Por qué la aposterioridad no (basta, según Kripke, ni) basta*

(Why Aposteriority Is Not (Enough according to Kripke, Nor Is) Enough)

Dan LÓPEZ DE SA

Recibido: 05.10.2005
Versión Final: 06.09.2006

BIBLID [0495-4548 (2006) 21: 57; pp. 245-255]

RESUMEN: Es conocido que Kripke argumentó que la ilusión de contingencia en el caso de la conciencia no puede explicarse del modo en que se explica en el resto de casos familiares de enunciados necesarios a posteriori. En un artículo reciente, Pérez Otero (2002) argumenta que hay una explicación alternativa, en términos de mera aposterioridad. Argumento en contra de la corrección exegética y de la verdad de esta tesis.

Descriptores: aposterioridad, consciencia, ilusión de contingencia, verdades necesarias aposteriori, bidimensionalismo, Saul Kripke.

ABSTRACT: *Kripke famously argued that the illusion of contingency cannot be explained away, in the case of consciousness, in the way it is explained away in the rest of familiar cases of necessary aposteriori statements. In a recent paper, Pérez Otero (2002) argues that there is an alternative way of explaining it a way, in terms of mere aposteriority. I argue against the exegetical accuracy and the truth of this contention.*

Key Words: *aposteriority, consciousness, illusion of contingency, necessary aposteriori truths, two-dimensionalism, Saul Kripke.*

En la tercera conferencia de *Naming & Necessity*,[1] Kripke ofrece un argumento en contra de la teoría de la identidad mente-cuerpo. De modo pertinentemente abstracto, y dejando implícitos muchos de los aspectos que, aunque controvertibles y de hecho a menudo controvertidos, no serán problematizados aquí, el argumento puede formularse así:

> Enunciados como 'dolor es excitación de las fibras C,' aunque fuesen verdaderos, parecen contingentes.
>
> Enunciados como 'dolor es excitación de las fibras C,' si fuesen verdaderos, serían necesariamente verdaderos.

* El origen de este trabajo es mi comentario a Manuel Pérez Otero en sesión de la Societat Catalana de Filosofia del Institut d'Estudis Catalans, Barcelona 2002. Agradezco mucho su réplica entonces y discusión previa y posterior, así como las discusiones con otros miembros de LOGOS, particularmente Óscar Cabaco, José A Díez, Manuel García-Carpintero, David Pineda, y Josep L Prades. Gracias también a Esa Díaz-León, José Gil-Férez, y un evaluador anónimo de esta revista. La investigación ha sido financiada por los proyectos BFF2003-08335-C03-03 (MEC) y BFF2002-10164 (ESF), y la beca EX2004-1159 (MEC).

[1] Kripke (1980). Me referiré a este trabajo de ahora en adelante con '*N&N*.'

> Un enunciado que parece contingente es contingente a menos que podamos explicar la ilusión de contingencia.[2]
>
> El modo en que, de acuerdo con Kripke, explicamos la ilusión de contingencia en casos de enunciados necesarios que parecen contingentes como 'agua es H_2O' o 'Héspero es Fósforo' no se aplica a enunciados como 'dolor es excitación de las fibras C.'
>
> No hay modo alternativo de explicar la ilusión de contingencia.
>
> Por tanto, enunciados como 'dolor es excitación de las fibras C' no son verdaderos.

Para cada una de las premisas del argumento, hay detractores que la rechazan. En un artículo reciente en esta revista, Manuel Pérez Otero (2002) se suma a aquellos que rechazan la última premisa.

El único argumento que Kripke parece ofrecer para la misma —de hecho, del único tipo que parece ser posible ofrecer para afirmaciones de este tipo— es, parafraseándolo, que es difícil ver qué tipo de explicación alternativa pudiera haber (*N&N*, p. 100).[3] En la literatura se encuentran intentos de rechazar la premisa que nos ocupa proponiendo precisamente explicaciones alternativas de la ilusión de contingencia. Pérez Otero propone una, con la peculiaridad de que se basa justamente en la confusión entre nociones epistémicas como *aprioridad* y nociones metafísicas como *necesidad* que Kripke mismo ha denunciado en la primera y la segunda conferencias de *N&N*. Puesto muy brevemente, la tesis que Pérez Otero defiende explícitamente, y que según creo está implícita también en las reacciones al argumento de filósofos como Cristopher Hill (1997) y quizás Brian Loar (1997, 2003), es que el hecho de que los enunciados en cuestión sean aposteriori explica, cabalmente, la ilusión de contingencia. Con más detalle, Pérez Otero defiende que:

(1) Kripke mismo, si no en *N&N* al menos sí en su artículo "Identity and Necessity"[4], acepta el hecho de que ciertos enunciados sean aposteriori como explicación de la ilusión de contingencia con respecto a ellos (aunque, claro está, ello genere entonces una tensión con el argumento que estamos considerando).

