MODELOS CONEXIONISTAS: CONSECUENCIAS PARA LA CIENCIA COGNITIVA

RONALD DE SOUSA

Dos paradigmas dominan la investigación actual sobre la Ciencia Cognitiva y la Inteligencia Artificial: los modelos "clásicos" que se basan en las computadoras de Neumann, y los modelos "conexionistas" que han sido motivo de mucha atención en los últimos tiempos. Después de describir brevemente algunas de las características de estos modelos, quiero sugerir una perspectiva desde la cual sea posible visualizar las contribuciones respectivas de dichos modelos a la ciencia cognitiva, así como una explicación, de cierto tipo, sobre las áreas en las que tienen éxito y aquellas en las que fracasan.

Los dos paradigmas rivales ilustran una especie de principio de "Primero que entra último que sale" según el cual las aptitudes mentales que mejor comprendemos son aquellas que hemos inventado nosotros mismos, recientemente y deliberadamente. En cambio, aquellas que *realizamos* mejor son las que la naturaleza necesitó unos millones de años para perfeccionar; también son aquellas que son las más difíciles de comprender. Concluiré formulando algunas especulaciones sobre las implicancias de estas consideraciones acerca de la comprensión de la *comprensión*. Tendemos a asumir que estamos en condiciones de comprender en su totalidad aquello que construimos. Pero el conexionismo ofrece la posibilidad de que quizá podamos comprender la mente sólo de modo atenuado, y carecer de toda esperanza de predecir con precisiones los hechos mentales.

Comenzaré con un brevísimo esbozo de estos dos enfoques, clásico y conexionista. Las computadoras clásicas son demasiado familiares como para explayarse demasiado en ellas. Están programadas para realizar, típicamente en secuencias, tareas que han sido previamente analizadas y descompuestas en las tareas elementales que las componen. Estas tareas consisten esencialmente en procesar información, y la información con la que se ha trabajado se almacena en espacios identificables en la memoria de la computadora.

Las máquinas conexionistas (o "procesadores paralelos distribuidos") realizan tareas que no han sido programadas o analizadas previamente, pero que una red *ha aprendido* a realizar. La información no se al-

macena en espacios identificables, sino que está distribuida en el conjunto de la red.

Para ser un poco más específico, una máquina conexionista es una red de *nodos*, típicamente organizados en tres o más *capas*. Una capa se denomina capa del *input*, otra es la capa del *output*. Cada uno de los nodos en una capa se conecta con todos los nodos de la capa o capas adyacentes. Para cada nodo hay un *umbral* particular de activación; dado el umbral, la actividad del nodo se determina por la suma de nodos activos con los cuales está conectado y por la fuerza de estas conexiones. La *fuerza de conexión* entre dos nodos (tanto la cantidad o la probabilidad de activación transmitida) puede variar entre 0 y 1. Las fuerzas de conexión iniciales pueden atribuirse al azar, pero éstas son modificadas por ciertos artificios sistemáticos de modo tal que pueden modificar la función input/output. La conducta de la red puede pues modificarse a través del *aprendizaje*, y puede decirse que la fuerza de las conexiones en la red en su conjunto es lo que almacena información. La programación se limita al establecimiento de artificios adecuados para la modificación de las fuerzas de conexión.

La visión clásica se sustenta en lo que podríamos llamar consideraciones *cartesianas*: para comprender las aptitudes propiamente humanas del lenguaje y la razón, debemos concebir la actividad mental en términos de manejo de elementos que tengan sentido de acuerdo con reglas sintácticas.

