De causalidad mental y conexionismo

1. PARADOJAS Y DILEMAS

Consideremos la siguiente paradoja (Fodor [1989], Jackson y Pettit [1988] [1993], Dretske [1988], Block [1991], Lepore y Loewer [1987], Lewis [1986], Segal y Sober [1991]):

1. El contenido intencional de un pensamiento (o de cualquier estado intencional) es causalmente relevante con respecto a sus efectos conductuales;
2. el contenido intencional no es más que el *significado* de las representaciones internas. Pero,
3. los procesadores internos son sensibles únicamente a las estructuras sintácticas de las representaciones internas, no a los singnificados.

Parece que si queremos defender la idea, absolutamente plausible desde un punto de vista intuitivo, de que los estados mentales / intencionales son causalmente responsables de *outputs* conductuales, y lo queremos hacer sobre la base fisicalista propia de toda metodología científica, tenemos que renunciar al convencimiento de que son tales estados intencionales *qua* intencionales —*i. e.* en tanto que poseedores de un significado determinado— los causalmente responsables de nuestra conducta.

La línea que conduce al epifenomenalismo de lo mental es clara: 1. los poderes causales de cualquier evento están enteramente determinados por sus propiedades físicas; 2. aunque las propiedades intencionales sobrevienen sobre las propiedades físicas, no son idénticas a estas últimas; luego, 3. las propiedades

Anales del Seminario de Metafísica, N.º 27-1993. Editorial Complutense. Madrid

intencionales, en tanto intencionales, no son causalmente responsables de la conducta, porque no intervienen en los poderes causales de los estados que las poseen, *i. e.*, las propiedades intencionales son epifenoménicas.

Consideremos ahora una posición diferente en carácter, aunque paralela a la anterior. Existe un importante debate en el ámbito de la Inteligencia Artificial sobre si la clase de mecanismos a los que nosotros pertenecemos y a la que tendría que remitirse el proyecto de modelado computacional de los procesos cognitivos está mejor representada por los modelos clásicos o por los denominados modelos conexionistas (McClelland, Rumelhart, et al. [1986], Smolensky [1987] [1988], Fodor y Pylyshyn [1988], Pinker y Prince [1988], Clark [1989], Ramsey, Stich y Rumelhart [1991], Clark y Karmiloff-Smith [1993]). Frente a los modelos clásicos de procesamiento serial en los que la información está codificada en base a reglas de carácter lingüístico, los modelos conexionistas o de procesamiento distribuido de información en paralelo (PDP) contienen toda la información necesaria para explicar una cierta relación input-output sin referencia a ninguna categoría semántica. Las relaciones causales entre las unidades que componen el sistema describen suficientemente cómo la información es procesada por la red. Pero como esas unidades no tienen una interpretación semántica directa, todas las computaciones pueden ser explicadas sin referencia al *contenido* de la información procesada. Si en el debate mencionado optamos por el paradigma conexionista, la idea intuitiva de que las actitudes proposicionales —deseos, creencias y, en definitiva, los estados mentales con un cierto contenido semántico— funcionan como causas de nuestra conducta parece ser una idea falsa. Las creencias o deseos, en tanto estados funcionalmente discretos, no son ni individualizables como estados de activación del sisteman, ni individualizables en relación con sus efectos, porque la información es codificada por la red en representaciones distribuidas y superpuestas (Ramsey, Stich y Garon [1991])[1].

La polémica entre los paradigmas clásico y conexionista en inteligencia artificial no hace pues sino reflejar el estado filosófico de discusión en ciencia cognitiva en lo relativo a la relevancia explicativa de las propiedades semánticas. En este sentido, el paralelismo entre las dos líneas de argumentación mencionadas resulta evidente. Pero, mientras la primera línea nos conduce a una paradoja, la

1. Una representación *distribuida* es una representación constituida por un amplio conjunto de micropropiedades que tienen un carácter sub-simbólico, *i. e.* que no son semánticamente interpretables en sí mismas. Se habla de representación *superpuesta* cuando la misma unidad o unidades que contribuye(n) a codificar la información presente en esa representación, contribuye a codificar información presente en muchas otras representaciones diferentes.

segunda nos conduce a un dilema: Si las hipótesis conexionistas son correctas, o bien claudicamos ante la posición eliminativista —lo que en términos filosóficos es equivalente a la posición epifenomenalista descrita más arriba— o defendemos la eficacia causal de los estados mentales en tanto que estados con contenido semántico mostrando que, después de todo, las redes conexionistas nunca podrán ser realmente buenos modelos cognitivos (Davies [1991]).

Desde mi punto de vista, sin embargo, ambas líneas de argumentación necesitan ser revisadas. Mi objetivo en este artículo es encontrar una fórmula de legitimación filosófica en favor de la eficacia causal del contenido de los estados mentales que eluda las objeciones epifenomenalistas y que no requiera la utilización de ítems simbólicos en la construcción de los modelos computacionales de tales estados mentales. En definitiva, el objetivo es encontrar un punto de reconciliación entre el modelo filosófico representacional y el modelo computacional conexionista (Clark [1989]).

2. RELEVANCIA *VERSUS* EFICACIA

¿Cómo compaginar el hecho de que las propiedades de nivel semántico parecen ser explicativamente **relevantes**, pero no causalmente **eficientes** —ya que el poder causal de los estados que las poseen reside en los estados microfísicos que las realizan?

Este problema supone una discusión de algunas de las posiciones más representativas de la Ciencia Cognitiva actual, básicamente de aquellas que se engloban bajo los rótulos de Teoría Representacional (Fodor [1975] [1987]) y Teoría Sintáctica de la Mente (Stich [1983], Churchland [1986] [1989]). La primera se caracteriza fundamentalmente por conceptualizar los estados mentales en términos de sus relaciones con alguna suerte de entidades representacionales —oraciones cifradas en algún código mental o lenguaje del pensamiento— que se conciben, por tanto, como semánticamente interpretables y como causalmente eficaces en virtud de su contenido semántico. Los modelos clásicos en Inteligencia Artificial representan el apoyo empírico a las posiciones agrupadas bajo este rótulo.

