, E. y Williams, Richard J. 1986. Learning representations by back-propagating errors. *Nature 23*; 533-536.

Searle, John. R. 1980. Minds, brains, and programs. *Behavioral and Brain Sciences 3*, 417–57.

Stacewicz, Paweł y Greif, Hajo. 2021. «Concepts as decision functions. The issue of epistemic opacity of conceptual representations in artificial computing systems». *Procedia Computer Science 192*: 4120–4127.

Symons, John y Calvo, Paco. 2014. *The Architecture of Cognition: Rethinking Fodor and Pylyshyn's Systematicity Challenge*. Cambridge: MIT Press.

Smolensky, Paul. 1988. «On the proper treatment of connectionism». *Behavioral and Brain Sciences* 11: 1 – 23.

Tolman, Edward. C. 1932. *Purposive behavior in animals and men*. Nueva York: Appleton-Century-Crofts.

Turing, Allan. 1936. «On computable numbers, with an application to Entscheidungs problem». *Proceedings of the London Mathematical Society 42*: 230–65 y 43, 544–46.

Turing, Allan. 1950. «Computing Machinery and Intelligence». *Mind 59*: 433-460.

Wundt, Wilhelm y Pintner, Rudolf. 2016. *An Introduction to Psychology*. Leopold Classic Library.

Mauricio Molina-Delgado: (mauricio.molina@ucr.ac.cr) es Licenciado en Estadística y Máster en Ciencias Cognoscitivas por la Universidad de Costa Rica y Doctor en Psicología por la Universidad Aristotélica de Salónica, Grecia. Cuenta con estudios en Ciencia de Datos en la Universidad de Boulder, Colorado. Entre el 2020 y el 2023 fue director de la Escuela de Filosofía de la Universidad de Costa Rica y actualmente es investigador del Instituto de Investigaciones Filosóficas, profesor de la Escuela de Filosofía y de la Escuela de Estadística de la Universidad de Costa Rica. Es además escritor y en el 2016 recibió el Premio Nacional Aquileo Echeverría en Poesía.

Eugenia Gallardo-Allen: (eugenia.gallardo@ucr.ac.cr) es docente de la Escuela de Estadística de la Universidad de Costa Rica. Realizó un doctorado en Gobierno y Políticas Públicas, además de una maestría académica y un bachillerato en Estadística, todos por la Universidad de Costa Rica. Entre los intereses de estudios se encuentran políticas educativas, aplicación de técnicas de machine learning e investigación educativa relacionada con la enseñanza de la Estadística.

Recibido: 30 de octubre, 2023.
Aprobado: 6 de noviembre, 2023.