Δαίμων 3 – 1991. Págs. 81-111

Davidson, Fodor, Dennett y conexionismo/s: razones y causas en los marcos actuales de explicación causal de la acción racional

JESÚS EZQUERRO

ABSTRACT

The aim of this paper is to show that some current theoretical frames for the causal explanation of rational action, that is, the accounts of Davidson and Fodor, cannot be successful in their attempts to answer satisfactorily the following question: «Why does our strategy in explaining rational action by ascribing beliefs and desires works?». The reason of the failure, in my opinion, could be that one: the proposals mentioned above try to account simultaneously for the causal and rational features of human action. Taking a different path, I propose, with D. Dennett, that there is room in cognitive science for two different projects, each one accounting in its own way for the rational and causal features separately. Moreover, if we want to preserve a realist conception of representations, a three level (at least) cognitive architecture must be posited, being the two lower levels of connectionist sort, and classical the highest one, however modified in its ideal assumptions.

INTRODUCCIÓN

¿Por qué resulta tan eficaz la práctica de adscribir creencias, deseos y otras actitudes proposicionales a las personas con el fin de explicar y predecir sus decisiones y acciones?. Esta es la cuestión a la que, desde aproximadamente tres decenios, tratan de responder los filósofos de la accion y, en general, los filósofos de la ciencia cognitiva sin que se haya podido llegar hasta la fecha, según mi punto de vista, a una solución satisfactoria. Desde que Donald Davidson publicó su polémico escrito «*Acciones, ra-*

zones y causas» (1963) [1], y G. H. von Wright su no menos polémico libro «*Explicación y comprensión*» (1971) [2], han corrido ríos de tinta acerca de el estatuto teórico de los modelos de explicación racional de la acción o de los modelos de explicación de la acción racional (veremos que estos últimos modelos forman una subclase de los primeros), y sin embargo continúan los desacuerdos acerca de las cuestiones más básicas.

Si uno es naturalista, es decir, si uno piensa que los estados y procesos internos (psicológicos, cognitivos, mentales, o como los queramos denominar) a los que se alude en la explicación de la acción humana son procesos que tienen lugar en algún tipo de sustancia material, es decir, si uno es monista o materialista, entonces se verá inmerso en serios problemas para diseñar un marco conceptual adecuado en el que se pueda hablar propiamente de explicación causal de la acción. La razón es la siguiente: el hecho de pretender hacer explicaciones causales de la acción racional impone, a primera vista, algunas restricciones a las que no tienen por qué someterse aquellos que no comparten dicho compromiso. Más específicamente: debe responder de forma mínimamente satisfactoria a la cuestión con la que hemos comenzado este escrito, es decir, debe garantizar, o mostrar, que la acción a explicar fue **causada** por las «razones» o «actitudes proposicionales» que se adscriben a los agentes en las premisas (*explanans*) del modelo explicativo, lo cual requiere a su vez:

(i) ofrecer algún marco conceptual en el que se articule con cierta plausibilidad la idea de que las «actitudes proposicionales», las «razones» (creencias, deseos...), poseen propiedades causales. Es decir, intervienen en relaciones causales con el entorno (*inputs*), con otras actitudes (o estados mentales internos), y con la conducta abierta (*outputs*).

(ii) ofrecer algún tipo de garantía de que las «razones» o «actitudes proposicionales» adscritas para explicar o predecir una acción son las causas de la conducta en que consiste dicha acción, y de que lo son ***qua* razones**, y que actúan del **modo adecuado** (*in the right way*), con el objeto de evitar el problema de las cadenas causales caprichosas (*wayward causal chains*) [3].

(iii) ofrecer, igualmente y en correspondencia con el punto anterior, alguna clase de

---

1 Además de muchos escritos dispersos provocados por la tesis que D. Davidson mantuvo en éste, puede hallarse un elenco representativo en las dos recopilaciones siguientes dedicadas exclusivamente a analizar la obra de Davidson: Vermazen & M. Hintikka, 1985 y E. Lepore & B. P. McLaughlin, 1985.

2 Buena parte de la literatura más representativa correspondiente a la polémica generada por esta publicación de von Wright puede verse en Manninen & Tuomela (eds.), 1976. La tendencia representada por von Wright, conocida como corriente Neo-hermenéutica o Nuevo dualismo, defiende la tesis de la imposibilidad de explicación científica de la acción humana. En tal medida, y también por razones de espacio, no la trataré, puesto que mi interés en este trabajo se centra en evaluar las dificultades de las concepciones naturalísticas para la explicación de la acción «por razones».

3 Para una formulación, ejemplos y propuestas de tratamiento de este problema, pueden verse los escritos de R. Chisholm 1964 y 1970.

garantía (en el sentido de contrastación, falsación, etc...), acerca de nuestra capacidad, como individuos que pretenden explicar las acciones de los demás, para adscribir actitudes proposicionales a las personas de modo fiable, en el sentido de «acertar» con las razones que causan su conducta (lo que D. Davidson denomina las razones 'primarias'). Recuérdese a este respecto que, como G.M. Anscombe señaló acertadamente, toda acción es intencional (o racional) bajo una u otra descripción. La solución a este problema conlleva a su vez tener de algún modo resuelto el no menos agudo problema de la individuación y/o identidad de las actitudes proposicionales. Veremos que va a ser precisamente este problema el talón de Aquiles de los diversos enfoques a analizar.

Bien, podrá parecer que estas son unas condiciones excesivamente restrictivas, pero por el momento no he hecho otra cosa que tratar de reflejar, de forma resumida, las principales constricciones señaladas por los diversos autores acerca del estatuto teórico de nuestras explicaciones racionales de acción. Las explicaciones de la **acción racional**, que como ya he anticipado, forman una subclase de las anteriores, se hallan sometidas, naturalmente, a un conjunto de restricciones más severo. En lo que sigue me centraré principalmente en este último caso, en la medida en que permite, dadas sus condiciones más exigentes, servir de *test* o piedra de toque para poner de manifiesto las limitaciones de los diversos enfoques hasta ahora intentados para explicar causalmente la accion intencional.

CONDICIONES DE LA EXPLICACIÓN DE LA ACCIÓN RACIONAL

Recientemente Jon Elster, en su *The Nature and Scope of Rational-Choice Explanation* (1985), expresaba las condiciones que a su juicio deben cumplir las que él denomina «explicaciones ideales de la elección racional». Son las siguientes:

(a) Deben mostrar que la acción a explicar es la mejor forma (y la única) de satisfacer el conjunto total de los deseos del agente, dadas las mejores (y únicas) creencias que el agente haya podido formar, relativas a la cantidad óptima (y únicamente determinada) de la evidencia.

(b) Deben mostrar que la acción fue causada (de la forma adecuada) por los deseos y las creencias, y las creencias causadas (de la forma adecuada) por la evidencia considerada (Cfr. J. Elster, op, cit. p. 71).

J. Elster denomina a la condición (a) el aspecto de la **optimalidad** y a la condición (b) el aspecto **causal** de las explicaciones de acción. La explicación a intentar debiera ser, naturalmente, aquella que cumpliera ambos requisitos. No obstante se muestra bastante pesimista (al igual que D. Davidson) respecto a las perspectivas de hallar pistas fiables acerca de la historia causal de cualquier acción (desde la línea causal

evidencia-creencias, hasta la relación causal deseos-creencias-acción). Hacerlo supondría tener algún tipo de acceso a la «maquinaria» psíquico-cerebral de los agentes, y eso es algo que difícilmente podemos hacer. A pesar de ello, Elster opina que, por razones puramente pragmáticas debiéramos conformarnos con explicaciones de acción que incorporen sólamente el primer aspecto, el de optimalidad, aunque a renglón seguido también se muestra pesimista respecto a las posibilidades de llevar a buen puerto esta empresa más modesta. Incluso un modelo explicativo que soslaye los aspectos causales de la acción y tome en consideración únicamente los de optimalidad (o racionales) tiene una viabilidad ciertamente dudosa, al descansar en tres postulados, no sólo excesivamente fuertes, sino también insolubles, según los casos, desde las propias bases teóricas de la Teoría de la Decisión. Se trata de los postulados de **unicidad** señalados en (a): determinación única de la evidencia óptima; de las creencias óptimas, dada la evidencia; y de la acción óptima, dadas las creencias y los deseos [4].

Christopher Cherniak señalaba, en su *Minimal Rationality* (1986), que ha sido precisamente la asunción (tácita) de una concepción de la racionalidad excesivamente idealizada la que ha conducido o bien a sostener la tesis de la **autonomía de lo mental** (su carácter no-nomológico), o bien a mantener concepciones **instrumentalistas** del modelo creencias-deseos. La argumentación de Cherniak dicurre en una línea bastante próxima a la que acabamos de ver en Elster. En esencia, señala que esta o estas concepciones ideales de la racionalidad no son realizables en los seres humanos de carne y hueso, ya que requieren una habilidad deductiva ideal y no son compatibles, por consiguiente, con el carácter finitario de las capacidades cognitivas humanas . En consecuencia, el único camino que Chr. Cherniak contempla para tratar de sortear el dilema anterior, es intentar formular concepciones de la racionalidad realizables en los seres humanos, lo que él denomina **racionalidad mínima**, de modo que alguna de ellas pudiera ser candidata a figurar como marco explicativo real de los procesos de acción y decisión humanos. Y adviértase que, puestos a evitar las opciones de la autonomía de lo mental y la concepción instrumentalista de los modelos de explicación «por razones», la única opción que resta, a primera vista, es la realista, que es la que parece favorecer Cherniak [5]. Personalmente no creo que el hecho de haber asumido concepciones de la

---

4 Entre las dificultades señaladas por J. Elster son destacables las siguientes: respecto al proceso mismo de formación de creencias a partir de la evidencia disponible existen con frecuencia casos en los que, por decirlo así, no se sabría cuándo parar de recopilar información para poder afirmar que el proceso de formación de creencias ha sido **racional**. Por contra, en muchas situaciones de decisión podría ser **irracional** consumir un tiempo excesivo en formar **racionalmente** las creencias sobre las que se va a decidir. Tenemos pues, que existen casos (no poco frecuentes), en los que ser racional en la adopción de decisiones implica actuar en base a unas creencias adquiridas irracionalmente y viceversa. Hay otros casos para los que, por definición, no existe una solución óptima, por lo que difícilmente se podrían cumplir los postulados de unicidad, etc...Cfr. J. Elster, Op. cit.

5 «And I attempted to construct more psychologically and computationally realistic models of the «minimal agent»»(Cfr. Chr. Cherniak, Op. cit. **Preface**). «I will further propose that such minimal rationality conditions are indispensable for adequate cognitive theory. What is at stake concerns the very possibility of a cognitive science and of **a realist interpretation of it**». Op. cit. p. 3 (el énfasis es mío).

racionalidad excesivamente ideales haya sido la única razón, ni siquiera la principal, por la que muchos filósofos han dudado de la interpretación realista del modelo creencias-deseos. Creo más bien que ha sido básicamente el problema de la **intencionalidad** (el hecho de que las actitudes proposicionales posean propiedades semánticas y las múltiples y complejas consecuencias que se derivan de ello) lo que ha empujado a muchos filósofos -naturalistas y no-naturalistas- a considerar muy poco plausible el intento de hacer la psicología popular científicamente respetable bajo una interpretación realista. En cualquier caso pienso, como veremos más adelante, que la propuesta de Cherniak de formular modelos de racionalidad mínima es un trabajo necesario para poder desarrollar modelos computacionales de razonamiento y acción dentro de un marco conceptual diferente al de los ordinariamente conocidos como funcionalistas o **clásicos** [6].

Las propuestas de Elster son, en cierto modo, similares a las de Cherniak, pero en aspectos fundamentales difieren sensiblemente. Por una parte, Elster critica como implausible, al igual que Cherniak, la posibilidad de realización de los modelos (ideales) de la elección racional, y las consecuencias que extrae de esas críticas asimismo guardan semejanzas, en la medida en que también propone debilitar los requisitos de racionalidad exigibles a los agentes. Por otra parte, sin embargo, su decisión de prescindir de los aspectos causales, de no contemplar la posibilidad de dar cuenta nómicamente de los vínculos causales entre evidencia, actitudes proposicionales y acciones, no descansa, como vendría a sostener Cherniak, en la constatación de las dificultades de implementación de los modelos (excesivamente idealizados) de racionalidad, sino en una observación previa: las dificultades de acceso a la causalidad psíquica de los agentes (Cfr. J. Elster, op. cit., pp. 62 y 71).