La idea básica de la Teoría Sintáctica es, por otro lado, que los estados cognitivos pueden proyectarse sistemáticamente sobre objetos sintácticos abstractos, de tal manera que las cadenas causales entre estímulos y eventos conductuales pueden describirse exclusivamente en términos de las propiedades y relaciones sintácticas de estos objetos sin necesidad de apelar a ningún contenido semántico. No debe resultar extraño, pues, que sean en este caso los

modelos conexionistas los que aporten el soporte empírico apropiado para defender las tesis típicas de la Teoría Sintáctica.

Una revisión exhaustiva de esta paisaje filosófico-computacional, incluso una revisión que se limite al tema específico de la causación mental, queda fuera de los límites de este artículo. Mi propuesta es mucho más modesta. Se reduce al intento de mostrar la posibilidad de leyes intencionales con la ayuda de herramientas conceptuales pertenecientes al ámbito computacional. Después de todo —y aunque esta no sea una tesis ajena a cierta controversia—, la cuestión sobre si una propiedad P, semántica o no, es causalmente responsable de cierta conducta se reduce a la cuestión de si existen leyes causales acerca de P[2]. Así, si se puede mostrar que existen leyes causales acerca de propiedades semánticas, y se puede hacer sin necesidad de recurrir a ningún código mental o lenguaje del pensamiento a lo Fodor, el doble objetivo propuesto habrá sido alcanzado.

De acuerdo con la tesis anterior, podemos decir que P es una propiedad causalmente eficaz si los individuos que la instancian pueden ser subsumidos bajo leyes causales. Para poder afirmar correctamente que un evento **c** ha causado un evento **e** deben existir propiedades de tipo **F** y **G** tal que **c** instancia **F** y **e** instancia **G** y «las instanciaciones de **F** son suficientes para las instanciaciones de **G**» es una ley causal (Fodor [1989]). El compromiso ontológico subyacente a esta tesis es claro. Sólo eventos o hechos individuales pueden ser causas. Pero, a la vez, el carácter necesario de las regularidades expresadas por una ley causal depende de que tales regularidades sean establecidas no entre eventos particulares, sino entre *tipos* de eventos. Ahora bien, puesto que los eventos particulares pueden ser designados de formas muy diferentes, algunas de las cuales no recogen en absoluto ninguna de las propiedades que los convierten en causa o efecto de otros, los criterios para agrupar eventos individuales en eventos del mismo tipo, eventos del tipo que puede aparecer en una **ley** causal, han de concentrarse sólo en aquellas propiedades que pueden mostrarse como causalmente eficaces. Por ejemplo, aunque puede ser verdadero que la cena del domingo causó mi dolor de estómago, la ley causal de la que depende tal verdad no establece ninguna relación entre eventos del tipo *cenar en domingo* y *dolor de estómago*, sino más bien entre eventos de un tipo físico determinado, tales como una cierta composición de la carne y una alteración de los jugos gástricos.

2. La influencia de Davidson en la *forma* de este planteamiento resulta evidente. Después de todo, su monismo anómalo, *i. e.*, la posición de que los eventos mentales entran en relaciones causales sólo bajo su descripción física porque ésta es la única descripción en la que quedan subsumidos bajo leyes, supone asumir la presente tesis.

En definitiva, lo que convierte una ley causal en una ley, y no en una mera regularidad con suficiente soporte estadístico, es la existencia de un mecanismo o estructura micro-física común a los diferentes macro-tipos de eventos que aparecen subsumidos por tal ley. O, dicho en otros términos, para que un macro-tipo de eventos pueda ser considerado causalmente eficaz debe sobrevenir en eventos microfísicos —quizá diferentes en diferentes ocasiones— de tal manera que los poderes causales de aquéllos se expliquen a través de los poderes causales de éstos[3]. A la hora de individualizar leyes causales parece pues que, hablando estrictamente, sólo las leyes microfísicas lo son. El resto —todas las leyes de las denominadas ciencias especiales, como las Ciencias Sociales, o las *leyes* intencionales de la psicología, basadas en el contenido semántico de los estados mentales— son sólo formulaciones aproximadas que incluyen inevitablemente cláusulas *ceteris paribus*.

La distinción que Jackson y Pettit (Jackson y Pettit [1993]) establecen entre *eficacia* causal y *relevancia* causal tiene sus raíces en el mismo tipo de consideraciones. La única diferencia es que lo que ellos denominan relevancia *causal* es, desde mi punto de vista, un término poco afortunado para denominar lo que aquí yo he llamado relevancia *explicativa*. En efecto, para Jackson y Pettit, tanto las historias causales a nivel intencional, como las leyes causales con las que trabajan las ciencias especiales, sólo son causalmente relevantes —o, en mi terminología, explicativamente relevantes—, mientras que la eficacia causal propiamente dicha reside únicamente en el nivel más básico de la (micro)física. Su argumentación es similar a la expuesta previamente: Sólo eventos individuales desde el punto de vista microfísico pueden ser causas. Ahora bien, las historias causales de nivel superior al físico no se desarrollan en términos de eventos singulares, sino que generalizan sobre conjunciones y disyunciones de eventos de nivel inferior. Por tanto, las leyes causales de nivel superior al físico sólo pueden ser verdaderas en virtud de la eficacia causal de los eventos microfísicos a los que se refieren en última instancia los términos del vocabulario macrofísico o intencional que aparece en tales leyes.

---

3. La noción de sobreveniencia fue introducida originalmente por G. E. Moore en su caracterización de las relaciones existentes entre propiedades evaluativas y descriptivas. Que las propiedades evaluativas sobrevienen sobre propiedades descriptivas quiere decir que no puede haber una diferencia en las primeras sin que la haya también en las segundas. En este sentido la noción de sobreveniencia es una variación de la tesis general de que los hechos físicos determinan todos los hechos. Y en este sentido general es utilizada aquí. En la sección cuatro se ofrece un tratamiento más detallado en los términos de quien ha hecho de esta noción una de las más importantes de la filosofía de la mente contemporánea: Jaegwon Kim.

La argumentación, aunque persuasiva, dirige desgraciadamente sus beneficios justo a la cuenta epifenomenalista que pretendíamos liquidar. Sin embargo, aunque se le pueda conceder al científico realista que nada existe aparte de las entidades descritas por la física, ello no significa que debamos devaluar el status causal de las leyes que no incorporen exclusivamente tales entidades. Ello supondría malinterpretar el carácter mismo del método y la explicación científicas. La sección siguiente representa el intento de resistir este tipo de argumentación a través de la defensa de una tesis anti-reduccionista.