---

[2] Por *ilusión de contingencia* entiendo el parecer contingente de enunciados que, no obstante, son necesarios (si son verdaderos). Como se verá más adelante, parte de lo que está en juego es precisamente si se logra explicar la ilusión de contingencia en este sentido, por oposición a explicar juicios (incorrectos) acerca de que los enunciados son contingentes. Estoy en deuda en este punto con un evaluador anónimo de esta revista.

[3] Es por tanto inapropiado afirmar, como cita aprobadoramente Pérez Otero (2002, p. 476), que "Kripke en ningún lugar proporciona una defensa del supuesto de que … el paradigma explicativo ilustrado por el ejemplo del calor sea el único paradigma para explicar las apariencias de posibilidad" (Hill 1997, p. 65).

[4] Kripke (1971). Me referiré a este trabajo de ahora en adelante con "I&N."

(2) Con independencia de si la tesis exegética anterior es correcta o no, el hecho de que ciertos enunciados sean aposteriori realmente explica la ilusión de contingencia con respecto a ellos (y así es posible rechazar la última premisa de la reconstrucción del argumento que nos ocupa).

Adicionalmente, Pérez Otero defiende que esta explicación (a la que llama 'E1') no sólo es una explicación *distinta* sino *mejor* que la explicación que Kripke mismo da en la tercera conferencia de la ilusión de contingencia con respecto a ciertos otros enunciados (a la que llama 'E2'), puesto que:

(3) De hecho, la explicación que propone Kripke de la ilusión de contingencia sólo lo es en la medida que añade, de modo explicativamente irrelevante, elementos a la explicación en términos de la aposterioridad de ciertos enunciados.

(4) Y en cualquier caso, la explicación que propone Kripke de la ilusión de contingencia no se aplica a casos como el de 'Cicerón es Tulio,' por oposición a 'Héspero es Fósforo.'

Mi propósito principal en este trabajo es exponer las razones que creo que hay para rechazar (1) y (2); y observar, en segundo término, qué me parece que hay adicionalmente de incorrecto en (3) y (4).

En contra de (1)

Las razones en contra de (1) son de dos tipos, positivas y negativas. En cuanto a las positivas, parece que Kripke es claro acerca de que no acepta algo así; en cuanto a las negativas, parece que las citas que Pérez Otero ofrece para sostener (1) no permiten al cabo defenderla. Vayamos por partes.

La principal de las razones positivas en contra de (1) es, a mi entender, una que el mismo Pérez Otero menciona:

> Kripke no puede pensar en E1 como una explicación plausible en sí misma. De otro modo, no podría —cuando razona contra el monismo psicofísico— hacer las aseveraciones que acabo de citar [correspondientes a la última premisa de mi reconstrucción —DLdS] y por lo tanto no dispondría de una objeción bien fundamentada a la identificación de estados mentales con estados cerebrales. (Pérez Otero 2002, p. 468)

Efectivamente, si Kripke hubiese considerado alguna vez que para explicar la ilusión de contingencia respecto a un cierto enunciado bastara atender al hecho de que sea aposteriori, entonces es difícil ver cómo podría haber albergado la idea de su objeción a la teoría de la identidad mente-cuerpo. Después de todo, claramente todos los defensores de ésta con los que discute Kripke explotaban explícitamente el hecho de que las identificaciones que propugnaban eran aposteriori, —y de ahí concluían (incorrectamente, según Kripke) que eran contingentes. Ahora bien, la discusión de la teoría de la identidad en los términos que nos ocupan es claramente algo que para Kripke informa *toda* la discusión, tanto en *Naming and Necessity* como en "Identity and Necessity", como lo muestran sendas afirmaciones en las introducciones respectivas:

> [Nuestros puntos de vista sobre el nombrar y la necesidad] tienen realmente implicaciones de amplio alcance para otros problemas en filosofía que tradicionalmente podrían haberse creído alejados, como argumentos sobre el problema mente-cuerpo y la llamada tesis de la identidad. (*N&N*, pp. 22-3)