El problema que se suscita con la propuesta de Elster vuelve a ser, de nuevo, que al obviar el aspecto causal de la explicación, cierra el camino de antemano a la posibilidad de hallar pistas respecto a la cuestión con la que hemos comenzado, pues, o se adopta una solución 'a la Davidson', según la cual el hecho de que las explicaciones «por razones» funcionen, es completamente independiente, desde el punto de vista epistemológico, del hecho de que las actitudes proposicionales que los humanos poseemos correspondan (en el sentido de ser idénticas) a estados cerebrales. Si corresponden, cosa que parece necesaria para poder adjudicarles propiedades causales, será por haber hecho una opción ontológica previa, por no ser dualistas ontológicos, en una palabra. Pero resulta que es justamente la tremenda dificultad de integrar los aspectos causal y racional de la acción humana lo que ha conducido a otros a afirmar el dualismo [7].

---

6 La denominacion de «modelos clásicos» en referencia a los modelos de psicología cognitivo-computacional que usualmente asumen la Teoría Representacional de la Mente desde una interpretación realista, es ya moneda corriente sobre todo en la literatura dedicada a contrastar dichos modelos con los conexionistas. Puede verse, por ejemplo, J. Fodor & Z. Pylyshyn, 1988.

7 Arthur C. Danto (1976): sostiene, con palabras ajustadas, la permanencia de la dualidad cartesiana precisamente por esa razón: «But representations, just because they admit of assessment in terms of truth

Bien, creo que con lo que llevamos dicho se van perfilando los extremos del problema. De un lado, las dificultades que plantean los modelos de racionalidad ideales conducen a Cherniak a la búsqueda de modelos de racionalidad mínima, supuestamente realizables, **dentro de un marco conceptual realista-representacionalista**. De otro, J. Elster, además de aducir razones similares, añade la dificultad *per se* de acceder a la «maquinaria» psicológica de los agentes para poner en cuestión la viabilidad de los modelos explicativos de acción que cumplan las dos condiciones (a) y (b) anteriores. Ello le conduce a situarse en un marco conceptual aparentemente muy próximo al de Donald Davidson con su **monismo anómalo** [8]. Mi propósito para la siguiente sección va a consistir en mostrar (o al menos aportar algunos elementos de juicio), con argumentos independientes a los de J. Elster, que el cumplimiento simultáneo de las dos condiciones anteriores desborda ampliamente tanto el marco davidsoniano como el realista-funcionalista de J. Fodor, lo que me conducirá a una opción instrumentalista *sui generis* en un sentido muy cercano al expuesto últimamente por Daniel Dennett que, aun planteando dificultades respecto a ofrecer una respuesta satisfactoria a la cuestión con la que hemos comenzado, pienso que es rehabilitable desde un marco conceptual un tanto diferente.

MARCOS CONCEPTUALES HEREDADOS

1. La tesis de la autonomía de lo mental (D. Davidson)

El **monismo anómalo** de D. Davidson, como es bien conocido, se presenta como resultado de la reconciliación de tres supuestos aparentemente inconsistentes: (1) existe interacción causal entre eventos mentales y eventos físicos; (2) la causalidad posee un carácter nomológico (es decir, allá donde se den relaciones causales deben existir leyes generales que las conecten o subsuman bajo una u otra descripción); (3) no existen leyes causales que conecten eventos mentales con eventos físicos.

Sobre esta base ontológico-epistemológica Donald Davidson construye su teoría de la acción, la cual se podría resumir en los dos siguientes principios:

P1. La etiología causal de una acción es determinante para su caracterización como

---

and falsity, are logically external to the world in which they may in every other respect be located. As agents and knowers, indeed, we are within the world under the concept of causation, and external to it under the concept of truth. Within and without the world at once: that is the philosophical nature of man», (Cfr. A.C. Danto, op. cit. p. 24)

8 Al contrario de lo que deja entrever J. Elster en el escrito de referencia, el argumento del incumplimiento de los axiomas de la Teoría de la Decisión, sí creo que desempeña un papel relevante en D. Davison para su negación de la posibilidad de leyes psicofísicas (Cfr. ver D. Davidson 1970, 1973 y 1974). En este sentido, el análisis de Elster podría ser considerado como un buen desarrollo de la argumentación de Davidson.

una acción intencional: una acción será intencional si y sólo si está causada de cierta manera (con este requisito se pretende atender al objetivo de que sean las razones, en su calidad de eventos físicos, y no otro tipo de eventos no racionales o no intencionales, las que causan la conducta a explicar);

P2. Explicar una acción consiste en señalar o indicar sus causas (que son «razones»), puesto que, en virtud del punto anterior, se supone que la conducta intencional está causada por las mismas actitudes proposicionales que, simultáneamente, la «racionalizan» y la constituyen como intencional [9].

En este contexto se comprende que Davidson atribuya a la racionalidad un **papel constitutivo** como requisito para moldear nuestro pensamiento acerca de las actitudes proposicionales, pues el propio carácter de las explicaciones de acción intencional requiere la consideración de los agentes cuya conducta se pretende explicar, y a nosotros mismos, en tanto que sujetos de la explicación (o explicadores) como «animales racionales»(Cfr. D. Davidson, 1984). Dado este marco conceptual, podemos pasar a examinar brevemente en qué medida cumple las condiciones (i), (ii) y (iii) propuestas al comienzo y las expresadas idealmente por J. Elster. A primera vista se podría conceder que un principio de racionalidad mínimo, reducido en la práctica a la mera consistencia entre actitudes y conducta, funciona de modo ordinario y constante en la acción racional-intencional, con las consabidas excepciones de la «conducta acrática» y la «debilidad de la voluntad» [10], casos que, eventualmente, podríamos incluir en la cláusula *ceteris paribus* del modelo explicativo correspondiente. De este modo, las restricciones (i) y (ii) parecen tener cabida en el marco davidsoniano: por una parte, los supuestos definitorios del monismo anómalo pretenden satisfacer (i), en el sentido de que las actitudes proposicionales podrían poseer propiedades, y entrar en relaciones causales, en virtud de ser idénticas a estados y eventos físicos, aunque, por supuesto, el carácter nómico de estas relaciones causales sólo tendria cabida al nivel de su descripción física y no de su descripción intencional; por otra parte, los principios P1 y P2 de su teoría de la acción, responden al propósito de cumplir (ii).

Los problemas, sin embargo, surgen con la condición (iii), ya que las garantías que nos puede ofrecer el **principio constitutivo de racionalidad** a la hora de identificar o individuar las actitudes proposicionales de las demás personas o agentes descansan en un acto con un sesgo claramente empático que convierte en dudoso el carácter explicativo-causal del modelo. La razón es interna a los propios supuestos conceptuales de Davidson al prohibirnos la formulación de leyes psico-físicas (o más bien negar su posibilidad). Sin duda, la información que nos permite adscribir actitudes proposicionales a los agentes se encuentra en su entorno, con lo que tendríamos que nuestras

9 La exposición y defensa del monismo anómalo por parte de D. Davidson puede verse en sus tres escritos, 1970, 1973 y 1974 ya citados.

10 Davidson trata estos problemas en diversos escritos, y a mi modo de ver, de forma bastante aceptable. Pueden consultarse especialmente sus 1970a, 1982 y 1985.

posibilidades como adscribientes de actitudes (estados mentales-cerebrales) a los agentes no pueden aspirar a ser nómicas, y sin este requisito no se cumple, ni la condición (a) de J. Elster, ni, lo que es más grave, el propio principio de explicación P1 que Davidson propone. Nos encontramos así con un marco conceptual imposible de llenar: las explicaciones que podrían caber en él no son factibles bajo sus propios supuestos.

Este problema puede verse con más claridad, en la medida en que se agudiza, si pasamos del contexto de la explicación de la simple acción intencional al de la explicación de la acción racional. La conclusión, a mi juicio, vale tanto para la una como para la otra, pues la exigencia de racionalidad (ideal o no) en los agentes no elimina la necesidad que tenemos, como agentes externos, de basarnos en el principio constitutivo de racionalidad a la hora de adscribir actitudes. No olvidemos que nuestro problema en este momento es el cumplimiento de la condición (iii). En el caso de la explicación de la acción racional se requiere, como señalaba Elster, no sólo la consistencia entre actitudes y conducta, se requiere igualmente la racionalidad de las actitudes (dada la evidencia), y entre las actitudes internamente.

Como hemos visto en la sección anterior, la exigencia de consistencia interna entre creencias y deseos constituye un requisito excesivamente fuerte por razones teóricas internas a la Teoría de la Decisión, pero también lo es por razones, llamémosles, prácticas, en lo que respecta a la viabilidad del modelo explicativo, pues correríamos el riesgo de confundir la irracionalidad con la incompetencia mental. Existen muchos casos de actitudes intransitivas, inconsistentes en el tiempo, etc..., además de los que subyacen a las críticas de Cherniak y Elster ya mencionados que, en principio, no sería correcto incluir en la cláusula *ceteris paribus* del modelo [11]. Tal maniobra tendría todo el aspecto de un recurso *ad hoc* y lo vaciaría de contenido empírico. Si nos centramos, por otra parte, en el problema de la relación evidencia-creencias, que es la que más nos interesa de cara a evaluar la fiabilidad de la adscripción de contenido y el cumplimiento de la cláusula (iii), podremos comprobar hasta qué punto la exigencia de racionalidad, o de fundamentación de las creencias, resulta incompatible con los supuestos doctrinales del propio D. Davidson.

Para poder afirmar que una creencia (o conjunto de creencias) está fundamentada, debería cumplir las tres condiciones (ideales) siguientes:

---

11 M. Minsky, 1975 realizó una durísima critica a la idea misma de utilizar aproximaciones lógicas para construir modelos de representación del conocimiento. El problema que señalaba es que un marco que cumpla las metapropiedades de la lógica clásica no puede dar cuenta del aprendizaje (en nuestra terminología, de la formación o adquisición de creencias); hacerlo podría poner en peligro la consistencia, en la medida que en se introducen premisas nuevas, con lo que la propia idea de utilizar sistemas monótonos para representar el conocimiento también se pone en cuestión. En su lugar, él propone su conocida teoría de los *Frames*. Por otra parte, debemos tener en cuenta que las diversas presentaciones de la teoría de la decisión buscan (y en tal medida se considera una virtud): cumplir las metapropiedades de la lógica clásica. El problema que se deriva de todo ésto es si la tradicional identificación de la racionalidad con una teoría ideal de la decisión que cumple las metapropiedades de la lógica clásica debe ser revisada.

a) C tiene que ser la mejor creencia, dada la evidencia (E) disponible.
b) C tiene que estar causada por la evidencia (E) disponible.
c) E tiene que causar C «del modo adecuado» (Cfr. J. Elster, op. cit., pp. 63-64).

Si estos requisitos son necesarios para la explicación de la acción racional, está claro que en el marco davidsoniano difícilmente se pueden cumplir. Para Davidson, adquirir una creencia consiste en formar un un estado mental de carácter lingüístico. Como sabemos, para Davidson el lenguaje es holista y, por consiguiente, nuestros pensamientos o estados mentales también tienen ese carácter. Precisamente en su aceptación del holismo y de su consecuencia, la indeterminación de la traducción [12], es donde se asienta su doctrina de la traducción radical, en la que desempeñan un papel fundamental el **principio constitutivo de racionalidad** y el **principio de caridad**. Mediante estos principios se imponen las restricciones apropiadas a la racionalidad de modo que resulte manejable para habérnoslas con la árdua tarea de comprender a los demás, para que la comunicación resulte posible en un grado aceptable. Pero, por contra, son estos mismos principios los que excluyen las leyes psicofísicas y la reducibilidad de lo mental a lo físico. En ellos se asienta la argumentación de Davidson para negar su posibilidad. Para que no hubiese indeterminación de la traducción se necesitaría una teoría causal del significado mental, que en los casos ideales debería cumplir las tres condiciones anteriores. Pero una teoría así demandaría la existencia de leyes psicofísicas: es decir, relaciones nómicas entre estímulos o estados y acontecimientos del mundo descritos en terminología física con estados y acontecimientos descritos en terminología mentalística o intencional.