3. LEYES ESTRICTAS Y LEYES *CETERIS PARIBUS*

No creo estar sola en la creencia de que el programa reduccionista ha probado ser un programa equivocado. Una buena exposición de esta línea anti-reduccionista puede encontrarse en *The Scientific Image* de van Fraasen (van Fraasen [1980]). ¿Por qué deberían las propiedades semánticas de nuestros estados mentales aparecer en el nivel físico? Nadie espera encontrarlas allí. Miles de objetos de existencia incuestionable —mesas, huracanes, montañas...— no existen en el vocabulario (micro)físico y no por ello quedan desprovistos de sus poderes causales. Las leyes que permiten predecir la formación de un huracán en el Pacífico o que permiten explicar las consecuencias de tal formación, una vez reducidas a un vocabulario puramente físico, no mencionan ninguna entidad que podamos seguir denominando huracán. Y, sin embargo, ciencias como la geología, la meteorología y otras igualmente respetables basan sus explicaciones en la existencia de leyes causales que conectan fenómenos descritos en el vocabulario de esas ciencias. Especialmente relevante es el caso de la biología o la química. No es que los biólogos o los químicos nieguen que existe algo más aparte de los procesos físicos subyacentes a los procesos biológicos o químicos. Es simplemente que las explicaciones en términos de acidez o alcalinidad, o de variación y selección, son buenas explicaciones en virtud de la existencia de leyes que relacionan propiedades cuya eficacia causal sólo tiene sentido en el vocabulario propio de la química o la biología.

La tesis de que creencias, deseos y, en general, los estados mentales objeto de la psicología, funcionan como causas internas de nuestra conducta y lo hacen así en virtud de su contenido semántico se justifica en base al mismo tratamiento que el ofrecido para las ciencias especiales. El deseo de una cerveza fría más la creencia de que las cervezas están en la nevera causó el hecho de que Miguel fuera a la cocina y abriera el frigorífico. O, en general,

(x) (p) (q) [(x desea p) $\wedge$ (x cree (q $\rightarrow$ p)] $\rightarrow$ *ceteris paribus* x realizará q

Esta ha sido, por ejemplo, la línea de argumentación defendida por Fodor en dos importantes artículos: «Making Mind Matter More» (Fodor [1989]) y «You Can Fool Some of The People All the Time, Everything Else Being Equal: Hedged Laws and Psychological Explanations» (Fodor [1991]). En el primero de ellos, Fodor pone de manifiesto cómo la legitimación de leyes causales formuladas en términos pertenecientes a un nivel que no sea el nivel básico de la física es un problema que afecta no sólo a la psicología, sino a cualquiera de las ciencias especiales o no básicas. Este problema, sin embargo, no parece ser irresoluble si tenemos en cuenta que la única diferencia entre leyes causales básicas y no-básicas es que, en el caso de estas últimas, tiene que haber un *mecanismo* en virtud del cual la satisfacción del antecedente garantice la satisfacción del consecuente, o, en otros términos, tiene que haber un mecanismo que implemente tales leyes. Y el punto esencial es: aunque los mecanismos que implementen leyes intencionales —leyes relativas al contenido de nuestros estados mentales— tienen un carácter físico, las leyes en sí mismas son esencialmente intencionales y justifican la adscripción de eficacia causal a tales estados en tanto semánticamente interpretables.

Ahora bien, la existencia de mecanismos (físicos) de implementación, en el caso de las leyes no-básicas, las convierte en leyes no estrictas. Estas leyes incorporan siempre cláusulas *ceteris paribus* y, por tanto, la cuestión que surge inmediatamente es la de cómo es posible garantizar la adscripción de eficacia causal a las propiedades semánticas de los estados subsumidos bajo leyes intencionales cuando incorporan de forma inevitable las mencionadas cláusulas. El problema se agrava si, en base a consideraciones sobre la múltiple realizabilidad de los estados intencionales, argüimos, como hace, por ejemplo, Schiffer (Schiffer [1991]) que la propia noción de ley *ceteris paribus* no tiene ningún sentido en psicología, entre otras razones, porque siempre es posible encontrar excepciones significativas a esas leyes, significativas en el sentido de hacer imposible la determinación de sus condiciones de verdad[4].

4. La posición de Schiffer es, de hecho, i) que la defensa de leyes intencionales es absolutamente implausible dada la múltiple realizabilidad física de los estados mentales, y ii) que no obstante, la capacidad explicativa y el carácter predictivo de las teoríass intencionales no depende en absoluto de la existencia de tales leyes con o sin cláusulas *ceteris paribus*. La segunda de estas tesis es compartida igualmente por A. Clark (Clark [1993]), si bien compartida desde presupuestos diferentes. Y el comentario de ambas lamentablemente empieza y acaba en esta nota. Espacio obliga.

La respuesta de Fodor, recogida en el segundo de los artículos mencionados, es complicada y requiere de una cierta terminología especial, pero de manera esquemática, se reduce al siguiente argumento:

a) Se asume que cualquier tipo de estado intencional (*A*) en virtud del cual un organismo satisface el antecedente de una ley causal *ceteris paribus* es un estado funcional cuya realización física puede ser diferente en distintos organismos o en el mismo organismo en momentos diferentes.

b) El hecho de que las leyes intencionales incorporen cláusulas *ceteris paribus* indica que la eficacia causal de un tipo de estado intenciona (*A*) con respecto a una conducta determinada (*B*) requiere la existencia conjunta de alguna de las posibles realizaciones físicas de ese estado **y** de las condiciones adicionales (*C*) que se agrupan bajo esas cláusulas. Así, si R1 es una realización del tipo evento *A*, tales condiciones adicionales vienen representadas por un tipo de evento arbitrario *C* si:

   i) *A* (R1) y *C* es (estrictamente) suficiente para la producción de *B*.
   ii) No es el caso que sólo *A*(R1) es suficiente para la producción de *B*.
   iii) No es el caso que sólo *C* es suficiente para la producción de *B*.