Irónicamente, para este punto de vista cartesiano, lo que estos modelos no han logrado explicar muy bien son aquellas aptitudes humanas que comparten con los animales —a los que Descartes mismo consideraba simples máquinas—. Lo que Descartes consideraba como rasgos esencialmente humanos —como la razón— se han logrado —por lo menos en parte— mecanizar, mientras que resulta muy difícil mecanizar aquellos que el consideraba obviamente mecánicos. La razón es que el segundo tipo de tareas, las que implican posibilidad de *aprendizaje*, aun cuando puedan *analizarse* adecuadamente, el análisis no siempre proporciona un método de realización. Este hecho no se basa en el supuesto de que estas tareas sean intrínsecamente *inanalizables*: pero aun cuando se puedan establecer reglas explícitas, puede resultar imposible para nosotros aplicarlas. Voy a darles por ejemplo el caso de cómo aprender a andar en bicicleta:

> Dé vuelta a los manubrios de modo que la curva de su trayectoria sea proporcional al ángulo de su equilibrio dividido por el cuadrado de su velocidad ([3], p. 95)

Es obvio que aun cuando esta regla sea correcta no serviría para aprender a andar en bicicleta.

Este es el tipo de hecho explorado por los conexionistas. Rechazan la concepción clásica según la cual una acción deba basarse en un análisis o programación previa; además sus modelos parecen capaces de simular algunas de nuestras "aptitudes animales" —tales como la capacidad de ver, caminar, andar en bicicleta— mejor que cualquier programa clásico.

Muchos investigadores de cualquiera de estas dos posiciones (aunque no todos) opinan que los dos enfoque son incompatibles. Algunos abogados de los modelos clásicos conceden que los modelos conexionistas pueden contribuir a la *implementación* de modelos clásicos, mientras insisten en que aun el "conocimiento infraverbal" se describe mejor en términos de Inteligencia Artificial clásica. Por otra partes, algunos conexionistas piensan en el modelo clásico como si describiera epifenómenos, basados en procesos más fundamentales, de naturaleza conexionista. Sin embargo, lógicamente, la aplicabilidad de algún tipo de modelo no parece excluir al otro. La intransigencia de ambas partes debe explicarse en términos de la importancia que tienen los temas en discusión.

Un tema de controversia fundamental es la tesis según la cual la actividad mental se comprende textualmente como procesamiento de información basado en *representaciones*. Esta es una tesis importante de la inteligencia artificial clásica, en la medida en que no intenta *simular* procesos mentales sino *reproducirlos*. Para comprender esto comparemos una computadora clásica y una piedra que cae. La computadora puede computar la trayectoria de una piedra que cae, recurriendo al cálculo pero la piedra en sí no computa nada. O bien pensar en "una computadora de pompa de jabón" en la que una película de jabón se configura en una forma que efectivamente resuelve un problema de geometría. Podemos preguntarnos si una pompa de jabón puede realmente procesar información en mayor medida de lo que podría hacerlo una piedra que cae.

Robert Cummins, en su libro *Psychological Explanation*, ha aclarado esta distinción en los siguientes términos. Necesitamos distinguir, sugiere, entre un *análisis* de un proceso y una *causa* de que el proceso esté ocurriendo. El análisis revelará lo que está ocurriendo. A veces, puede además hacer una contribución causal al proceso. Pasa así en la computadora programada con reglas de cálculo. Por contraste, la mera *realización* de un proceso (como la computadora de pompa de jabón que realiza una solución geométrica) no necesita contener ninguna descripción analítica de aquel proceso ([1] 46)

En estos términos podemos explicar la distinción entre los dos enfoques en cuestión. En forma somera, la visión clásica representa la actividad cerebral textualmente como un manipuleo de la información, instanciado en *un sistema de símbolos físicos*. Las representaciones mentales complejas están construidas con partes simples de acuerdo con reglas sintácticas, y el contenido semántico de una representación molecular es una función de los contenidos semánticos de sus partes y de su estructura sintáctica. Dentro de este esquema, dicen Fodor y Pylyshyn, "las contrapartes físicas de los símbolos y sus propiedades estructurales, *son causa* de la conducta del sistema" ([2] 11-12). (No me ocupo aquí de la cuestión polémica de cómo un sistema físico *puede adquirir sentido* y de este modo convertirse en un sistema verdadero de símbolos.)