Dentro del marco davidsoniano la condición a) para la fundamentación de las creencias debería ser sustituída por la siguiente:

a') C tiene que ser la mejor creencia, dada la evidencia disponible, y **dado el conjunto parficular de creencias y estados mentales que el agente posee**,

pues, qué significa E para un agente, depende de sus creencias previas y, a falta de una teoría causal del significado (que presupone la existencia de leyes psicofísicas), la explicación de la acción racional, y también la de la mera acción intencional, quedan indeterminadas. J. Elster dice que, a diferencia de lo que sucede con las creencias, todavía no contamos con una definición adecuada de la racionalidad de los deseos y las preferencias, sin embargo, por lo que llevamos dicho, cabría suponer, *mutatis mutandis*, que tendrían problemas similares.

En resumen, es el carácter holista e intencional de los estados mentales lo que conduce a Davidson a rechazar la reducibilidad de lo mental a lo físico y a afirmar la autonomía de lo mental. Esta tesis, sin embargo, no puede por menos que causar cierta

12 Para una interpretación, que personalmente creo la más correcta, en el sentido de que en Quine la tesis de la indeterminación de la traducción y el resto de sus tesis, se derivan de su aceptación previa del holismo y no al revés, pueden verse G. D. Romanos, 1983 y P. Roth, 1984.

desazón en la medida en que por mucho que la negación de la nomicidad de lo mental sea consistente con los supuestos davidsonianos anteriores y con su propia concepción de la causalidad, la autonomía de lo mental no lo es tanto con respecto a su afirmación monista. Alternativamente, algunos autores han sostenido que el monismo anómalo está abocado o bien a la inconsistencia o bien al epifenomenalismo [13]. Desde mi punto de vista, sin embargo, la concepción davidsoniana admite una evaluación bastante más consistente si se la dota de una lectura instrumentalista en un sentido similar al que ofrece Daniel Dennett, 1987, permitiéndole conservar intacta la fuerza de su argumentación en contra de la reducibilidad de lo mental [14].

2. **Realismo-representacionalista (J. Fodor)**

La empresa de J. Fodor [15] se podría resumir, en sus propias palabras, como un intento de mostrar que los deseos y creencias son reales, intervienen en relaciones causales y son estados cargados de contenido semántico. En resumen: mostrar que es posible una psicología científica que reivindica la explicación de la conducta, practicada en la psicología popular, por creencias y deseos. De este modo, si para D. Davidson el Principio de Racionalidad es un principio constitutivo para poder hablar de conducta intencional, para J. Fodor, dicho principio, o más bien, una Teoría de la Decisión, forma parte de la base de datos estructural del **Lenguaje del Pensamiento** que, según presupone, todos los humanos poseemos como dotación genética (Cfr. J. Fodor, 1975, cap. 1, pp. 47-49).

La idea de que las actitudes proposicionales poseen propiedades causales (condición (i)) se intenta resolver por J. Fodor en el marco del **Fisicalismo de Casos** (*Token Physicalism*). Dentro de este marco, la hipótesis del Lenguaje del Pensamiento (en adelante, LP) intenta construir los casos (*tokens*) de las actitudes proposicionales como relaciones con casos (*tokens*) de símbolos. Más concretamente, al poseer los agentes un

---

13 Por ejemplo, Ted Honderich, 1982, J. Kim, 1979. Pero ver Cynthia and Graham MacDonald, 1986 para una contraargumentación y una defensa de la consistencia del monismo anómalo.

14 El propio D. Davidson parece favorecer últimamente una lectura de este tipo, o incluso yo diría que más radical. La concepción de la adscripción de actitudes como una especie de métrica acerca de los estados mentales de los demás, en un sentido similar al de P. Churchland, 1979, a quien se remite expresamente, es lo que propone en su 1989.

15 Aunque para tratar esta concepción me centraré en J. Fodor, por ser, desde mi punto de vista, quien más la ha desarrollado, teóricos de la acción causalistas tan diversos como puedan ser A.I. Goldman (1970): o R. Tuomela, (1977), se han remitido expresamente al marco funcionalista computacional para dar cabida a la idea de que las razones pueden funcionar como causas, además de intentar solucionar otros problemas como el de las cadenas causales caprichosas. Así pues, la crítica presentada aquí, debe valer, en principio, también para ellos. La única diferencia es que Fodor ha reducido en los últimos años su enfoque funcionalista a las actitudes, y lo ha eliminado del contenido de las actitudes, con objeto de evitar los problemas del holismo. En este sentido, la crítica a realizar aquí deberá afectar a los autores mencionados por partida doble.

LP con su léxico, su sintáxis y su semántica combinatorias, tener una actitud proposicional, por ejemplo, hallarse en un estado de creencia, consiste en poseer una determinada relación con una representación, donde:

a) la relación es computacional,
b) la representación es una sentencia en el LP del agente,
c) la participación de la representación (o sentencia en el LP) en la relación (es decir, en lo que respecta a su eficacia causal) depende únicamente de sus propiedades sintáctico-estructurales (y no de sus propiedades semánticas o intencionales).
d) la sintaxis postulada para el LP es isomorfa con la de los lenguajes externos, abiertos o públicos, el mismo lenguaje que se utiliza para adscribir actitudes(Cfr. J. Fodor, 1987, cap. 4).

A lo largo de los últimos años se han venido ofreciendo diversos argumentos en contra de la legitimidad de esta propuesta de J. Fodor [16], en el sentido de que, si bien pudiera resultar legítimo, a falta de otra alternativa, postular la existencia de algún mecanismo innato para dar cuenta de los fenómenos cognitivos, de ahí no se infiere que dicho mecanismo deba poseer la estructura y propiedades que Fodor le atribuye. Pero entrar ahora a evaluar la plausibilidad de la hipótesis del LP en general desbordaría ampliamente los objetivos de este escrito. Me limitaré, por tanto a cotejarla estrictamente en cuanto a su viabilidad como marco teórico para acoger las explicaciones de acción mediante la adscripción de contenido, dadas las condiciones propuestas anteriormente. No obstante, al final se verá cómo los problemas en que se ve envuelta ponen en cuestión la propia hipótesis. La razón es aparente, puesto que la eficacia del modelo Creencias-Deseos es, junto con el argumento para la mejor explicación, la principal razón aducida por Fodor en defensa de la hipótesis del LP.

En la sección anterior hemos visto cómo el carácter holista de los estados mentales constituye el problema fundamental en el marco causalista de D. Davidson, y quizá ello explique la escasa simpatía que siempre ha mostrado este autor hacia el funcionalismo. En Fodor, sin embargo, la sombra del holismo emerge precisamente por esta última vía: el carácter funcional de los procesos cognitivos. Evidentemente, mientras que la naturaleza funcional de los estados mentales (pensamientos) favorece la aplicación del enfoque computacional y, por consiguiente, de un marco teórico-causal plausible a primera vista, sin embargo conduce inexorablemente a una caracterización holista del contenido de los estados mentales: poseeer una creencia con determinado contenido —creer que p—, consiste en, o es dependiente de, poseer las relaciones adecuadas con los correspondientes *inputs, outputs* **y otros estados mentales**. Pero si esto es así, entonces la identificación de los estados mentales, o si se quiere, la individuación de las causas de la conducta, se torna harto problemática, puesto que el contenido de cualquier actitud elemental se encontraría, por decirlo así, desparramado a lo largo de

16 Véase, por ejemplo, D. Dennett, 1987; A. Clark, 1988; Braddon-Mitchell & Fitzpatrick, 1990.

un complejo entramado de estados mentales, lo que resultaría incompatible con el funcionamiento discreto y secuencial de las arquitecturas computacionales clásicas y, por consiguiente, con el propio enfoque computacional que constituye la esencia del marco teórico que se pretendía proponer.

Perfectamente consciente del fatal efecto que el holismo podría ocasionar a su sistema, Fodor opta por considerar la tesis holista como una especie de trampa que los escrupulosos filósofos colocan con el fin de entorpecer el desarrollo de una psicología verdaderamente científica, en lugar de tratar de buscar un marco teórico capaz de acoger al mismo tiempo el carácter holista de las actitudes junto con el carácter causal de las explicaciones de acción. El problema radicará entonces en intentar, una vez desechada la caracterización funcional de los estados mentales [17], una explicación alternativa para dar cuenta del significado mental y posibilitar, de ese modo, la identificación del contenido de las actitudes. Fodor la encuentra en el baul de los recuerdos más cercanos de la historia de la semántica: las teorías clásicas o semántica **denotacional**. Según Fodor, los conceptos (mentales o intencionales) son individuados por las propiedades que expresan o denotan, y los pensamientos por los estados de cosas a los que corresponden. Y ello, en virtud de las relaciones causales que vinculan propiedades y estados de cosas con los conceptos y pensamientos respectivos (Cfr. J. Fodor, 1987, Cap. 3). Podría parecer irónica esta vuelta hacia atrás de más de medio siglo, después de todo lo ocurrido en filosofía con las conocidas y asumidas refutaciones a la tesis de la neutralidad teórica de la observación y las no menos demoledoras críticas a los supuestos de la semántica clásica efectuadas por S. Kripke y H. Putnam. En realidad, el panorama dibujado por Fodor guarda una casi estricta similaridad con el que estábamos acostumbrados a encontrarnos en los textos más ortodoxos de los autores de la Posición Heredada, incluida su forma de dar cuenta de los términos teóricos [18], pero quizá sea éste el precio a pagar por una teoría que, de ser viable, podría responder adecuadamente a la cuestión con la que hemos comenzado. Veámoslo.

Si J. Elster era pesimista con respecto a la posibilidad de acceder a los mecanismos de la causalidad psíquica de los sujetos, y, por consiguiente, con respecto a las posibilidades de explicación causal de la acción, J. Fodor parece, por el contrario, muy optimista. Lo que nos ofrece es una «Teoría Causal del Contenido Refinada» (*Slightly Less Crude Causal Theory of Content*), según la cual, «una condición suficiente para

---

17 En realidad, lo que Fodor propone, es eliminar la caracterización funcional del **contenido** de las actitudes, no de las actitudes mismas. Es decir, el hecho de que un estado mental sea de creencia, o de deseo, etc... dependerá funcionalmente de otras actitudes, mientras que las proposiciones, sentencias o «pensamientos» objeto de tales actitudes no.

18 Puede ilustrar esta idea el siguiente párrafo de J. Fodor: «A consequence of the present view is that although theoríes mediate symbol/world connections, still Meaning Holism is not thereby implied. That's because the content of a theory does not determine the meanings of the terms whose connections to the world the theory mediates. What determines their meanings is which things in the world the theory connects them to. The unit of meaning is not the theory; it's the world/symbol correlation *however mediated*». (J. Fodor, 1987, p. 125).Creo que sobran los comentarios.

que los '**A**' (casos -*tokens*- de contenidos mentales) expresen **A** (propiedades o estados de cosas), es que sea nomológicamente necesario que: (1) toda instancia de **A** cause casos de '**A**' cuando, (i) los **A** son responsables causalmente de las marcas o trazos psicofísicos con respecto a los cuales, (ii) el organismo se encuentra en una relación psicofísica óptima; y (2) si los **no-A** causan '**A**', tal hecho es dependiente asimétricamente de que los **A** causen '**A**'»(Cfr. J. Fodor, 1987, p. 126). Según Fodor, esta teoría debería despejar automáticamente todas las dudas de los escépticos acerca de las posibilidades de naturalizar la intencionalidad y la razón, desde el momento en que proporciona una condición suficiente para que una parte del mundo (las representaciones o estados mentales) se relacione semánticamente con otra (expresen propiedades y estados de cosas) y lo hace sin recurrir a una terminología intencional, ni semántica, ni teleológica, y sin incurrir en circularidad.