c) De acuerdo con la formulación anterior, si R1 es una realización física del estado *A*, existen tres formas imaginables en que la presencia de R1 podría conllevar, sin embargo, excepciones a la ley «*ceteris paribus A* → *B*»:

   i) En primer lugar, cuando las condiciones adicionales (*C*) necesarias para la producción de *B*, aunque determinables, no se producen **conjuntamente** con la realización R1 de *A*. Tenemos entonces lo que Fodor denomina una *mera* excepción.
   ii) En segundo lugar, cuando no hay tales condiciones adicionales, *i. e.*, cuando no hay ningún evento de tipo *C* tal que conjuntamente con la realización R1 de *A* constituya una condición suficiente para la producción de *B*. En ese caso, todos los eventos particulares de tipo *A* realizados por R1 son excepciones *absolutas* a la ley.
   iii) Finalmente, cuando la realización R1 de *A* es una excepción absoluta no sólo con respecto a la ley «*ceteris paribus A* → *B*», sino con respecto a **todas** las leyes en las que *A* aparece en el antecedente («*ceteris paribus A* → *C*»; «*ceteris paribus A* → *D*», etc.).

De acuerdo con Fodor lo que distingue a las leyes estrictas de las leyes *ceteris paribus* es que estas últimas pueden tener excepciones, incluso excepciones absolutas en el sentido explicado en ii). Lo que distingue las leyes *ceteris paribus* de meras proposiciones vacías es que en las primeras, pero no en las segundas,

las realizaciones físicas que corresponden a los estados intencionales subsumidos bajo el antecedente no son excepciones absolutas a **todas** las leyes en las que aparece el mismo antecedente, *i. e.*, no son excepciones absolutas en el sentido explicado en iii).

La justificación última de por qué tales excepciones no son posibles y, en definitiva, la justificación del carácter nomológico de las leyes *ceteris paribus*, a pesar de la existencia de los otros tipos de excepciones, reside en que, en el caso de las leyes intencionales, la noción de realización física se define *funcionalmente*. Una vez asumida su naturaleza funcional, un estado físico que fuera una excepción absoluta al conjunto de leyes en las que un estado intencional *A* aparece como antecedente, simplemente no podría ser individualizado como una realización de tal estado intencional, porque no habría ningún criterio externo para determinar de qué estado estamos hablando. El problema, pues, no es tanto un problema relativo a la posibilidad de leyes *ceteris paribus* cuanto un problema relacionado con la correcta individualización funcional de los estados que aparecen en tales leyes.

Veamos cómo funciona esta estrategia en un caso concreto. Supongamos que existen únicamente tres leyes en las que el estado «querer perder peso» aparece como antecedente. Estas leyes son:

— *ceteris paribus*, las personas que quieren perder peso controlan su dieta alimenticia;
— *ceteris paribus*, las personas que quieren perder peso hacen ejercicio físico;
— *ceteris paribus*, las personas que quieren perder peso profieren frases del tipo «me gustaría perder algunos kilos».

Supongamos además que la realización física propuesta para tal estado es la de tener una configuración neuronal S y, por último, supongamos que las personas que presentan tal configuración son excepciones absolutas a las tres leyes mencionadas, *i. e.*, a **todas** las leyes que involucran el estado mental «querer perder peso».

Como hemos visto, esto quiere decir no sólo que, de vez en cuando, la gente que quiere adelgazar no controla su dieta o no hace ejercicio o no profiere frases de marras. Significa, más bien, que no existen condiciones adicionales C1 tales que si se tiene la configuración neuronal S y se poseen tales condiciones, entonces se controla la dieta **y** no existen condiciones adicionales C2 tales que si se tiene la configuración neuronal S y se poseen tales condiciones, entonces se hace ejercicio **y** no existen condiciones C3 tales que si se tiene la configuración neuronal S y se poseen tales condiciones, entonces se profiere «me gustaría perder algunos kilos».

En este caso, *i. e.*, cuando no existen condiciones que *completen* el estado de tener la configuración neuronal S con respecto a ninguna de las leyes en las que «querer perder peso» aparece como antecedente, tenemos razones suficientes, no para cuestionar el carácter nomológico de nuestras generalizaciones, sino para rechazar la idea de que «estar en S» es la realización física de «querer perder peso». Esto es así porque, después de todo lo que define el estado mental «querer perder peso» es su función, y su función es causar al menos alguna de las conductas mencionadas en nuestro ejemplo. De esta manera quedan excluidas las únicas excepciones que presentarían problemas en el proceso de validación de las leyes no estrictas.

Desde mi punto de vista, sin embargo, la defensa que hace Fodor de la legitimidad de las leyes *ceteris paribus* es, en cierto sentido, circular[5]. La razón de ello es la siguiente. Si el papel funcional de una representación o estado mental es —al menos en una versión amplia del funcionalismo— su papel causal, una defensa de las leyes *ceteris paribus* pensada para garantizar la adscripción de eficacia causal a las propiedades semánticas de tales estados que descansa, en última instancia, en su correcta individualización funcional —*i. e.*, **causal**—, no parece ser la mejor defensa posible. Una argumentación equivalente, e igualmente circular sería una que defendiera la validez de la ley «*ceteris paribus* si corta vidrio, entonces es un diamante» apelando al hecho de que la definición funcional de lo que sea un diamante incluye la propiedad de cortar vidrio.

Existe, no obstante, una línea argumentativa en favor de la legitimidad de las leyes *ceteris paribus* que no conlleva tales problemas. Tal línea se basa en una definición de causalidad a través de condicionales contrafácticos que tiene su origen en Lewis (Lewis [1986]), y en la reivindicación de una cierta noción de sobreveniencia que no sólo permite dar cuenta de la múltiple realizabilidad de las estructuras micro-físicas que subyacen a los mismos tipos de estados intencionales, sino que lo hace estableciendo una relación nomológica entre los estados así caracterizados. Examinemos pues en qué consiste este planteamiento alternativo.

5. «En cierto sentido» significa en el sentido que expongo a continuación y no en el sentido en que Schiffer cree que es circular, sentido que aparece contestado —satisfactoriamente— por Fodor en las últimas páginas del artículo mencionado, *i. e.*, Fodor (1991), pp. 31-33.