Por contraste, el modelo conexionista hace su camino de un modo puramente físico. Parece no haber posibilidad de establecer un esquema estricto entre los pasos y elementos de un argumento lógico y cualquier elemento identificable en lo hechos físicos que realizan ese argumento.

Desde el punto de vista clásico, entonces, el conexionismo se enfrenta a un dilema: o bien transforma el cerebro en algo demasiado parecido a la computadora "pompa de jabón", que no puede lograr actividad mental, o bien hace una contribución informativa genuina y entonces no hace otra cosa que *implementar* el modelo clásico.

Sin embargo tal dilema exclusivo debería ser rechazado. Paul Smolensky argumentó que existe un nivel adicional, intermedio entre el puramente físico y el simbólico, que necesita ser comprendido en términos de "paradigma *subsimbólico*". Nuestras aptitudes subverbales, o lo que Smolensky denomina "el procesador intuitivo", hay que comprenderlos en términos de este paradigma subsimbólico. Además, Smolensky piensa que el procesador digital "lógico" modelado por la Inteligencia Artificial clásica es sólo un nivel macro de aproximación que se explica, en parte, en términos de modelo conexionistas.[6]

Quiero sugerir que las glorias y miserias de ambos modelos pueden considerarse en términos del principio de FILO: *First In, Last Out*: entra primero, sale último. Lo que comprendemos mejor sobre nuestras mentes es lo que nuestros cerebros inventaron al final; y es muy posible que comprendamos en última instancia los secretos relativos a aquellas aptitudes que nuestros cerebros desarrollaron primero.

La atracción del paradigma simbólico deriva del poder de la lógica y del razonamiento secuencial: una pieza de tecnología intelectual más fa-

bulosa que la rueda, y más reciente. Porque es esencialmente la elaboración de un invento de Aristóteles: a saber, de la *representación lingüística formal* de *modelos abstractos de inferencia.* (Y esto a su vez se basa en el descubrimiento de Platón referente a los objetos abstractos que pueden ser modelados en el mundo real, pero nunca identificados con ningún objeto del mundo real.) El concepto moderno de *funcionalismo* es en realidad el concepto de Platón según el cual ciertas realidades pueden ser modeladas indistintamente con muchos soportes materiales diferentes sin ser identificadas con ninguno de ellos.

Bastante naturalmente, pues, la ciencia de la computación ha abordado en *primer lugar* a este último, muy reciente logro (imposible de imaginar sin el lenguaje). Esto fue así no sólo porque era tan poderoso, sino también porque era, en un sentido, tan *fácil.* Es fácil, naturalmente, porque lo hemos inventado (o descubierto los principios básicos que lo gobiernan). Sin embargo, éstos no son procedimientos en los que somos muy prácticos: exige justamente mucha práctica ejercer una inferencia lógica estrictamente deductiva, hasta cuando cometemos errores.

Por otra parte destrezas como *ver* y *moverse* —según Descartes aptitudes puramente animales— son cosas que nos resulta extremadamente fácil aprender. Las hemos estado practicando durante millones de años. Nos parece muy fácil ver, pero enormemente más difícil hallar los procedimientos que utilizamos para ver.

El principio de FILO puede parecer sospechosamente mágico. *Stacking,* que implica el principio de FILO, es utilizado ampliamente en la ciencia de la computación, pero ¿qué razón existe para pensar que se aplica en este caso? En psicoanálisis suele recurrirse a algo parecido al principio de *stacking*: se da por sentado que los supuestos planteados en una fase anterior de la infancia serán mucho más difíciles de reparar y modificar que aquellas capas instaladas en fases posteriores. Pero desconozco mecanismos claros para este efecto de *stacking* y aun cuando se aplica a escala de ontogenia individual esto no constituiría una razón para que el principio se sustente en la escala evolutiva.