A tenor de lo que acabamos de decir, la condición (i) se cumple en caso de que la hipótesis del LP sea correcta, ya que esta ofrece un marco conceptual en el que las actitudes proposicionales pueden poseer eficacia causal. Asimismo se podría aceptar, bajo el mismo supuesto que, en un grado bastante aproximado, las actitudes, ***qua* razones**, pueden causar la conducta —condición (ii)—, en la medida en que las propiedades sintácticas de los estados psicofísicos en que consisten las actitudes son las responsables de su procesamiento causal. La condición (iii), sin embargo, ofrece más problemas, los cuales, en mi opinión, podrán en cuestión la propia plausibilidad del marco fodoriano para cumplir (ii). Acabamos de decir que Fodor rechaza la tesis holista con respecto al contenido de las actitudes. Brevemente, su argumentación consiste en una reducción al absurdo de una determinada versión de la tesis del *Meaning Holism* que podríamos calificar como perdedora de antemano. Según esta versión, el contenido de cualquier actitud de un agente, una actitud de creencia, por ejemplo, vendría determinado por el conjunto completo de sus vínculos epistémicos —«*epistemic liaisons*» es la expresión escogida, no casualmente, por J. Fodor—. A su vez, la noción de vínculo epistémico es definida de la siguiente manera: «Cuando un sistema intencional considera el valor semántico de **P** como relevante para la evaluación semántica de **Q**, diremos que **P** es una *epistemic liaison* de **Q** (para dicho sistema en un tiempo determinado» (J. Fodor, 1987, p. 56). La noción de *epistemic liaison* no será, por consiguiente, una noción epistemológica, sino **psicológica**, en la que no podrán contar las dependencias objetivas entre las proposiciones, sino sólo aquellas supuestas por el agente en cuestión. Supongamos que Vd. no cree, pero yo sí, que Jerry A. Fodor es un gran experto en temas operísticos, de modo que para mí, pero no para Vd., la opinión de Jerry A. Fodor es relevante para la evaluación semántica de la proposición «Pavarotti es mejor que P. Domingo». Tendremos entonces que nuestros respectivos vínculos epistémicos serán diferentes. Ahora bien, si según la tesis del *Meaning Holism* la identidad de cualquier creencia depende de la totalidad de sus vínculos epistémicos, entonces, dados los supuestos anteriores, la creencia anterior es una creencia que Vd. y yo no podemos compartir. Por otra parte, podríamos añadir, dado que cualquier creencia de un agente deberá poseer vinculos epistémicos con otras

creencias, y éstas con otras, y así sucesivamente, habría que admitir que el contenido de esa tal creencia está determinado por el conjunto completo de sus creencias. Como parece un absurdo en sí mismo pensar que pudieran existir en el mundo un par de sujetos con exactamente el mismo **conjunto completo** de creencias, entonces nadie en el mundo podrá compartir una sóla creencia. Finalmente, dado que todo criterio de identidad para creencias debería ser capaz de establecer cuándo dos creencias cualesquiera son la misma creencia o no, entonces no habrá criterio de identidad para creencias. Pero si no hay criterio alguno de identidad para creencias, entonces tampoco será posible hacer explicación alguna en psicología, en la medida en que toda explicación debe estar fundamentada en algún tipo de generalización acerca de las acciones a llevar a cabo por los sujetos que poseen las mismas actitudes. Por otra parte, según hemos asumido, las explicaciones basadas en la adscripción de actitudes funcionan. Conclusión: la tesis del *Meaning Holism* debe estar equivocada. Hasta aquí el argumento de Fodor.

Señalaré tres problemas en la argumentación anterior especialmente relevantes, desde mi punto de vista, para el tema que nos ocupa. El primero, y quizá el más importante, tiene que ver con el alcance del ataque de Fodor a la tesis del *Meaning Holism*. Acabamos de ver cómo la crítica de Fodor va dirigida a reducir al absurdo una determinada versión de la tesis definida previamente por él mismo —la que sostiene que el contenido de una actitud viene determinado por sus *epistemic liaisons*— y, naturalmente, existen otras versiones. Concedamos, sin embargo, por causa del argumento, que su crítica hace blanco. ¿Qué es lo que demostraría?. Pues ni más ni menos eso, que el contenido de las actitudes, o si se quiere, la base de datos que llena la estructura del Lenguaje del Pensamiento en las mentes/cerebros de las personas no puede poseer un carácter holista. Pero en ningún momento demuestra que la **adscripción de contenido** no posea intrínsecamente dicho carácter. Tan absurda como la situación diseñada por Fodor sería imaginar que podemos atribuir a un agente la creencia de que «Pavarotti es mejor que P. Domingo» para explicar, por ejemplo, su elección de una determinada versión en una tienda de discos, sin adscribirle al mismo tiempo un cúmulo de creencias difícilmente definible acerca de la música, la ópera, qué es un tenor, criterios acerca de cómo se dan bien los tonos agudos y los graves, etc... Después de todo, para Fodor es perfectamente concebible la idea de que alguien posea como todo y único contenido mental el pensamiento «Pavarotti es mejor que P. Domingo», sin tener idea alguna acerca de qué es un cantante, o un tenor, o la ópera, etc..., del mismo modo que asegura que la idea de que alguien posea el pensamiento «el tres es un número primo» sin poseer ningún otro pensamiento, por ejemplo, acerca de que es ser múltiplo o divisor, no le parece una idea autoevidentemente falsa (Cfr. J. Fodor, 1987, p. 89) [19]. De

19 Resulta un tanto complicado entender cabalmente lo que Fodor quiere decir en este punto. En caso de pretender que el argumento tuviese una aplicación general, habría que admitir que si dichas personas son capaces de aprender en el futuro qué es la ópera, o qué es ser cantante, o las nociones de múltiplo y divisor, entonces, por definción de la propia hipótesis del LP, ya deberían poseer dichos conceptos en la base de datos de su LP, puesto que de otro modo no podrían adquirir dichos conceptos ya

hecho, Fodor reconoce que los argumentos de autores como St. Stich (1983) demuestran que la adscripción de contenido es holista, pero denuncia, al propio tiempo, como ilegítimo, el paso que con frecuencia se da del holismo de la adscripción al holismo del pensamiento (Cfr. J. Fodor, 1987, not. 1, cap. 3). Quizá pudiera tener razón al denunciar esta maniobra, pero nos encontraríamos entonces con un panorama un tanto extraño: el atomismo semántico vendría a ser un hecho profundo, ontológico, de nuestra vida mental, pero al mismo tiempo incognoscible si se admite el carácter holista de la adscripción. Uno se sorprendería de que la ciencia cognitiva pudiera avanzar bajo tales supuestos, pero, en todo caso, **la explicación de acción presupone, por definición, la adscripción de actitudes**, si la adscripción es holista, entonces falla la identificación o individuación de las actitudes y, por consiguiente, siempre bajo los propios supuestos teóricos de Fodor, no se cumpliría la condición (iii).

El segundo problema está relacionado con la propia legitimidad de la hipótesis del LP. Naturalmente, aunque asumiéramos de nuevo que la refutación de la tesis del *Meaning Holism* es correcta, de ahí no se sigue la corrección de la hipótesis del LP, tal y como Fodor la propone, ni tampoco la justificación del carácter causal-denotacional de los contenidos mentales, a no ser que se aporten argumentos independientes. Personalmente intuyo que es al revés, es decir, que Fodor refuta el holismo y propone su atomismo semántico porque previamente ha dado por supuesta la corrección de la hipótesis del LP y no renuncia a un enfoque estricto de la causalidad. Si esta intuición resulta plausible, entonces deberíamos concluír que, en cierto modo, Fodor pone la carreta delante de los bueyes. El argumento explícito esgrimido por Fodor en favor de que la hipótesis del LP posea las características anteriormente descritas (véase p. 91) es el de su supuesta capacidad para dar cuenta de la productividad y sistematicidad de nuestras capacidades cognitivas (Cfr. J. Fodor & Z. Pylyshyn, 1988). Pero primero hay que ver con qué elementos se cuenta para construir la hipótesis, y lo cierto es que el único material a mano es precisamente la interpretación de la conducta abierta de las

---

que, también por definición, dentro de la hipótesis del LP no se puede hablar de aprendizaje en sentido estricto, sino sólo de **traducción** del LP a un lenguaje externo. Siendo ésto así, lo más razonable sería pensar que, para activar dichos conceptos, es necesario que éstos pasen a formar el contenido de una actitud determinada —de creencia, deseo, temor, etc..— ya que según Fodor, las actitudes sí poseen un carácter funcional, es decir, están vinculadas a sus inputs, outputs y otros estados mentales. Pero si ésta fuera la interpretación correcta, entonces parece trivial, puesto que absolutamente todos los conceptos de la base de datos del LP estarían en la misma situación. Lo que está en juego aquí es, por utilizar el mismo ejemplo que Fodor, si es concebible que alguien pueda poseer el concepto de «padre» sin poseer al mismo tiempo los conceptos de «progenitor», «varón» y otros varios pertenecientes a su red semántica (obsérvese que este ejemplo está en la misma situación que los anteriores). Una solución propuesta por el mismo Fodor, cito textualmente, es la siguiente: «What's needed to avoid this embarrassment is a distinction between having a concept 'free' and having it 'bound'»(cfr. Fodor, 1987, not. 8, cap. 3). Se supone que todos los conceptos de la base de datos del LP se encuentran, en principio, 'libres', mientras no hayan tenido ocasión de ser 'ligados' al haber sido sometidos a los estímulos adecuados. Lo que no se alcanza a ver fácilmente es cómo Fodor pudiera evitar un holismo, siquiera moderado, una vez que habla de conceptos ligados.

personas (de sus acciones) por medio de la adscripción de actitudes en un lenguaje externo o abierto. Uno puede tomar ese o esos lenguajes externos, postular una sintaxis combinatoria de la que la interpretación de ese lenguaje puede ser modelo, y pasar seguidamente a pensar que dicha sintaxis instancia ontológica y causalmente los procesos cognitivos correspondientes a las adscripciones de actitudes que hacemos a las personas, pero, en realidad, no habrá aportado ningún argumento independiente para afirmar que la sintaxis postulada es la única correcta, la estructura real responsable de nuestros procesos cognitivos. Fodor y Pylyshyn dan la impresión de sufrir esta ilusión, pero no aportan ningún argumento independiente para apoyar el tipo de sintaxis que defienden, salvo el de que es «*the only game in the town*», pero no exploran la posibilidad de que puedan existir otros tipos de estructuras sintácticas igualmente compatibles, en principio, con la interpretación de la conducta y el lenguaje abiertos habiendo ya candidatos como el conexionismo [20]. Lo curioso es que es precisamente una estructura de este tipo la que resulta incompatible con la admisión del carácter holista de los contenidos mentales, y les conduce, por consiguiente, al intento de refutación de esta tesis y a proponer su alternativa, la semántica denotacional, puesto que encaja mucho mejor con el carácter composicional del LP que se deriva de su hipótesis.

Al hilo del problema anterior, el tercero afecta a la propia viabilidad del programa psico-semántico de Fodor. Según viene a afirmar, su «Teoría Causal del Contenido Refinada» (SLCCTC) es capaz de ofrecer un marco causal que da cuenta del contenido de los estados mentales sin incurrir en circularidad ni utilizar una terminología teleológica, ni semántica, ni intencional (ver p. 19). Sin embargo, como acabamos de ver, su procedimiento para extraer la sintaxis del LP a partir de la semántica de los lenguajes externos usados para adscribir actitudes, y sin otros apoyos independientes para dicha sintaxis, ofrece claros síntomas de circularidad. Por otra parte, el recurso a las **condiciones óptimas** y a sistemas intencionales (o agentes) **intactos**, también ofrece dudas acerca de si su enfoque está libre de supuestos teleológicos. Fodor sale al paso de esta posible objeción afirmando que, después de todo, el recurso a organismos intactos y a condiciones óptimas es común, por ejemplo, en toda teorización biológica (Cfr. J. Fodor, 1987, p. 127). El problema, sin embargo, es si resulta posible especificar condiciones óptimas para sistemas intencionales sin recurrir a terminología semántica o intencional. Según Fodor, toda instancia de **vaca** debe activar causalmente el símbolo o trazo psicofísico del LP '**vaca**' en condiciones óptimas. Supongamos que un día, a la puesta del sol, Vd. se encuentra en las cercanías de un prado y un ejemplar de **caballo** activa en su LP el símbolo '**vaca**', entonces, nos dice Fodor, es que han fallado las condiciones epistémicas (falta de luz adecuada) y, en todo caso, tal hecho depende asimétricamente de que, en condiciones óptimas, los ejemplares de **vaca** siempre cau-

---

20 En realidad Fodor & Pylyshyn, 1988 es un largo artículo donde aparentemente se comparan las virtudes y defectos del conexionismo frente a la arquitectura clásica, pero el problema es que ellos no ven la arquitectura conexionista como alternativa a la clásica, sino en todo caso, como una posible implementación de esta última.

san '**vaca**'. Uno no termina de despejar sus dudas acerca de si este tratamiento está libre de circularidad, puesto que, evidentemente, en condiciones óptimas no cabe el error. Los problemas se acrecientan cuando, por ejemplo, alguien falla en identificar un electrón mirando una cámara de niebla, con luz adecuada y todo lo que queramos. El punto, podría decirnos Fodor, es que para que un ejemplar de **electrón** active el símbolo '**electrón**' en el LP de un agente, es necesario que se den las condiciones epistémicas adecuadas, y entre ellas están, naturalmente, la posesión o no de ciertos conocimientos de física, o si se quiere, en terminología fodoriana, tener activados otros estados psicofísicos, o tener «ligados» un cúmulo de otros conceptos. Después de todo, admite la mediación teórica, aunque al final es la correlación símbolo/mundo lo que cuenta (ver nota 18). No voy a entrar de nuevo en esta discusión, sin embargo, en cuanto al tema que nos afecta, el problema que se plantea aquí es que si la posesión o no de otras creencias, o estados mentales o como les queramos llamar, cuenta a la hora de especificar las condiciones epistémicas óptimas, entonces el reto de Fodor está en poder hacerlo sin recurrir a terminología intencional y obviando las consecuencias holistas que se ven venir, y no lo ha hecho ni tampoco se perciben caminos transitables para escapar de este atrolladero [21]. Después de todo, dentro de este marco teórico difícilmente se puede cumplir la condición (b) de Elster, en la medida en que no es capaz de dar cuenta, de un modo estrictamente causal, de la formación de creencias.