> [S]e ha supuesto que [las identificaciones teóricas] son un ejemplo muy importante dada su conexión con el problema mente-cuerpo. … Muchos filósofos actuales han mantenido que es muy importante para nuestra comprensión teórica del problema mente-cuerpo el que pueda haber enunciados de identidad contingentes con esta forma. ('I&N', pp. 76-7)

Así, por contraposición, parece razonable rechazar que Kripke haya considerado alguna vez que basta para explicar la ilusión de contingencia respecto a un cierto enunciado el atender al hecho de que es aposteriori, y por tanto rechazar (1).[5]

No obstante, Pérez Otero cree encontrar en 'Identity and Necessity' cierta evidencia a favor de (1). Según él, Kripke parece aceptar ('I&N', pp. 89-91) el hecho de que ciertos enunciados sean aposteriori como explicación de la ilusión de contingencia con respecto a ellos. A mi entender, esto no es correcto. El contexto de la discusión es el siguiente:

> Si se acepta la distinción que he hecho [entre aprioridad y necesidad], no se tiene por qué concluir ninguna de ambas cosas [a saber, que puede haber enunciados de identidad que involucran nombres propios que sean contingentemente verdaderos o que todos los enunciados de identidad que involucran nombres propios genuinos verdaderos son cognoscibles apriori.] Uno puede sostener que ciertos enunciados de identidad entre nombres, aunque con frecuencia conocidos aposteriori, y quizás no cognoscibles apriori, son de hecho necesarios, si es que son verdaderos. ('I&N', p. 89)

Y entonces, tras motivar de nuevo que los nombres propios como 'Nixon' son designadores rígidos, se pregunta:

> A parte de la identificación de la necesidad con la aprioridad, ¿qué es lo que ha hecho que la gente piense de otra manera? ('I&N', p. 89)

Esto ciertamente implica que, de acuerdo con Kripke, la distinción entre aprioridad y necesidad —y así el hecho de que ciertos enunciados de identidad verdaderos que involucran nombres (y son por tanto necesarios) sean aposteriori— explica *algo* que involucra de algún modo la contingencia de los mismos.

Por otro lado, tras preguntarse qué puede estar queriendo decir la gente cuando asevera cosas como 'Héspero podría no haber sido Fósforo' responde:

> [P]ueden estar queriendo decir que no sabemos apriori que Héspero es Fósforo. Esto ya lo he concedido. ('I&N', pp. 90-1)

De nuevo, esto ciertamente implica que, de acuerdo con Kripke, la distinción entre aprioridad y necesidad explica *algo* que involucra de algún modo su contingencia.

---

[5] La razón principal en contra de la tesis exegética de Pérez Otero es por tanto no meramente que "introduce cierta tensión en su argumento dualista" (2002, p. 468) sino que no hace siquiera inteligible que Kripke albergara los propósitos generales que explicita. Si esto es así entonces no se me ocurre casi nada argüible a favor de una interpretación tal. Sin embargo, Manuel Pérez Otero me ha hecho saber en discusión que sigue sin considerarla una razón con peso suficiente. No puedo en este punto más que apelar al juicio del lector al respecto y evocar las palabras de Kripke cuando, en otro contexto, declara su "sorpresa al oír una objeción que concede tan poca inteligencia al argumento" (*N&N*, p. 149).

Ninguna de ambas cosas, sin embargo, entra en contradicción con lo que he defendido antes, y en esta medida, no logran constituir ninguna evidencia a favor de (1). La razón es que para que así fuese, Kripke debería estar no sólo aseverando que el hecho de que ciertos enunciados sean aposteriori explica *algo* que involucra de algún modo la contingencia de los mismos, sino que explica precisamente *la ilusión de contingencia* respecto de los mismos. Pero esto *no* es lo que ocurre, porque lo que Kripke asevera es que el hecho de que los enunciados en cuestión sean aposteriori explica por un lado que la gente haya dado en creer que son contingentes, y por el otro, relacionadamente, haya dado en creer que enunciaciones de su contingencia eran verdaderas. Sin embargo la ilusión de contingencia involucrada en el argumento que se ha de explicar, según Kripke, no consiste ni en lo uno ni en lo otro. Tendré ocasión de volver sobre esto justo a continuación.