Sin embargo, existen algunas consideraciones en favor de este principio de FILO. La más importante es que nuestra aptitud para adquirir diferentes destrezas resultó de la evolución biológica: no podemos por lo tanto esperar que sean tan especialmente simples como para resolverlas mediante "reverse engineering". La selección natural es la última *anarquía de programadores*; cualquier programador sabe que un programa con el que han lidiado unas cuantas docenas de programadores —sin hablar

de unos cuantos millones— puede volverse desesperadamente opaco. Las consignas programadas y unidas por evolución pueden a veces ser barrocas al extremo.

Más aún, la diferencia entre dos modelos puede también estar relacionada con la importancia para el lenguaje de la *digitalidad*.

En verdad, la digitalidad es también un descubrimiento de Platón. Consiste esencialmente en definir la *semejanza* como relación de tres términos en vez de dos. Permítaseme explicar: para copiar una letra o un número, no imitamos estrictamente el original sino que simplemente reproducimos otra letra o número *de este tipo*. En lugar de reproducir el original mismo, podemos copiar una especie de forma platónica de la que la forma "original" es meramente una especie de copia. Eso es la copia digital.

Podemos preguntarnos bajo cuales circunstancias sería aquello particularmente útil. La respuesta es: Cuando uno desea que una taxonomía permanezca estable a través de múltiples réplicas. Cualquier información de *tipo público* tendrá necesidad de sobrevivir muchas repeticiones: si no es digital, se degradará inexorablemente, y tomará una distancia arbitraria de su origen.

Esto, sugiero, puede radicar la insistencia de Fodor y Pylyshyn, según quienes debe existir un lenguaje "composicional" —uno compuesto sistemáticamente por partes—. Pero, si esto es cierto, el argumento que ellos sostienen no viene del todo al caso. Porque para lo que se necesitan los procesos digitales es para procesos *culturales*. La información procesada "internamente" puede no formular las mismas exigencias ([6] 4). Esto deja abierta la posibilidad, por la que abogan los conexionistas, de que exista un modo diferente de procesar la información. Aquel "procesador intuitivo" será, naturalmente, sistemático, pero no deberá ser necesariamente digital. Por cierto, las leyes de causalidad que pueden descubrirse en la naturaleza son sistemáticas sin ser digitales.

Entre las razones que existen para pensar que este procesador intuitivo trabaja en la línea de una red conexionista está el hecho de que tales redes han podido, en algunos casos, desarrollar destrezas y aprender tareas de categorización mejor que cualquier computadora "programada"[5]. De modo que aun si rechazamos la hegemonización de este modelo a áreas de Inteligencia Artificial, es posible argumentar que los modelos conexionistas están para quedarse.

He sugerido también que los modelos conexionistas pueden describir un nivel genuinamente diferente de procesamiento de la información, si-

tuado en cualquier punto entre el modelo clásico de computadora y el nivel de realización puramente física en que habló Cummins. Sin embargo, esta perspectiva puede implicar el pago de un precio sumamente elevado en términos de la comprensión de la actividad mental a la que podemos aspirar.

Lord Kelvin escribió una vez:

> Jamás me siento satisfecho hasta lograr un modelo mecánico de algo. Si logro crear un modelo mecánico, puedo comprenderlo. En tanto no logre establecer un modelo mecánico completo no puedo comprender de qué se trata ([3] 24).

Pero en la ciencia cognitiva, puede resultar que *aun cuando podamos construir un cerebro no seamos capaces de comprenderlo.*

La observación de Lord Kelvin conjura una imagen de un mundo del *hágalo usted mismo*, en el que todo puede construirse en el propio laboratorio de casa. Tal vez, en un sentido relevante, tal construcción es rara vez posible ya. Pero el *tipo* de construcción que fuera concebido por la Inteligencia Artificial era ciertamente eso: uno podía comprender todo, precisamente porque se podían construir modelos perfectamente transparentes, hechos de elementos platónicos y lógicos. Esto no es así en el programa de investigación conexionista, donde los constructores se sienten orgullosos de *no haber programado* a sus computadoras.