Por si esto fuera poco, los problemas para dar cuenta de la individuación de conceptos se incrementan cualitativamente cuando pasamos a la individuación de acciones. Según hemos visto, Fodor restaura con su psicosemántica el marco de la semántica clásica con su tradicional dicotomía entre sentido y referencia, donde los también tradicionales vínculos entre estas dos últimas nociones se solventan en base a las relaciones causales entre el mundo y los símbolos mentales. Naturalmente, para poder hacer ésto sin perjuicios para su enfoque reductivo y naturalizador del contenido intencional, debe hacer frente a las objeciones familiares planteados por los casos de las tierras gemelas, etc... La razón es obvia, puesto que su enfoque reductivo debe garantizar, de un lado, la identidad de los estados psicofísicos, es decir, que el estado cerebral de los agentes difiere cuando su estado mental difiere, y de otro que, dada la dirección de la fecha causal en su teoría causal del contenido refinada, a saber, del mundo a la mente, debe garantizar que la extensión restringe el contenido. La solución ofrecida por Fodor a la brecha abierta entre sentido y referencia por los ejemplos de las tierras gemelas no es nueva [22], el contenido entendido en sentido estricto, contenido mental o *Narrow Content*, depende del estado causal del organismo y de su relación causal con el entorno, mientras que el contenido entendido en sentido amplio o *Broad Content* se obtiene como una función del *narrow content* relativizada a un contexto.

---

21 En cuanto a las dificultades en general del enfoque de Fodor para escapar del holismo y del uso de terminología semántica o intencional, pueden verse Brian McLaughlin (1987): y Barry Loewer (1987).

22 De hecho ya había sido introducida por otros autores como D. Dennett años atrás. Véase, por ejemplo, su 1982.

Así es posible garantizar que yo y mi gemelo, hallándonos en un estado psicofísico con exactamente el mismo *narrow content* de 'agua'(en nuestros respectivos LPs, estemos en relación con diferentes extensiones, H_2O y XYZ respectivamente. Así pues, para Fodor es el contexto el encargado de 'anclar' el contenido, de modo que, una vez fijado un contexto, la identidad de extensión garantiza la identidad de intensión.

El problema con esta solución es que, si bien parece funcionar aceptablemente para la individuación de los conceptos de clase natural, no se ve cómo pudiera hacerlo en el caso de la individuación de acciones. La forma usual (y prácticamente inevitable) de adscribir actitudes a las personas es identificar (interpretar?) sus acciones. Unas veces adjudicamos actitudes proposicionales a las personas *ex post facto,* interpretada ya la acción (en realidad habría que decir que la propia adscripción de actitudes forma parte de la interpretación), y otras contamos con algún tipo de acceso previo a las actitudes de los agentes, y así podemos prever, mas o menos, sus acciones. El punto es que el proceso de adscripción de actitudes, requisito necesario para poder mantener una simple conversación, es un proceso esencialmente comunicativo. Según acabamos de ver en el esquema de Fodor, el poder fijar los contextos resulta esencial para la identificación del contenido, pero por lo que atañe a los procesos comunicativos no necesitamos trasladarnos a la tierra gemela para modificar el contexto, es algo que ocurre con asiduidad incluso dentro de un mismo proceso. La frecuencia de indéxicos, el hecho de que exactamente las mismas aparentes acciones obedezcan a diferentes propósitos y, por consiguiente, sean distintas y las actitudes que las causan también, de unos contextos a otros, así lo atestigua. Así pues, nos encontramos con que mientras Fodor no ofrezca una solución al problema de fijar los contextos en las situaciones de comunicación y acción, no habrá posibilidad de individuar las actitudes, y por consiguiente, tampoco será posible establecer las generalizaciones necesarias para que se pueda hablar propiamente de explicación de acción [23].

---

23 Según acabamos de ver, Fodor necesita del narrow content, o de una noción de significado internalista para dar sentido a su problema de la individuación de actitudes en-el-cerebro que le permita a su vez introducir generalizaciones causales. D. Dennett (1988a, p. 387): ha planteado un problema a su solución contextual que me parece pertinente. Supongamos que algún ejemplar de XYZ causa en el LP de un agente terrestre la activación del símbolo 'agua'. En ese caso, y ceteris paribus, no se ve cómo Fodor podría afirmar que tal hecho es asimétricamente dependiente de que sean los ejemplares de H_2O los que causan eso mismo. Así pues, 'agua' en el LP de dicho agente terrestre significa H_2O ó XYZ, y por consiguiente el agente y su doble serían gemelos en cuanto a broad content después de todo. Más importante todavía, esta vindicación del externalismo tiene su correlato en el caso de la individuación de acciones, donde, según se ha dicho, la posibilidad de fijar el contexto, dado el esquema de Fodor, debería ser suficiente para hacer transparente la semántica de las actitudes que forman el plan de acción o inferencia práctica de los agentes. De hecho, ésto es lo que viene sucediendo en los enfoques tradicionales o clásicos de la acción y teoría de planes en Inteligencia Artificial. En estos enfoques los planes son concebidos o como estructuras formales que controlan el desarrollo de las acciones particulares, situadas, o bien como abstracciones de instancias de acciones situadas, sirviendo posteriormente dichas instancias para satisfacer la estructura abstracta en cada ocasión particular. Ahora bien, si este enfoque fuera correcto, entonces una situación en la que estuvieran perfectamente especificados el contexto y las condiciones bajo las cuales los constructos mentales (planes de acción) se realizan como acciones, debiera ser suficiente para tener acceso

LA DOBLE ESTRATEGIA: D. DENNETT

Anteriormente hemos dicho que las dificultades del marco fodoriano para acoger las explicaciones de acción mediante la adscripción de actitudes se podían convertir en dificultades que afectan a la hipótesis misma del LP. Resulta hasta cierto punto irónico constatar que, siendo el éxito del modelo creencias-deseos el principal argumento aducido por Fodor en favor de su hipótesis, sea precisamente en ese campo donde la hipótesis fracasa, pues así como ofrece un marco para individuar conceptos, por muy discutible que sea, no lo ofrece para individuar acciones. Es más, el hecho de asumir la hipótesis del LP, con su **realismo-en-el-cerebro** de un lenguaje isomorfo a los lenguajes públicos, con su sintaxis y su semántica combinatorias, es precisamente lo que le impide aceptar la tesis holista derivada del carácter gradual del aprendizaje conceptual y del carácter esencialmente contextual de las adscripciones de contenido.

Recuérdese que en el marco davidsoniano las actitudes causan la conducta, pero no en tanto que descritas como «razones», es decir, como procesos cerebrales de estructura lingüística, sino bajo alguna descripción física. La razón es que para Davidson sólo hay explicaciones causales cuando existen enunciados nomológicos vinculando los términos de la relación causal. Fodor piensa exactamente lo mismo respecto a las relaciones causales. La diferencia entre ambos es que, mientras Davidson acepta el carácter intrínsecamente holista del lenguaje y de las atribuciones de actitud, y por consiguiente, no puede admitir la formulación de enunciados nomológicos vinculando estados físicos y estados mentales, y éstos con otros estados mentales, Fodor rechaza el holismo para poder formular, coherentemente, enunciados de ese tipo. El problema, no obstante, es que el holismo no desaparece simplemente con cerrar los ojos, ni con derribar una caricatura suya. Quizá sea el momento oportuno de recordar un antiguo escrito de Fodor del año 1974. En él argumentaba en favor de la especificidad de las ciencias sociales y humanas —«*Special Sciences (or the Disyunity of Science as a Working Hypotheses)*» fue su revelador título— aduciendo que las generalizaciones nómicas en

---

a las actitudes del agente. Sin embargo, no parece ser así. L. A. Suchman (1987): ha realizado un sencillo e interesante experimento diseñado bajo los anteriores supuestos. En el experimento una máquina fotocopiadora está dotada de un sistema tutorial para enseñar su propio manejo a los usuarios. Como se podrá notar, se cuenta aquí con un contexto perfectamente fijado y con un sistema de representación de un plan de acción diseñado expresamente que conforma las 'estructuras mentales' de la máquina. Por otra parte, también se supone que los usuarios desean ejecutar con éxito la acción de obtener fotocopias de la máquina. Pues bien, los resultados del experimento mostraron que en el transcurso de la interacción usuario-máquina, unas mismas instrucciones eran interpretadas por los mismos agentes de forma diferente o ambigua, y que la máquina tampoco interpretaba siempre igual las mismas acciones de los usuarios, llegando a situaciones de impasse. Es decir, en un contexto tan restrigido y acotado como el relatado, se producen fallos en la identificación de las «actitudes» del interlocutor. La conclusión más importante de la autora es que la coherencia de las acciones no se puede explicar adecuadamente en base a esquemas cognitivos preconcebidos, y que, por el contrario, la organización de la acción situada (y todas lo son): es una propiedad emergente de las interacciones momento-por-momento entre los agentes, y entre los agentes y el entorno de su acción (Cfr. L. A. Suchman, 1987, p. 179).

estas ciencias deben poseer un carácter **especial** debido a su relación de **múltiple instanciación** con respecto a las relaciones causales de nivel inferior que las soportan. ¿Por qué entonces su posterior empeño en hacer de las leyes psicológicas enunciados causales genuinos bajo una concepción tan estricta de la causalidad? Mi sospecha es que para acreditar el carácter científico de la psicología —sus generalizaciones no tratan de entidades fantasmagóricas en ningún sentido— y para evitar el reduccionismo, cosa que al final se ha visto obligado a hacer por demandas de su propio esquema conceptual: proponer un enfoque reductivo de la intencionalidad que no utilice terminología semántica. Sin embargo, para conseguir esto lo único que necesitaba es mostrar que al nivel de descripción psicológico se obtienen regularidades no detectables a niveles más bajos —fisiológicos—, teniendo en cuenta que su carácter explicativo viene garantizado por el hecho de que las generalizaciones psicológicas funcionan en la medida en que son instanciadas por procesos causales de nivel inferior.

Esto es lo que viene sosteniendo D. Dennett desde hace bastantes años al reconocer, por una parte, con Quine, la indispensabilidad práctica del lenguaje intencional o modelo creencias-deseos en psicología y en la vida cotidiana, y no renunciar, al mismo tiempo, al empeño de estudiar científicamente la mente. Davidson, como hemos visto, adopta el mismo punto de partida, sólo que al reconocer las limitaciones del lenguaje intencional para la formulación de enunciados nomológicos genuínos, se queda en la afirmación del carácter anómalo de lo mental sin ofrecer otras posibilidades. Desde esta perspectiva no hubiera perdido nada reconociendo el carácter instrumental de dicho lenguaje en las explicaciones de acción, como vino haciendo (quizá se le ha achacado de forma exagerada, como él mismo dice) D. Dennett. La ventaja de la concepción instrumentalista del lenguaje intencional es que deja la puerta abierta a otros enfoques y a posibles desarrollos que vayan puenteando el abismo entre el lenguaje psicológico y el físico.