4. CONTRAFACTICOS Y SOBREVENIENCIA

En su formulación más básica, una definición de causalidad en términos de condicionales contrafácticos puede darse de la manera siguiente:

> «Si **c** y **e** son dos eventos actuales tales que **e** no hubiera ocurrido si **c** no hubiera tenido lugar, entonces **c** es causa de **e**» (Lewis [1986], p. 167).

Por supuesto, esta definición necesita precisiones importantes. Supongamos, por ejemplo, que una brasa al rojo vivo causa que mi cigarrillo se encienda. Podríamos decir que si la brasa no hubiera sido roja, mi cigarrillo no se hubiera encendido. Pero, aunque el condicional resultante es verdadero, ello no establece ningún tipo de relación causal entre las propiedades *ser rojo* y *estar encendido.* La rojez del carbón sólo es relevante para la producción del efecto en tanto propiedad resultante de la obtención de una cierta temperatura y, a su vez, la propiedad de tener tal temperatura no es más que una propiedad microfísica, un cierto estado de agitación molecular.

Si queremos garantizar que las propiedades de nivel macroscópico utilizadas para describir los eventos conectados contrafácticamente en una explicación causal son realmente eficaces con respecto a los outputs conductuales descritos, necesitamos establecer algún tipo de conexión que articule con carácter nomológico las relaciones existentes entre esas propiedades macroscópicas y los mecanismos que median en su eficacia causal. Esa conexión viene proporcionada por la noción de sobreveniencia, expuesta en su forma más general por Kim en los siguientes términos:

> «La sobreveniencia de una familia de propiedades A sobre una familia de propiedades B puede explicarse como sigue: necesariamente para cualquier propiedad P perteneciente a A, si un objeto x tiene la propiedad P, entonces existe una propiedad Q, perteneciente a B, tal que x tiene Q y necesariamente cualquier objeto que tenga la propiedad Q tiene la propiedad P. Cuando propiedades del tipo P y Q están relacionadas de acuerdo a esta definición, podemos decir que P sobreviene en Q y Q es la base de sobreveniencia de P» (Kim [1984], p. 262).

Kim utiliza la noción de superveniencia como idea base para argüir en contra del carácter epifenoménico atribuido a las propiedades mentales. La causación mental no es más que un caso de causación sobreveniente y, como tal, las propiedades semánticas de los estados que entran en relaciones causales sobrevenientes son causalmente eficaces (cfr. Kim [1984] [1988[).

Sin embargo, como dije anteriormente, lo que necesitamos para garantizar que las propiedades macroscópicas descritas en las explicaciones intencionales son causalmente eficaces es una articulación de carácter **nomológico** entres esas propiedades y los mecanismos que median su eficacia causal. La posición de Kim expuesta resulta en ese sentido demasiado débil, porque no requiere la existencia

de leyes que conecten los eventos correspondientes, mientras que los casos paradigmáticos de sobreveniencia —casos como ser líquido y causar humedad, etc.— requieren del establecimiento de tales relaciones nomológicas. Lo que necesitamos es añadir a la definición mencionada el requisito expuesto anteriormente, *i.e.*, que una propiedad es causalmente eficaz si es una propiedad en virtud de la cual los individuos que la poseen pueden ser subsumidos bajo leyes, posiblemente con cláusulas *ceteris paribus*. De hecho, es esta conjunción de posiciones la que parece estar detrás de los que Kim denomina «sobreveniencia mereológica», una noción que sólo pueden entenderse como un tipo de sobreveniencia que implica necesariamente una relación nomológica.

Si retomamos ahora la definición de causalidad en términos de contrafácticos junto con esta noción de sobreveniencia mereológica, podemos establecer una condición suficiente para garantizar la eficacia causal de una propiedad macroscópica cualesquiera (cfr. Segal y Sober [1991]):

Si

i) existe una ley causal (posiblemente con cláusulas *ceteris paribus*) que conecta eventos del tipo F con eventos del tipo G y que sustenta contrafácticos del tipo «G no hubiera ocurrido si F no hubiera tenido lugar» y
ii) en cada caso en que un evento de tipo F causa un evento del tipo G existen micro-propiedades m(F) y m(G) tales que las m(F) que constituyen la causa —o un subconjunto de m(F) en el caso de que m(F) sea un conjunto de tales micro-propiedades— causan las m(G) que constituyen el efecto y
iii) F sobreviene mereológicamente en m(F) y G sobreviene mereológicamente en m(G),

entonces,

F es causalmente eficaz en la producción de G.

Una vez formulada esta condición, el paso siguiente es mostrar que las propiedades semánticas de los estados mentales la cumplen.

5. PROCESOS COGNITIVOS Y PROCESOS COMPUTACIONALES

Si hubiera que establecer un objetivo general al conjunto de disciplinas que se agrupan bajo el rótulo de Ciencia Cognitiva, éste sería probablemente el de especificar las leyes intencionales que gobiernan los procesos cognitivos y el de establecer los tipos de mecanismos que implementaran esas leyes. En ese sentido,

el argumento que conduce a la reivindicación de la eficacia causal de las propiedades semánticas, *i. e.*, el argumento que muestra cómo las propiedades semánticas cumplen la condición expuesta en el apartado anterior, parece caer naturalmente bajo el ámbito de esta disciplina.

No es necesario, sin embargo, entrar en grandes precisiones para ver cómo se desarrolla este argumento. El pilar básico lo constituye la caracterización de los procesos cognitivos como procesos computacionales. Los procesos computacionales se definen, a su vez, en términos de representaciones. Las representaciones-input forman los argumentos de una función. Las representaciones-output constituyen los valores de la función computada. Una representación es así un tipo de configuración física muy especial, una configuración física que tiene una lectura sintáctica y una lectura semántica. Y el punto importante es que, aunque los procesadores de un computador son sensibles únicamente a la sintaxis de las representaciones, la máquina puede ser diseñada de tal modo que la producción de ítems sintácticos tiene sentido a la luz de la interpretación semántica impuesta por los problemas que se supone ha de resolver.