En un sentido, y por contraste con Kelvin, se ufanan de no comprender las máquinas mismas que han construido. De modo que construir algo parece no ser condición suficiente para la comprensión. Lo mejor que podemos hacer es comprenderlo "en principio", de la misma manera que podemos comprender ciertos hechos físicos en términos de gravedad, viento, resistencia, fricción, etc., sin ser realmente capaces de leer todos los parámetros relevantes para establecer una predicción útil.

En alguna medida, esta disminución de las expectativas puede atribuirse a la naturaleza estadística de la actividad neurológica subyacente al conocimiento. Compare la relación entre la termodinámica y la estadística: sabemos qué tipo de hechos estadísticos subyacen a los fenómenos de volumen y presión en un gas. Pero no existe modo en que podamos saber jamás en detalle lo que ocurre en un caso particular. Pero entonces no nos importa demasiado. Porque la reducción ya está completa: la información estadística nos dirá lo que queramos saber sobre presión y volumen.

En la ciencia cognitiva la situación quizá es peor. El argumento de Smolensky es que existe un nivel genuino de procesamiento de la infor-

mación que puede comprenderse en términos de "paradigma subsimbólico" que se relaciona al "nivel simbólico" de la Inteligencia Artificial clásica de un modo que no es sistemático. Si esto es correcto, entonces tenemos razón para aceptar las siguientes tesis:

1. Existen representaciones mentales en el nivel subsimbólico.

2. Estas representaciones resultan *inaccesibles* para el nivel consciente y explícito de representaciones en términos en los que pensamos y analizamos las tareas.

3. Están *relacionadas con imprecisión* a aquellas representaciones de nivel simbólico.

4. Sin embargo, dados los requerimientos de nuestros código lingüístico, necesitaremos para siempre *hablar* como si el nivel simbólico de representación fuera toda la representación con la que nos tenemos que preocupar para comprender la mente.

Si aceptamos estas cuatro tesis, entonces tendremos una razón fundamental para creer que en un sentido muy poderoso nosotros no podemos comprendernos a nosotros mismos.

UNIVERSITY OF TORONTO, CANADA

REFERENCIAS

[1] Cummins, Robert. *Psychological Explanation* (MIT: 1983).

[2] Fodor, J. and Pylyshyn, Z. "Connectionism and Cognitive Architecture", in *Connections and Symbols,* ed. Steven Pinker and Jacques Mehler (Reprint of *Cognition*, vol. 28, 1988)

[3] Johnson-Laird, P. *The Computer and the Mind: an Introduction to Cognitive Science* (Harvard: 1988).

[4] McClelland, D.E., Rumelhart, James and the PDP Research Group. *Parallel Distributed Processing: Explorations in the Microstructures of Cognition.* 2 vols. (MIT: 1986)

[5] Sejnowski, T.J. and Rosenberg, C.R. "Parallel networks that learn to pronounce English". *Complex Systems* 1:145-168 (1987).

[6] Smolensky, Paul. "The Proper Treatment of Connectionism". *Behavioral and Brain Sciences,* 8:1-23.

ABSTRACT

Two classes of models dominate current Cognitive Science and Artificial Intelligence research: those I shall call "classical", based on von Neumann machines, and the "connectionist" models which have recently attracted much attention. After briefly describing some features of these models, I suggest a perspective in which we might view their respective contributions to cognitive science, and an explanation, of sorts, for their different areas of success and failure. The two competing paradigms illustrate a kidn of "First In Last Out" principle: those of our own mental skills that we understand best are those which we have ourselves consciously and recently invented. Those that we *perform* best, by contrast, are those it took Nature a few million years to perfect; those are also the most difficult to understand. I conclude with some sopeculations about the implications of these considerations for our understanding of *understanding*. We tend to assume that anything we can build, we can fully understand. But connectonism offers the prospect that perhaps we might understand the mind only in some attenuated sense; we would lack any hope of precise prediction.