En este nuevo contexto, la pregunta con la que hemos comenzado este escrito es respondida por D. Dennett de la siguiente forma: «¿Por qué funciona la estrategia intencional? Primero, porque somos aproximaciones suficientemente cercanas de un diseño cognitivo óptimo (i.e., racionalidad). Sin embargo, ésto no dice nada acerca de cuáles son los detalles últimos de este diseño, que es una cuestión independiente. No deberíamos saltar a la conclusión de que la maquinaria interna de un sistema intencional y la estrategia que predice su conducta *coinciden* —esto es, no deberíamos concluir que la hipótesis del lenguaje-del-pensamiento es verdadera» (D. Dennett, 1988, p. 497). Alguien podría apreciar ciertos supuestos teleológicos ocultos en la referencia a la proximidad de nuestro diseño cognitivo al ideal de la racionalidad. Pero nótese la diferencia con Fodor. En Dennett las observaciones acerca de la racionalidad (y también acerca del holismo) son afirmaciones sobre los rasgos de **nuestra estrategia predictiva** para con los sistemas intencionales, no acerca de los mecanismos causales de su conducta. En el caso de Fodor, sin embargo, al postular como mecanismo causal, interno, un lenguaje con estructura isomorfa a los lenguajes utilizados para adscribir actitudes (la estrategia intencional), las afirmaciones de corte teleológico se convierten

en aserciones ontológicas acerca de la estructura o arquitectura cognitiva. Así, su alusión a mecanismos intactos y, sobre todo, su referencia a la condición asimétrica de las creencias causadas por estímulos que no constituyen sus extensiones, presupone el supuesto teleológico siguiente: «Nuestros mecanismos cognitivos están diseñados para adquirir las creencias óptimas dada la evidencia E disponible». Ahora bien, si ésto se admite respecto a las relaciones entre evidencia y creencias, debería admitirse igualmente de las relaciones entre actitudes y entre estas y la conducta. Después de todo, cuando sucede que los deseos y las creencias (estados psicofísicos) de un agente no causan la acción adecuada, siempre podríamos aducir que tal hecho es dependiente asimétricamente de que las actitudes causen la acción adecuada. Su implicación teleológica sería la siguiente: «Nuestros mecanismos cognitivos están diseñados para adoptar conductas únicas y óptimas, dadas las creencias y deseos». Presumo que Fodor no vería como autoevidentemente falsa una afirmación como la anterior. Si nos atenemos a su forma de dar cuenta de las actitudes de creencia, no debería parecerle ni más ni menos escandalosa. Con ello se verían cumplidas las condiciones (a) y (b) de Elster. El problema es que al hacerlo, Fodor estaría elevando a categoría de estructura cognitiva real, innata, en-el-cerebro, un diseño de racionalidad **que ni siquiera tiene resuelto la Teoría Normativa de la Decisión**.

En definitiva, la pretensión de obtener una teoría de la acción racional que satisfaga las condiciones (a) y (b) de Elster y las condiciones (i), (ii) y (iii) simultáneamente, no sólo desborda los marcos teóricos de Davidson y Fodor, sino que hay que pensar que constituye un objetivo inalcanzable bajo los propios presupuestos conceptuales de su planteamiento. ¿Qué es lo que se puede hacer? Dice Dennett que las cuestiones acerca de los fenómenos cognitivos demandan dos tipos de respuestas: una conceptual y otra causal. La conceptual caracteriza a la estrategia intencional, y en esta estrategia el mito del agente racional estructura y organiza nuestras atribuciones de actitud. La respuesta causal, sin embargo, debe ser reductiva y ocuparse de los detalles de la maquinaria interna. Estos dos tipos de respuestas encuentran su acomodo, según Dennett, en dos tipos de proyectos diferentes, aunque íntimamente relacionados, para trabajar en ciencia cognitiva. El primero debe perseguir la obtención de una **teoría pura de los sistemas intencionales**, de carácter ideal, holístico y abstracto; el segundo tipo de proyecto, denominado **psicología cognitiva a nivel sub-personal**, debe ocuparse de elaborar una teoría de la realización específica, a micronivel, de los sistemas intencionales cuya conducta viene especificada por la primera clase de teorías (Cfr. D. Dennett, 1981).

Unas observaciones finales a esta propuesta. Primero, la dicotomía de proyectos anterior parece destinar a la teoría pura de los sistemas intencionales a un mero papel instrumental, una teoría que cuantifica sobre entidades inexistentes. Dennett ha rechazado últimamente esta interpretación, alegando que en su opinión las creencias, deseos y demás actitudes son fenómenos perfectamente objetivos, declarándose, por consiguiente, realista acerca de las representaciones. Pero realista «con un grano de sal», tan realista como los físicos lo son respecto a los centros de gravedad o los planos sin fricción (Cfr.

D. Dennett, 1987). Para introducir esta idea, recurre a analogías con los casos de la cinemática y la dinámica, o a la vieja distinción de Reichenbach (1938) para los términos teóricos entre *illata* (entidades teóricas propuestas) y *abstracta* (constructos lógicos vinculados a entidades postulados para el cálculo). Desde esta perspectiva, las actitudes proposicionales serían entidades pertenecientes a la última clase, y tal concepción podría caracterizar, por ejemplo, a la posición expresada por D. Davidson en los últimos tiempos (ver not. 14). No obstante, las analogías propuestas, con ser iluminadoras, me parecen un tanto desafortunadas. En primer lugar, en las teorías científicas ordinarias, *illata* y *abstracta* no suelen ir netamente separados integrando proyectos teóricos diferentes. En segundo lugar, y reconociendo el carácter hasta cierto punto ambigüo de su posición general, cuando Dennett habla, siguiendo a Quine y a Davidson, de la indispensabilidad práctica de la estrategia intencional, da la impresión de que se refiere a algo más que a la indispensabilidad de *una métrica* para predecir el comportamiento de unos sistemas cuyos mecanismos causales desconocemos. Después de todo, resulta difícil evadir la idea de que, con bastante frecuencia, nos hallamos en estados cognitivos consistentes en pensamientos o actitudes cuyo contenido se encuentra explícitamente verbalizado, y utilizamos ese mismo supuesto para encontrar sentido o explicar la conducta de los demás. Esta idea es recogida por el propio Dennett (1987) cuando propone establecer técnicamente la distinción entre '*opinions*' —o estados informados lingüísticamente— y '*beliefs*' —o estados profundos, animales—. Ahora bien, si esta distinción responde a una intuición correcta, entonces habría que pensar que las creencias y deseos son algo más que meras entidades teóricas, abstractas e instrumentales para predecir conducta. De no hacerlo así, no veo cómo Dennett podría sortear la crítica de que, en última instancia, la teoría pura de los sistemas intencionales es una teoría de caja negra.

Mi sospecha es que, en el fondo, Dennett tampoco ha terminado de desterrar una concepción excesivamente restrictiva de la causalidad y de su relación con la explicación. Las dos únicas vías de escape ante esta situación que se me ocurren son, de un lado, admitir la posibilidad de explicaciones científicas genuínas que no se basen en leyes causales estrictas [24], y de otro, proponer alguna noción de emergencia capaz de acomodar la idea de cómo los estados psicológicos informados lingüísticamente (los que Dennett llama *opinions*) son supervinientes de los estados profundos, animales (*beliefs*), y éstos de los estados de máquina o neurofisiológicos. Es cierto que las nociones de emergencia y superviniencia son, hoy por hoy, unas nociones bastante

---

24 Esta propuesta es un tanto diferente de la efectuada por algunos psicólogos sociales (véase, por ejemplo, J. Macnamara et al., 1988): que sostienen que la explicación científica de las acciones humanas no requiere leyes (op. cit. p. 1), no obstante suponer que la psicología puede investigar científicamente la acción y aportar resultados importantes. Así pues, la diferencia estriba en sostener que puede haber explicación psicológica de la acción sin leyes causales, o bien que se debe intentar apoyar la posibilidad de explicación psicológica basada en leyes causales no estrictas.

oscuras [25], pero para dar sentido a lo que actualmente se está haciendo en ciencia cognitiva creo que es suficiente con demostrar que el nivel de descripción intencional captura regularidades (*de facto y de iure*), que se pierden al micronivel de la descripción del funcionamiento causal de la maquinaria interna.

El segundo, y último, grupo de observaciones tiene que ver con el carácter ideal y holístico de la teoría pura de los sistemas intencionales. A nadie se le escapa que los seres humanos nos comportamos, frecuentemente, de forma irracional, y sin embargo también solemos dar sentido a este tipo de acciones. Dennett se ocupa explícitamente de este problema en su *«Making sense of ourselves»* (1981a) en respuesta a St. Stich (1981), quien le critica, en primer lugar, que desde el supuesto de racionalidad ideal con que trabaja la estrategia intencional, no es posible dar cuenta (explicar o predecir) de las confusiones o fallos en el comportamiento racional de los agentes, y en segundo lugar, que Dennett no ofrece ninguna definición específica de racionalidad. Dennett contesta a la última, creo que correctamente, que no es posible ofrecer una definición específica (como consistencia, completud, cierre deductivo, etc..., ni tampoco como aquello de lo que la evolución-nos-ha-dotado) por que no se ha demostrado, ni que la consistencia, etc... sean virtudes cognitivas en cualquier circunstancia, ni que la evolución no produzca errores. En su lugar, ofrece una noción de racionalidad para la estrategia intencional un tanto delicuescente: el concepto pre-teórico de excelencia cognitiva («sea lo que sea lo que esto suponga en circunstancias particulares») (Cfr. Dennett, 1988, p. 498). A la primera objeción contesta que los fallos o desajustes en la racionalidad deben ser explicados desde los niveles (causales) del diseño o físico, pero que, en cualquier caso, para dar sentido a los errores de los agentes, siempre tendremos que echar mano de nuestro sistema holístico de atribución de creencias y deseos.

Bien, aun estando de acuerdo, en términos generales, con las respuestas anteriores, quedan algunos detalles por especificar en lo que respecta a sus consecuencias acerca de los posibles marcos de trabajo en ciencia cognitiva. En primer lugar, la referencia al supuesto de «excelencia cognitiva» con que trabaja la estrategia intencional, relegando la explicación de los fracasos a niveles cognitivos más básicos o causales, no hace justicia, ni a sus propias manifestaciones de realismo respecto a las actitudes, ni tampoco al extraordinario trabajo teórico y empírico que se está desarrollando desde los campos de la teoría descriptiva de la decisión (psicología de las preferencias), la psicología social y la lógica. En todas estas áreas se está realizando un enorme esfuerzo tratanto de hallar regularidades en las desviaciones con respecto a los marcos teóricos ideales de la teoría de la decisión y de la lógica o intentando nuevos formalismos que se ajusten a las peculiaridades de los procesos cognitivos reales. Como correctamente

---

25 A. Clark, 1988a y 1989 ha intentado legitimar la idea de propiedades emergentes (por ejemplo, la propiedad de un coche de tomar bien o mal las curvas): aduciendo que son necesarias en diversas explicaciones (i.e., la explicación de los accidentes): y que, como los estados psicológicos, se encuentran en relación de múltiple instanciación con respecto a diferentes materiales y diseños físicos, de modo que las generalizaciones que permiten no son deducibles a partir de descripciones de nivel inferior.

han puesto de manifiesto diversos autores [26], estos marcos ideales resultan irrealizables en los seres humanos de carne y hueso por razones ya mencionadas: o bien presuponen memorias infinitas, omnisciencia, etc..., o bien, dadas estas restricciones, ser prudente, por ejemplo, puede significar ser racional, y ser racional en este sentido puede implicar no respetar la consistencia, la transitividad , etc...