En contra de (2)

La razón principal en contra de (2) se puede enunciar de un modo muy simple: no puede ser que la ilusión de contingencia con respecto a un cierto enunciado se explique en términos de la aposterioridad del mismo y una confusión acerca de las nociones de *aprioridad* y *necesidad*, porqué la ilusión de contingencia con respecto a un cierto enunciado persiste incluso cuando toda confusión eventual se disipa. Kripke es en *Naming & Necessity* completamente explícito a este respecto (lo que por otro lado confiere justificación ulterior al punto exegético anterior, si es que fuese necesaria):

> Ahora bien, *a pesar de todos los argumentos que di anteriormente a favor de la distinción entre verdad necesaria y verdad apriori*, la noción de verdad necesaria aposteriori puede seguir siendo de cierto modo desconcertante [*puzzling*]. (*N&N*, p. 140, énfasis añadido)[6]

En efecto: hoy en día estamos razonablemente familiarizados con la distinción en cuestión, y somos presumiblemente poco susceptibles de confundir sus términos. Es más, muchos de nosotros estamos de hecho convencidos, seguramente por las razo-

[6] En discusión, Manuel Pérez Otero me ha sugerido que podría interpretarse este pasaje en la línea de "a pesar de tener considerable fuerza todos los argumentos que di, la noción de verdad necesaria aposteriori puede seguir siendo desconcertante," lo que equivaldría aproximadamente a "incluso quien crea que los argumentos que di tienen fuerza considerable puede no quedar del todo convencido por ellos, y continuar rechazando su conclusión y seguir teniendo la ilusión de contingencia." Me parece claramente una interpretación excesivamente forzada, atendiendo sólo al pasaje. Pero en cualquier caso, el contexto del mismo la excluye sin lugar a dudas. Lo que ocurre más exactamente en la discusión que prosigue a las palabras de Kripke es lo siguiente. El objetor potencial razona del siguiente modo: "Entiendo que Héspero podría haber resultado no ser Fósoforo. ¿Qué quieres decir entonces cuando dices que tales eventualidades son imposibles? Si Héspero podría haber resultado no ser Fósoforo, entonces Héspero podría no haber sido Fósoforo" (*N&N*, p. 141). Kripke concede la generalización de este último condicional (*ibid.*) y se compromete por tanto con que no es verdad que Héspero podría haber resultado no ser Fósforo. Pero claramente parece que sí podría, y así tal apariencia resulta ilusoria y debe por tanto ser explicada. Y efectivamente Kripke se pregunta: "¿En qué consiste entonces la intuición de que las cosas en cuestión podrían haber resultado ser de modo distinto?" (*N&N*, p. 142) Y su respuesta es la explicación que conocemos y que explota en contra de la teoría de la identidad mente-cuerpo.

nes de Kripke mismo, de que los siguientes enunciados expresan verdades que, a pesar de no ser cognoscibles apriori, son *necesarias*:

Héspero es Fósforo.

Agua es H_2O.

Ahora bien, estos enunciados *son* aposteriori y así nos siguen pareciendo contingentes: nos sigue *pareciendo* que podrían ser falsos, aunque claro, no lo *creamos*, y de hecho sepamos que *no* pueden ser falsos . La distinción a la que apelo aquí, de aplicación general en el ámbito de las ilusiones, es aquella que hay entre que ciertas cosas *parezcan* de distinto modo a como en realidad son, y que alguien *juzgue* (incorrectamente), quizás sobre esta base, que son de cierto modo, que resulta ser distinto a como realmente son. Ciertamente ambos son hechos distintos, y lo que eventualmente explica uno de ellos puede muy bien no explicar el otro. En particular: la explicación de por qué alguien que se enfrente por primera vez a las flechas de Müller-Lyer juzgue que los segmentos de recta relevantes tienen la misma longitud presumiblemente consiste en que, entre otras cosas, así es como le parece. Pero difícilmente alguien ofrecería este hecho como explicación del hecho de que ambos segmentos de recta le parezcan tener la misma longitud. En este sentido, el carácter aposteriori de algunos enunciados necesarios no puede explicar la ilusión de contingencia con respecto a los mismos porque precisamente es lo que genera el fenómeno que requiere explicación: sería un error concluir que el que ciertos enunciados (necesarios) sean aposteriori, junto con una confusión entre nociones epistémicas y metafísicas, explica la ilusión de contingencia, sobre la base de que explica que haya gente que juzgue que son contingentes. Y éste es un error que, a mi juicio, Kripke *no* comete.