En el marco del paradigma clásico, la imagen de la mente como un computador supone una interpretación de los estados intencionales en términos de estados que involucran símbolos de un lenguaje *mental* o Lenguaje del Pensamiento (Fodor [1975]). Mi creencia de que hay un pastel de manzana en la nevera conlleva mi estar en algún tipo de relación computacional con el símbolo de *mentalese* correspondiente a «Hay un pastel de manzana en la nevera». El contenido de tal estado intencional es el contenido de ese símbolo en *mentalese* y el hecho de que sea una creencia —en lugar de un deseo o una duda— está determinado por la naturaleza de la relación computacional con el resto de mis estados mentales y/o mi conducta. Los símbolos de este lenguaje mental están dotados de una sintaxis combinatoria del tipo de la que rige en el cálculo proposicional y son implementados físicamente por patrones de excitación celular. De esta manera los procesos que subyacen a las relaciones entre estados intencionales son, en última instancia, procesos físicos.

Un planteamiento como este permite establecer leyes causales que relacionan estados intencionales en virtud de sus propiedades semánticas. El carácter nomológico de esas relaciones viene garantizado por la existencia de esos mismos estados bajo un nivel de descripción puramente físico, un nivel que constituye la base sobreviniente de las propiedades descritas a nivel intencional. Esas leyes contendrán muy probablemente cláusulas *ceteris paribus*, ya que fallos o alteraciones en el nivel físico pueden impedir el correcto funcionamiento del sistema, pero como se ha puesto de manifiesto, esto no es un problema importante. Lo importante es la existencia de una base de sobreveniencia adecuada y, bajo esta perspectiva, tal condición se cumple ya que la base de sobreveniencia

incluye todas aquellas propiedades físicas de las representaciones que explican los poderes causales de éstas a un nivel micro-físico.

La respuesta a la pregunta sobre si existen leyes intencionales de la forma «Toda instanciación de P causa una instanciación de Q» (posiblemente con cláusulas *ceteris paribus*) en las que «F» se refiere a una propiedad semántica es, a la luz de este planteamiento, claramente positiva.

Las propiedades semánticas de los estados mentales pueden verse como causalmente eficaces con respecto a conductas determinadas, aunque la clase de equivalencia a la que pertenecen no es descriptible en términos de características proyectables en el vocabulario físico. Por supuesto debe existir una descripción física, debe existir un mecanismo físico de implementación de las propiedades responsables de la causalidad semántica, pero tal caracterización física no es en absoluto la descripción relevante en la explicación de la eficacia causal. (Horgan [1989], McLaughlin [1989]).

Ahora bien, este planteamiento computacional representa una estrategia de acuerdo con la cual la adscripción de eficacia causal a un determinado estado intencional requiere algún tipo de *vehículo*, aislable a nivel del lenguaje del pensamiento, que pueda ser visto como el ítem modular responsable de la información contenida en tal estado y como el ítem implicado en todos los episodios cognitivos en lo que tal estado aparece. ¿Qué decir entonces acerca del programa conexionista, típicamente caracterizado por la ausencia de *objetos* estructurados sistemáticamente en la forma en que la sintaxis y la semántica combinatoria clásicas requiere? Como puse de manifiesto en la primera parte de este trabajo, la respuesta a esa pregunta parece conducirnos a un dilema: o bien aceptamos la inadecuación del conexionismo como modelo cognitivo o, en otro caso, nos vemos obligados a adoptar una posición eliminativista con respecto a las actitudes proposicionales que es equivalente a la negación de toda eficacia causal al contenido de las mismas.

En lo que resta, me propongo defender al conexionismo de ambas acusaciones y mostrar que el dilema es, como tantos otros, un falso dilema. La aceptación del conexionismo como modelo cognitivo no implica necesariamente la adopción de una postura eliminativista, siempre y cuando establezcamos claramente cuál es el nivel apropiado de análisis de tales modelos computacionales.

6. CONEXIONISMO Y NIVELES DE DESCRIPCIÓN

Los modelos conexionistas son redes complejas de elementos computacionales conectados en paralelo. Cada uno de estos elementos o unidades tiene un valor de activación que se establece numéricamente en función de los valores de

activación del resto de las unidades en la red y de la fuerza o carga (*weight*) asignada a las conexiones entre esas unidades.

La influencia de una unidad *a* sobre una unidad *b* es el resultado de multiplicar el valor de activación de la unidad *a* por la fuerza de conexión entre *a* y *b*. Así, si una unidad tiene un valor de activación positivo, su influencia en el valor de la unidad adyacente será positiva si la fuerza de conexión es positiva y negativa si la fuerza de conexión es negativa. En una alusión claramente neurológica, las conexiones con carga positiva se denominan conexiones excitatorias y conexiones inhibitorias las que tienen carga negativa.

Un modelo conexionista típico consta de tres conjuntos de unidades: las unidades input, las unidades output y las denominadas *hidden units* o unidades ocultas. La representación del input se establece a través de la imposición de valores de activación en las unidades input. Esta activación se prolonga a través de las conexiones entre el resto de las unidades input y las unidades *ocultas* hasta que un conjunto de valores de activación emerge en las unidades output. Las computaciones realizadas por la red en la transformación de los patrones de actividad del input hasta dar como resultado el patrón de actividad output depende así del conjunto de las fuerzas de conexión. Son estas fuerzas de conexión las que se consideran normalmente responsables del conocimiento del sistema y, en este sentido, juegan el mismo papel que un programa en un computador convencional (cfr. Smolensky [1988]).

Gran parte del interés despertado por las redes conexionadas reside en su capacidad de autoprogramación, *i. e.*, en la incorporación de procedimientos de aprendizaje a través de los cuales, y tras un período de entrenamiento en el que se somete a la red a un bombardeo de pares input / output, la red misma ajusta sus fuerzas de conexión y establece las funciones a computar por las unidades ocultas. Sin embargo, no es esta la característica más relevante en la discusión que nos ocupa. El punto crucial es el carácter distribuido y superposicional que tiene la codificación de la información en los modelos conexionistas. Efectivamente, los *mecanismos* responsables de las relaciones *input-output* en estos modelos son las unidades que conforman la red. Cada unidad contribuye a codificar información acerca de diferentes proposiciones —esto es lo que significa que las representaciones son superposicionales— y cada representación está así distribuida a lo largo de un conjunto amplio de micro-propiedades que son subsimbólicas, *i. e.*, que no son semánticamente interpretables en sí mismas. La producción de un determinado *output* supone la existencia de un determinado patrón de actividad de las unidades, aunque no necesariamente el mismo para representaciones del mismo tipo.