Por otra parte, el problema del holismo, tanto en la adquisición y manejo de conceptos (ver not. 19, p. 94), como el derivado del carácter esencialmente contextual de la acción (ver not. 23, p. 98), contrasta nuevamente con las declaraciones de realismo de Dennett si sólo sirve para relegar sus peculiaridades a la estrategia intencional (que es quien las detecta) y no se propone alguna arquitectura cognitiva que las soporte. Consideremos el ejemplo tratado anteriormente (ver not. 19), acerca de la posibilidad, admitida por Fodor, de que alguien posea el concepto de «padre» sin poseer al mismo tiempo los conceptos de «progenitor», «varón»... Una situación contrastable empíricamente si, por ejemplo, se somete a un niño a las pruebas correspondientes de modo que se comprueba que sabe usar la palabra «padre», reconoce a su padre, etc.. pero no da señales de conocer los conceptos anteriores. El recurso de Fodor para explicar una situación como ésta, un recurso en mi opinión *ad hoc*, consiste en decir que en este caso el niño sí posee dichos conceptos, pero no tiene acceso a su estructura interna. Dennett propone como punto de vista alternativo que, si uno es holista, entonces es libre para decir que el niño sólo gradualmente se va aproximando a la posesión del concepto adulto de «padre», en la medida en que va adquiriendo las creencias relevantes (Cfr. D. Dennett, 1988a, p. 388). Bien, en un caso como éste, lo razonable es pensar que no estamos hablando del holismo **como una propiedad de nuestro sistema de atribución de creencias, sino como una característica del sistema de adquisición y manejo de creencias de las personas**, por lo que demanda, para su mejor explicación, de algún tipo de arquitectura cognitiva que dé cuenta de su comportamiento.

OBSERVACIONES FINALES

¿Qué conclusiones cabe extraer de esta situación?. En los últimos años han venido apareciendo análisis efectuados por filósofos e investigadores en Inteligencia Artificial donde se ponen de manifiesto las limitaciones del paradigma clásico (*The Symbolic Search Space Paradigm in AI*) respecto al tratamiento de la toma de decisiones y la acción humanas. Fruto de estos análisis han sido la propuesta de diversas dicotomías muy pertinentes para el tema que nos ocupa. Winogard & Flores (1986) sostienen que mientras el enfoque clásico funciona aceptablemente bien para representar el procesamiento de solución de problemas abstractos, formales, cuenta con fuertes limitaciones para modelar el procesamiento de los problemas cotidianos. Hewitt (1985) se acerca

26 A este respecto remito a las obras de Davidson, Elster y Cherniak citadas, Wason & Johnson-Laird, 1972, Nisbett & Ross, 1980, Tversky, 1975, Tversky & Kahneman, 1974 , Kahneman & Tversky, 1982, entre otras muchas.

más al punto que estoy tratando de señalar aquí al distinguir entre dominios artificiales como el ajedrez o la prueba de teoremas, y dominios **abiertos**, que son objeto de comunicación y de restricciones desde fuera. Sobre esta idea, Partridge (1986 y 1987) ha propuesto distinguir los tipos de acciones o toma de decisiones humanas en términos de sensibilidad al contexto **débilmente asociada o fuertemente asociada**, con lo que trata de atender a casos concretos, cotidianos, para los cuales resulta posible especificar configuraciones en gran medida libres de contexto. Por otra parte, existen clases de acción, como el ajedrez, que a primera vista son buenos candidatos a ser representados mediante modelos clásicos, al estar perfectamente definidos los estados y las reglas de transformación, y sin embargo los esfuerzos realizados desde la Inteligencia Artificial no han conseguido el éxito esperado desde el punto de vista de la adecuación a la práctica real de este juego por los humanos. De hecho se ha comprobado que los expertos del ajedrez manejan sólamente un reducido subconjunto del conjunto total de reglas de transformación alternativas para cada estado, efectuando la mayoría de sus movimientos de forma intuitiva y en interacción comunicativa con el rival. Tratando de ofrecer una respuesta a este problema, Dreyfus & Dreyfus (1986) proponen una nueva distinción de clases de acción humana: **dominios intuitivos y no-intuitivos.** Mientras que en los segundos se hace un uso explícito y consciente de reglas, en los primeros no ocurre así.

Un dato relevante es que prácticamente todos estos autores están volviendo la mirada hacia el conexionismo como posible fuente de solución a las limitaciones de los modelos clásicos. En pocas palabras, lo que diferencia a los modelos conexionistas es que están basados en redes (*networks*) de unidades conectadas por vínculos. Las unidades poseen un valor de activación y las conexiones entre unidades también pueden poseer diversos valores que expresan el peso (*weight*) de la conexión. Tanto los valores de las unidades como los de las conexiones pueden variar con el tiempo en función del aprendizaje o del ajuste o desajuste de los *outputs*. De este modo, las representaciones (conceptos y proposiciones) emergen como modelos de activación (*activation patterns*) de conjuntos completos de unidades conectadas entre sí [27]. Hasta aquí lo común de los enfoques conexionistas, pues en cuanto a la noción misma de representación y en cómo se lleva a cabo existen notables diferencias no siempre explicitadas en la literatura al uso. Se cuenta con algunas distinciones —como las de redes simbólicas/subsimbólicas, modelos locales/masivamente distribuidos— que no hacen justicia a la variedad de modelos ofertados, de modo que una taxonomía adecuada de los modelos conexionistas está por hacer. Sin embargo, hay una que me parece especialmente relevante para nuestros propósitos. Se trata de diferenciar entre modelos en los que las unidades simples admiten interpretación semántica y los que no, es decir, entre modelos en los que el *pattern* de activación representa un concepto y sus unidades simples representan

---

27 Dado que no es posible hacer aquí una exposición que haga justicia mínimamente a la complejidad de los enfoques conexionistas, pueden verse con carácter general McClelland & Rumelhart (1986), P. Smolensky (1988): y A. Clark (1989), entre otros muchos.

micro-aspectos y aquellos en los que las representaciones son emergentes de un conjunto de unidades interconectadas, dándose el caso de que cada una por sí sóla resulta ininterpretable [28]. Los primeros son implementaciones en modelos conexionistas de las teorías de *propagación de activación* (*spreading activation*) en redes semánticas [29], y tienen la ventaja de ofrecer una solución sencilla y elegante al problema de la adquisición de creencias y al problema (insoluble en el marco de Fodor) de la identidad de los contenidos mentales para poder establecer generalizaciones psicológicas. ¿Qué es lo que se propagó por todo el mundo el día 20 de julio de 1969?. La creencia de que el hombre había puesto el pie en la luna. Según Dennett (1983), el efecto de la propagación de esta creencia no tuvo por qué ser, ni causal ni sintácticamente, el mismo en ningún par de humanos, y sin embargo, la afirmación de que ninguno de ellos tiene nada en común es una falsedad obvia. En realidad, en este nuevo marco no es necesaria la exigencia de identidad de estados mentales/cerebrales para dar cuenta de la misma conducta. Al contrario, al depender las representaciones de conceptos y actitudes del *pattern* completo de activación y no de las unidades atómicas o simples, que en este caso representan micro-aspectos, resulta perfectamente admisible que dados dos *patterns* de activación para un mismo concepto o actitud no necesiten ser exactamente iguales. En el marco de Fodor uno tiene el concepto de 'padre' o no lo tiene, en este nuevo marco cabe la posibilidad de que dos personas (o la misma en el transcurso del tiempo) posean el concepto de 'padre' sin que su *pattern* posea activados todos y exactamente los mismos micro-aspectos.

Así las cosas, en mi opinión resulta plausible la idea de distinguir tres niveles de arquitectura cognitiva con vistas a situar la explicación psicológica. El nivel-1, el menos básico, correspondería a lo que actualmente se entiende como arquitectura clásica, aunque refinada. Es decir, sería una especie de procesador virtual clásico semejante a lo que *Smolensky* (1988 y 1988a) denomina *Conscious Rule Interpreter*, capturando clases de equivalencia utilizables en procesos computacionales que no tienen por qué darse a niveles más básicos. En este nivel es donde debe situarse en principio la explicación de la acción racional desde la estrategia intencional. Pero debe tenerse en cuenta que el carácter ideal, normativo y holístico corresponden a la estrategia que detecta este nivel (la estrategia intencional), y no a la arquitectura cognitiva misma. Si se producen desajustes entre nuestras expectativas racionales sobre el comportamiento de los agentes y su conducta real, entonces se dará una asimetría con respecto a nuestro sistema de adscripción, pero no se tratará de una asimetría causal interna en el diseño de la arquitectura cognitiva del agente. La razón es que este nivel

---

28 W. Ramsey (1990), propone esta distinción, y sostiene que ninguna de las dos anteriores es capaz de dar cuenta de este aspecto.

29 Un buen panorama de la evolución del trabajo desarrollado en Redes Semánticas puede verse en Rumelhart & Norman (1985). Como ya clásicos en el tema pueden verse Collins & Quillian (1969), Collins&Loftus (1975): y Anderson (1983), por citar algunos, con la particularidad de que éste último es reivindicable expresamente como puente para los trabajos de implementación de redes semánticas (aunque su modelo ACT* se concentra en el aprendizaje y manejo de reglas) en redes conexionistas.

opera sobre clases de equivalencia formadas a partir del nivel inferior, pero no lo agota. En la tarea de configuración de este nivel es donde encuentran su lugar natural los trabajos en Lógica y en Inteligencia Artificial encaminados a formular sistemas **realizables** en humanos, guiados por los resultados de la Psicología y la Teoría Descriptiva de la Decisión.

El nivel-2 correspondería a una arquitectura conexionista del tipo representacional por micro-aspectos mencionado anteriormente. Este nivel sustenta al nivel-1, pero el conjunto de clases de equivalencia que lo conforma no tiene por qué coincidir con él, de modo que si dos agentes son descritos como creyendo que 'P', donde 'P' es el mismo pensamiento (concepto, proposición), sus estados mentales a este nivel no necesitan ser equivalentes en el nivel-2. Es precisamente esta posibilidad la que explica que las respuestas ante las mismas representaciones postuladas al nivel-1 difieran con frecuencia, *ceteris paribus*, de unos agentes a otros.

Por último, el nivel-3, o puramente conexionista, implica un cambio de dimensión con respecto a los anteriores, desde el momento que sus unidades simples no son interpretables. Es decir, no incorpora conocimiento, ni tácito ni explícito, de conceptos o reglas de transformación detectables en los otros niveles, por lo que en este sentido no requiere un Lenguaje del Pensamiento del corte propuesto por Fodor. Esto no quiere decir que no incorpore estructuras innatas de otro tipo que controlen su procesamiento, con lo que, a fin de cuentas, una versión **débil** del innatismo, como parecen demandar los argumentos del Chomsky y Fodor, es perfectamente compatible con este nivel [30]. A él corresponderían los estados profundos, animales (*beliefs*) de los que habla Dennett, así como la conducta intuitiva, etc... no codificada lingüísticamente con respecto a la que los autores mencionados ven tantas dificultades para su procesamiento por un sistema clásico. Por otra parte, este es el nivel ingenieril, de diseño de la estructura causal básica capaz de soportar un sistema cognitivo. Está constituído a partir de, y por consiguiente, se halla restringido por, las posibilidades de nuestro sistema neuronal. Y es subsimbólico, por lo que en él ya no es necesario exigir ningún tipo de identidad para pares de estados aproximados o idénticos de los niveles anteriores, debido a su relación de múltiple instanciación/realización respecto a ellos.

Quedan muchas cuestiones por resolver en este marco teórico, entre ellas, recurrente, la de dotar a los niveles simbólicos de un grado de autonomía suficiente como para garantizar la existencia de explicaciones genuínas en ellos. El motivo es que, como acabamos de ver, este marco asume que buena parte de nuestra conducta tiene lugar sin un procesamiento explícito, en-el-cerebro, de símbolos o reglas de transformación

---

30 W. Ramsey & St. Stich, 1990, han puesto de manifiesto, en mi opinión correctamente, que la supuesta incompatibilidad entre el conexionismo y el innatismo ha sido demasiado exagerada, en la medida en que todos los sistemas conexionistas hasta la fecha propuestos incorporan estructuras o reglas de algún tipo, y en muchos casos, específicas del dominio del lenguaje. Esto no quiere decir, naturalmente, que los sistemas conexionistas sean compatibles con la clase específica de LP propuesta por Chomsky y Fodor, sino sólo con la idea del innatismo en las distintas versiones conceptuales genéricas propuestas por Chomsky: innatismo mínimo, anti-empirismo y racionalismo.

lingüísticas, y en cambio hay otras partes de la conducta a las que no podemos dar sentido sin ese supuesto, aunque de todas formas, incluso este procesamiento simbólico debe estar soportado por el nivel básico, subsimbólico. A este respecto A. Clark (1990, pp. 219-21) ha realizado una propuesta muy sugerente: nuestra conducta real es dependiente de *dos sistemas*, uno básico, el conexionista, que procesa subsimbólicamente; el otro, adicional, creado quizá por la explotación de símbolos externos, simula una máquina clásica capaz de incorporar las estructuras conceptuales y reglas de transformación que especifican las arquitecturas clásicas al nivel de competencia. Con esta propuesta, el autor está sugiriendo una multiplicidad de arquitecturas virtuales interactivas como responsables de la conducta real.

Con este esbozo de arquitectura cognitiva a tres niveles espero haber mostrado, o al menos indicado, un marco conceptual a desarrollar en el que pueden tener cabida las explicaciones de acción mediante la adscripción de actitudes sin caer ni en el mero instrumentalismo respecto a estas entidades, ni tampoco en las inconsistencias a que se ve expuesto Fodor con su propuesta de la hipótesis del LP. Al final, lo que uno no termina de comprender es por qué razón Fodor, que reiteradamente ha manifestado que las características y propiedades del LP encuentran su lugar apropiado en los sistemas modulares, encapsulados y automáticos, argumenta desde el modelo creencias-deseos en favor de su hipótesis. El modelo creencias-deseos es un modelo de explicación de acción, y de explicación racional de acción, y en tal medida su procesamiento debe pertenecer a los sistemas centrales. Pero en lo que respecta al conocimiento de estos sistemas está casi todo por hacer.

* Agradezco al Gobierno Vasco (Proyecto GB 003.230-0004/88) la ayuda prestada para la realización del presente trabajo.

REFERENCIAS

ANDERSON, J. R. (1983): *The Architecture of Cognition*, Cambrigge, Mass.: Harvard University Press.

BRADDON-MITCHELL, D & FITZPATRICK, J. (1990): «Explanation and the Language of Thought», *Synthese* **83**, pp. 2-29.

CHERNIAK, Chr. (1986): *Minimal Rationality*, Bradford, MIT.

CHISHOLM, R. M. (1964): «The Descriptive Element in the Concept of Action», *The Journal of Philosophy* **LXI**, pp. 613-625.

— (1970): «The Structure of Intention», *The Journal of Philosophy* **LXVII**, pp.633-47.

CHURCHLAND, P. (1979): *Scientific Realism and the Plasticity of Mind*, Cambridge, Cambridge University Press.

CLARK, A. (1988): «Thoughts, Sentences and Cognitive Science», *Philosophical Psychology* vol. **I**, n.º 3, pp. 263-78.

— (1988a): *Critical Notice of 'Psychosemantics'*, *Mind*, vol. **xcvii**, pp. 605-617.
— (1989): *Microcognition: Philosophy, Cognitive Science and Parallel Distributed Processing*, Bradford, MIT.
— (1990): «Connectionism, Competence, and Explanation», *British Journal for the Philosophy of Science* **41**, pp. 195-222.
COLLINS, A. M. & LOFTUS, E. F. (1975): «A Spreading Activation Theory of Semantic Processing», *Psychological Review* **82**, pp. 407-28.
COLLINS, A. M. & QUILLIAN, M. R. (1969): «Retrieval Time from Semantic Memory», *Journal of Verbal Learning and Verbal Behavior* **8**, pp. 240-47.
ELSTER, J. (1985): «The Nature and Scope of Rational-Choice Explanation», E. Lepore & B. P. McLaughlin (eds.), 1985, pp. 60-72.
DANTO, A. (1976): «Action, Knowledge and Representation», M. Brand & D. Walton (eds), *Action Theory, Dordrecht*, D. Reidel, pp. 11-25.
DAVIDSON, D. (1963): «Actions, Reasons and Causes», *The Journal of Philosophy*, **LX**, pp. 685-700. Reimp. en D. Davidson (1980).
— (1970): «Mental Events», L. Foster & J. Swanson (eds.), *Experience and Theory*, London, Duckwort. Reimp. en D. Davidson 1980.
— (1970a): «How is the Weakness of the Will Possible?» Joel Feinberg (ed.), *Moral Concepts*, Oxford Readings in Philosophy. Reimp. en D. Davidson, 1980.
— (1973): «Material Mind», P. Suppes et. al.(eds) *Logic, Methodology and Philosophy of Science, IV*, North-Holland. Reimp. en D. Davidson, 1980.
— (1974): «Psychology as Philosophy», S. C. Brown (ed.), *Philosophy of Psychology*, London. Reimp. en Davidson, 1980).
— (1980): *Essays on Actions and Events*, Oxford, Clarendon Press.
— (1982): «Paradoxes of Irrationality», R. Wolhein & J. Hopkins (eds.), *Philosophical Essays on Freud*, Cambridge Univ. Press.
— (1984): «Rational Animals», *Dialéctica* **36**, pp. 317-27.
— (1985): «Decepcion and Division», E. Lepore & B. P. McLaughlin (eds.), 1985, pp.139-148.
— (1989): «What is Present to the Mind?» Ponencia del autor en la Segunda Conferencia de Sofia *On Consciousness*, Buenos Aires, Agosto, 1989.
DENNETT, D. (1981): «Three Kinds of Intentional Psychology», R. Healey (ed.), *Reduction, Time and Reality*, Cambridge Univ. Press. pp. 37-61. Reimp. en Dennett, 1987.
— (1981a): «Making Sense of Ourselves», *Philosophical Topics* **12**. Reimp. en Dennett, 1987.
— (1982): «Beyond Belief», A. Woodfield (ed.), *Thought and Object. Essays on Intentionality*, London, Clarendon, pp. 1-95. Reimp. en Dennett, 1987.
— (1983): «Styles of Mental representation», *Proceedings of Aristotelian Society*, vol. LXXXIII. Reimp. en Dennett, 1987.
— (1987): *The Intentional Stance*, Bradford, MIT.
— (1988): «Précis of The Intentional Stance», *Behavioral and Brain Sciences* **11**, pp. 495-546.

— (1988a): «Review of Psychosemantics», *The Journal of Philosophy* **LXXXV**, pp. 384-389.
DREYFUS, S. E. & DREYFUS, H. L (1986): *Mind over Machine*, McMillan/Freen Press.
FODOR, J. (1974): «Special Sciences (or the Disyunity of Science as a Working Hypotheses», *Synthese* **28**, pp. 97-115.
— (1975): *The Language of Thought*, N. York, Crowell.
— (1987): *Psychosemantics. The Problem of Meaning in the Philosophy of Mind*, Cambridge, Mass.:MIT.
FODOR, J. & PYLYSHYN, Z. (1988): «Connectionism and Cognitive Architecture», *Cognition* **28**, pp. 3-71.
GOLDMAN, A. I. (1970): *A Theory of Human Action*, N. Jersey Prentice-Hall, Inc., Englewood Cliffs.
HEWITT, C. (1985): «The Challenge of Open Systems», D. Partridge & Y. Wilks (eds.) (1988), *The Foundations of AI: A Sourcebook*, Cambridge: Cambridge Univ. Press.
HONDERICH, T. (1982): «The Argument for Anomalous Monism», *Analysis* **42**, pp. 59-64.
KAHNEMAN, D. & TVERSKY, A. (1982): «Psicología de las preferencias», *Investigación y Ciencia* **66**, pp. 100-106.
KIM, J. (1979): «Causality, Identity, and Supervenience in the Mind-Body Problem», *Midwest Studies in Philosophy IV*, Minneapolis, pp. 31-50.
LePORE, E. & McLAUGHLIN, B. (eds.) (1985): *Actions and Events. Perspectives on the Philosophy of Donald Davidson*, Oxford, Basil Blackwell.
LOEWER, B. (1987): «From Information to Intentionality», *Synthese* **70**, pp. 287-317.
MANNINEN & TUOMELA, R. (eds.) (1976): *Essays on Explanation and Undertanding, Dordrecht*: Reidel.
MACDONALD, C. & MACDONALD, G. (1986): «Mental Causes and Explanation of Action», L. Stevenson, R. Squires & J. Haldane (eds.), *Mind, Causation and Action*, Oxford, Basil Blackwell, pp. 35-48.
MACNAMARA, J., GOVITRIKAR, V. P. & DOAN, B. (1988): «Actions, Laws, and Scientific Psychology», *Cognition* **29**, pp. 1-27.
McCLELLAND, J., RUMELHART, D. & the PDP Research Group (1986): *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Cambridge Mass.:MIT.
MCLAUGHLIN, B. (1987): «What is Wrong With Correlational Semantics», *Synthese* **70**, pp. 271-286.
MINSKY, M. (1975): «A Framework for Representing Knowledge», Wiston, P. (ed), *The Psychology of Computer Vision*, N. York, McGraw-Hill.
NISBETT, R. E. & ROSS, L. D. (1980): *Human Inference: Strategy and Shortcomings*, Englewood Cliffs:Prentice Hall.
PARTRIDGE, D. (1986): *AI: Applications in the Future of Software Engineering*, Ellis Horwood/Wiley, Chichester.
— (1987): «Human Decision Making & The Symbolic Search Space Paradigm in AI», *AI & Society*, vol. I, pp. 103-114.

RAMSEY, W. (1990): «Connectionism and the Representation of Concepts», Ponencia presentada en I Jornada em Ciencia Cognitiva, Campinas, Brasil, agosto 1990.
RAMSEY, W. & STICH, St. (1990): «Connectionism and Three Levels of Nativism», *Synthese* **82**, pp. 177-206.
REICHENBACH, H. (1938): *Experience and Prediction*, Chicago Univ. of Chicago Press.
ROMANOS, G. D. (1983): *Quine and Analytic Philosophy: The Language of Languages*, Cambridge Mass.:MIT Press.
ROTH, P.(1984): «Critical Discussion: On Missing Neurath's Boat: Some Reflections on Recent Quine Literature», *Synthese* **61**, pp. 205-31.
RUMELHART, D. E. & NORMAN, D. A. (1985): «Representation of Knowledge», A. M. Aitkenhead & J. M. Slack, *Issues in Cognitive Modelling*, Hillsdale, N. Jersey: Lawrence Erlbaum, pp. 15-62.
SMOLENSKY, P.(1988): «On the Proper Treatment of Connectionism», *Behavioral and Brain Sciences* **11**, pp. 1-73.
— (1988a): «The Constituent Structure of Connectionist Mental States: A Reply to Fodor and Pylyshyn», *Tech. Rep. CU-CS-394-88*, Dep. of Computer Science & Institute of Cognitive Science, Univ. of Colorado. Publ. en *Southern Journal of Philosophy* **XXVI**, Sup. pp. 137-162.
STICH, St. (1981): «Dennett on Intentional Systems», *Philosophical Topics* **12**, pp. 38-62.
— (1983): *From Folk Psychology to Cognitive Science: The Case Against Belief*, Bradford: MIT.
SUCHMAN, L. (1987): *Plans and Situated Actions. The Problem of Human Machine Communication*, Cambridge: Cambridge Univ. Press.
TUOMELA, R. (1977): *Human Action and Its Explanation*, Dordrecht, D. Reidel.
TVERSKY, A. (1975): «A Critique of Expected Utility: Descriptive and Normative Considerations», *Erkenntnis* **9**, pp. 163-173.
TVERSKY, A. & KAHNEMAN, D. (1974): «Judgement Under Uncertainty: Heuristics and Biases», *Science* **185**, pp. 499-518.
VERMAZEN, B. & HINTIKKA, M. (eds.) (1985): *Essays on Davidson. Actions and Events*, Oxford, Clarendon.
VON WRIGHT, G. H. (1971): *Explanation and Understanding*, Ithaca, N. York: Cornell Univ. Press.
WASON, P. & JOHNSON-LAIRD, P. (1972): *Psychology of Reasoning: Structure and Content*, London, B. T. Batsford.
WINOGARD, T. & FLORES, F. (1986): *Understanding Computers and Cognition*, N. York, Ablex.

Jesús Ezquerro
Dpto. de Lógica y Filosofía de la Ciencia
Facultad de Filosofía y CC. de la Educación
Apartado, 1249
SAN SEBASTIÁN