Puesto que las computaciones realizadas por la red están completamente

determinadas al nivel de las unidades, y estas unidades codifican al mismo tiempo información de tipo diferente, no es posible identificar una entidad estable y recurrente que respalde la noción clásica de símbolo y que esté sujeta a las manipulaciones de un sistema de procesamiento independiente. Por otra parte, puesto que la misma información puede ser representada por redes con unidades y fuerzas de conexión diferentes, la clase de esas redes —posiblemente indefinida— no es más que un «conjunto caóticamente disyuntivo» que no corresponde en absoluto a la clase natural que resultaría, de acuerdo con la psicología popular, al considerar el conjunto de agentes cognitivos que tienen una creencia determinada. La conclusión eliminativista / epifenomenalista a propósito de estas consideraciones es que no tiene ningún sentido preguntar si la representación de una proposición determinada juega o no un papel causal en las computaciones de la red, porque no existe un estado discreto que sea la información contenida en tal proposición, ni nada que sea comparable a una clase natural en el sentido psicológico (cfr. Ramsey, Stich y Garon [1991]).

El primer paso en la refutación de estos argumentos es una concesión al eliminativismo: si el nivel de análisis adoptado en la explicación de la conducta de los modelos conexionistas se reduce al nivel de las unidades del sistema —lo que Smolensky denomina nivel neurológico—, entonces las conclusiones eliminativistas son perfectamente plausibles. Pero, y este es el punto importante, un tratamiento adecuado del paradigma conexionista, en tanto que **modelo cognitivo**, ha de desarrollarse a un nivel de descripción superior al de las meras activaciones numéricas de las unidades y las fuerzas de conexión.

Una de las primeras lecciones aprendidas en Ciencia Cognitiva es la de la existencia de múltiples niveles de descripción con respecto a un modelo computacional. La elección de uno u otro de estos niveles impone fuertes restricciones en el tipo de explicación que podemos proporcionar de la conducta del sistema. No es extraño, por tanto, que limitándonos al nivel de fuerzas y unidades, obtengamos explicaciones poco satisfactorias desde el punto de vista de una conducta *semánticamente* interpretada. Sin embargo, dentro de este paradigma existen técnicas de análisis que permiten ascender a un nivel de descripción superior, un nivel de descripción bajo el cual se pueden identificar representaciones y transformaciones que cumplen el mismo papel que las representaciones y reglas de los sistemas clásicos —si bien tienen un carácter completamente diferente a los símbolos y algoritmos clásicos ya que tales representaciones y reglas son implícitas y altamente distribuidas.

Una de estas técnicas es la denominada análisis de grupo (*cluster analysis*). La idea básica de este método de tipo estadístico es la de extraer regularidades en los patrones de actividad que presentan las unidades ocultas para cada una de

las relaciones input-output y utilizarlas para construir representaciones que agrupan relaciones con patrones similares. Esta técnica permite así unificar bajo categorías semánticas definidas patrones de actividad que, al nivel de las unidades y conexiones, tienen una estructura interna diferente. Y lo más importante, permite establecer esta unificación semántica en función de outputs producidos a partir de inputs diferentes, *i. e.*, permite agrupar lo que al nivel de unidades son diferentes estados de una red en base a su eficacia causal respecto a una conducta determinada[6].

La misma técnica de análisis de grupo muestra que la parte del argumento eliminativista que ponía en cuestión la existencia de clases naturales, en el sentido en que esta noción se entiende en psicología, es innecesariamente reduccionista. El hecho de que distintas redes puedan representar la misma información a pesar de estar constituidas por unidades con actividad y fuerzas de conexión diferentes deja de ser un problema si el análisis de tales redes se desarrolla a un nivel de descripción donde los *bloques* explicativos no son tanto las unidades cuanto sus patrones de actividad. Puesto que esos patrones resultan de la agrupación de mecanismos físicos diferentes que producen, sin embargo, los mismos output, el análisis del sistema a este nivel de descripción superior nos permite unificar lo que antes parecía un «conjunto caóticamente disyuntivo» en la misma clase de equivalencia (cfr. Clark [1990]).

Así, el análisis de grupo cumple, con respecto al problema de establecer la adecuación del conexionismo como modelo cognitivo, el mismo papel que el análisis desarrollado anteriormente en términos de contrafácticos y sobreveniencia cumplía con respecto al problema de la causación mental. Ambos enfoques nos ofrecen la posibilidad de reivindicar la eficacia causal de las propiedades semánticas de los estados mentales / computacionales a pesar de la múltiple realizabilidad de las estructuras físicas subyacentes a tales estados. Y, lo más importante, el hecho de que tal reivindicación sea posible dentro del paradigma conexionista muestra que la existencia de ítems modulares estructurados de acuerdo con la sintaxis y la semántica clásicas no es un requisito empírico imprescindible en el funcionamiento de la estrategia computacional.

---

6. Si la red es suficientemente compleja, como lo es NETtalk (Sejnowski y Rosenberg [1986]), que toma texto escrito como input y produce fonemas como output, esta partición del espacio representacional adoptará una estructura jerárquica en forma de árbol. Así, la división final en esta red en vocales y consonantes se justifica en virtud del hecho de que los patrones de actividad que funcionan como representaciones para cada una de las vocales son más parecidos entre sí que los que funcionan para cada una de las consonantes. A su vez,k en el grupo de consonantes, los patrones de actividad correspondientes a las palatales, por ejemplo, pueden agruparse como tales debido a la mayor similaridad existente entre ellos que la existente en relación con los patrones correspondientes a labiales o nasales. El mismo principio se aplica para cada una de las categorías fonéticas.

Después de todo, el proceso de *etiquetado* simbólico de los patrones de actividad de las unidades de un sistema conexionista es un proceso que se lleva a cabo desde el *exterior* del sistema. Los patrones, en sí mismos, no son simbólicos en ningún sentido que sea equiparable al paradigma clásico. No lo son desde el punto de vista semántico, porque tales patrones representan únicamente conjuntos de micro-propiedades, micro-propiedades que, en el paradigma clásico, no constituyen la referencia de un único símbolo, ni desde el punto de vista sintáctico, porque las operaciones combinatorias sobre estos patrones no están regidas por reglas de carácter lingüístico, sino por computaciones exclusivamente numéricas. No obstante, su función es la de caracterizar la información representada por la actividad de las unidades ocultas, una información que es causalmente responsable de la conducta del sistema.

Los rasgos de esta «segunda metáfora computacional» representada por el paradigma conexionista han sido sólo brevemente esbozados. Los detalles técnicos y conceptuales, así como los problemas, aparecen ampliamente recogidos en la literatura sobre el tema. Sin embargo, incluso una caracterización tan breve como ésta nos proporciona las herramientas teóricas necesarias para poder completar el objetivo de este trabajo. A través de ella espero haber puesto de manifiesto que es posible justificar filosóficamente la eficacia causal del contenido de los estados mentales sin tener que desarrollar por ello una estrategia computacional que postule la existencia de ítems simbólicos y, por tanto, que es posible *salvar* al conexionismo del falso dilema en el que las conclusiones eliminativistas / epifenomenalistas parecían haberlo situado.

Josefa Toribio Mateas
Departamento de Lógica y Filosofía de la Ciencia. Facultad de Filosofía.
Universidad Complutense. Madrid

REFERENCIAS BIBLIOGRÁFICAS

Block, N. (1991): «Can the Mind Change the World», en Boolos, G. (ed.), *Essays in Honour of Hilary Putnam,* Cambridge, Cambridge University Press.

Clark, A. (1989): *Microcognition: Philosophy, Cognitive Science and Parallel Distributed Processing*, Cambridge, Mass., M.I.T. Press.

Clark, A. (1990): «Connectionist Minds», *Proceedings of the Aristotelian Society,* vol. XC, pp. 83-102.

Clark, A. (1991): «Radical Ascent», *Proceedings of the Aristotelian Society,* vol. sup. LXV, pp. 211-227.

CLARK, A. (1993): *Associative Engines. Connectionism, Concepts and Representational Change,* Cambridge, Mass., M.I.T. Press, en prensa.

CLARK, A., y KARMILOFF-SMITH, A. (1993): «The Cognizer's Innards: a Psychological and Philosophical Perspective on the Development of Thought», *Mind & Language,* en prensa.

CHURCHLAND, P. (1986): *Neurophilosophy,* Cambridge, Mass., M.I.T. Press.

CHURCHLAND, P. (1989): *The Neurocomputational Perspective,* Cambridge, Mass., M.I.T. Press.

DAVIES, M. (1991): «Concepts, Connectionism and the Language of Thought», en Ramsey, W.; Stich, S., y Rumelhart, D. (1991), pp. 229-257.

DRETSKE, F. (1988): *Explaining Behaviour: Reasons in a World of Causes,* Cambridge, Mass., M.I.T. Press.

FODOR, J. (1975): *The Language of Thought,* New York, Thomas Y. Crowell.

FODOR, J. (1987): *Psychosemantics. The Problem of Meaning in the Philosophy of Mind,* Cambridge, Mass., M.I.T. Press.

FODOR, J. (1989): «Making Mind Matter More», *Philosophical Topics,* 17, 1, 59-79.

FODOR, J. (1991): «You Can Fool Some of the People all of the Time, Everything Else Being Equal; Hedged Laws and Psychological Explanations», *Mind,* vol. C (1), pp. 19-34.

FODOR, J., y PYLYSHYN, Z. (1988): «Connetionism and Cognitive Architecture», *Cognition,* 28, 3-71.

HORGAN, T. (1989): «Mental Quasation», *Philosophical Perspectives,* 3, pp. 47-74.

JACKSON, F., y PETIT, P. (1988): «Functionalism and Broad Content», *Mind,* 97 (387), pp. 381-400.

JACKSON, F., y PETIT, P. (1993): «Causation in the Philosophy of Mind», en Clark, A., y Millican, P. (eds.), *Proceedings of the 1991 Turing Colloquium,* Oxford, Oxford University Press, en prensa.

KIM, J. (1984): «Epiphenomenal and Supervenient Causation», en French, P., *et al.* (eds.), *Midwest Studies in Philosophy, IX,* Mineapolis, University of Minnesota Press.

KIM, J. (1988): «Supervenience for Multiple Domains», *Philosophical Topics,* 16 (1), pp. 129-150.

LEPORE, E., y LOEWER, B. (1987): «Mind Matters», *Journal of Philosophy,* 84, 11, pp. 630-642.

LEWIS, D. (1986): *Philosophical Papers,* vol. 2, Oxford, Oxford University Press.

MCCLELLAND, J.; RUMELHART, D., and the PPD Research Group (1986): *Parallel Distributed Processing: Explorations in the Microstructure of Cognition,* vols. 1 y 2, Cambridge, Mass., M.I.T. Press.

MCLAUGHLIN, B. (1989): «Type Epiphenomenalism, Type Dualism and the Causal Priority of the Physical», *Philosophical Perspectives,* 3, pp. 109-135.

PINKER, S., y PRINCE, A. (1988): «On Language and Connectionism. Analysis of a Prallel Distributed Processing», *Cognition,* 28, pp. 73-193.

RAMSEY, W.; STICH, S., y GARON, J. (1991): «Connectionism, Eliminativism and the Future of Folk Psychology», en Ramsey, W.; Stich, S., y Rumelhart, D. (1991), pp. 199-228.

RAMSEY, W.; STICH, S., y RUMELHART, D. (eds.) (1991): *Philosophy and Connectionist Theory,* Londres, Lawrence Erlbaum.

SCHIFFER, S. (1991): «Ceteris Paribus Laws», *Mind,* vol. C (1), pp. 1-17.

SEGAL, G., y SOBER, E. (1991): «The Causal Efficacy of Content», *Philosophical Studies,* 63, 1-30.

SEJNOWSKI, T., y ROSENBERG, C. (1986): *NETtalk: A parallel network that learns to read aloud,* John Hopkins University, Technical Report JHU/EEC-86/01.

SMOLENSKY, P. (1987): «Connectionist AI, and the Brain», *Artificial Intelligence Review,* 1, pp. 95-109.

SMOLENSKY, P. (1988): «On the Proper Treatment of Connectionism», *Behavioural and Brain Sciences,* 11, 1-74.

STICH, S. (1983): *From Folk Psychology to Cognitive Science,* Cambridge, Mass., M.I.T. Press.

VAN FRAASEN, B. (1980): *The Scientific Image,* Oxford, Clarendon Press.