En contra de (3)

¿Cuál es la situación en relación a la explicación kripkeana de la ilusión de contingencia en casos de enunciados necesarios que parecen contingentes como 'agua es H_2O' o 'Héspero es Fósforo' que, de acuerdo con Kripke, no se aplica a enunciados como 'dolor es excitación de las fibras C'? De acuerdo con Pérez Otero, esta explicación

> también nos atribuye una confusión. Al evaluar el estatuto modal de un enunciado *S*, asociamos con *S* —afirma *E2*— otro enunciado, *S'*, que es contingente; y concluimos, a partir de la contingencia real de *S'*, que *S* es contingente. Por lo tanto confundimos *S* con *S'*. O creemos, al menos, que ambos enunciados tienen el mismo estatuto modal. (Pérez Otero 2002, pp. 469-70)

Si esto fuese así, si la explicación kripkeana de la ilusión de contingencia postulara similarmente una confusión de este tipo, entonces la objeción que he estado considerado —que la ilusión de contingencia persiste una vez toda confusión eventual se disipa— se aplicaría también en contra de ella.[7] El condicional es, no obstante, contrafác-

[7] Debo esta consideración de nuevo a Manuel Pérez Otero en discusión. La tesis de que la explicación kripkeana postula asimismo una confusión de ese tipo es la presuposición de la discusión de la segunda mitad de Pérez Otero (2002) en donde argumenta que en definitiva la razón por la que se ha dado en creer, incorrectamente, que ciertos enunciados necesarios eran contingentes involucra confu-

tico. Como ya hemos visto antes, quizás haya confusiones que expliquen por qué se ha dado en juzgar, erróneamente, que ciertos enunciados eran contingentes, cuando no lo son. Pero la ilusión de contingencia es otro fenómeno, que persiste cuando las confusiones se disipan y no puede por tanto explicarse en términos de ellas. A mi entender, la explicación kripkeana consiste, dicho de un modo sumamente abstracto, en señalar aquello que es en efecto contingente y que está apropiadamente relacionado con los enunciados que por tanto, pese a no ser contingentes, lo parecen.

A mi juicio, la elaboración más convincente de esta explicación es la que han proporcionado de modo más reciente filósofos como Frank Jackson (1994) y David Chalmers (1996), que la formulan en el marco bidimensional desarrollado por David Kaplan (1989), Robert Stalnaker (1979) y David Lewis (1980), entre otros, para expresar los dos modos distintos en que la verdad de un enunciado depende de los hechos. Un mismo enunciado puede ser verdadero en un cierto contexto y falso en otro, en virtud de que diferentes hechos afecten a lo que con el enunciado se asevera en uno y otro contexto. Esto es claramente así en el caso de enunciados que contienen expresiones como 'yo' o 'ese hombre:' el enunciado 'yo estoy cansado' puede ser verdadero en un contexto en el que yo lo digo, y falso en otro en el que tú lo dices. La discusión filosófica en las últimas décadas, y particularmente en los últimos años, ha puesto de manifiesto que la clase de expresiones *dependientes del contexto* en este sentido —que pueden contribuir distintamente a lo que se asevera con enunciados que las contienen atendiendo a distintos rasgos de los contextos en los que se podrían decir— es mucho mayor que lo que pueda parecer a primera vista, y de hecho quizás contenga las expresiones filosóficamente más interesantes, como 'saber,' 'poder,' o 'bueno.' Por otro lado, lenguajes como el nuestro contienen expresiones como 'posiblemente' o 'estrictamente hablando' que requieren, para recibir un valor de verdad cuando se dicen en un contexto, que se evalúen los enunciados incrustados (o sus variantes) con respecto a rasgos posibles de contextos que no tiene por qué ser los del contexto en los que se dicen. Por ejemplo 'posiblemente haya gatos de color azul' requiere para ser verdadero que 'hay gatos de color azul' sea verdadero con respecto a un mundo posible, quizás distinto del mundo del contexto en el que se dice.

Así pues, aquello lingüístico convencionalmente asociado a un enunciado tipo determina un valor de verdad relativamente a un *contexto* —una circunstancia en la que el enunciado podría proferirse, que puede ser representado por un mundo posible centrado en un punto espaciotemporal— y un *índice* con respecto al que el enunciado en cuanto que proferido en un contexto se evalúa —que puede ser representado por una tupla de los posibles rasgos de contextos variables por operadores como los considerados, aunque no tiene por qué haber ningún contexto posible cuyos rasgos coincidan con los de un índice dado. (En las discusiones sobre consciencia a veces se simplifica identificando el contexto y el índice con sendos mundos posibles. Entonces se suele

siones entre la noción metafísica de que el enunciado sea contingente y la noción epistémica de que sea cognoscible sólo aposteriori. Como he dicho antes, no tengo nada en contra de esta idea: sencillamente ocurre que explicar por qué alguien ha dado en juzgar incorrectamente sobre la base de que algo parecía contingente que era contingente no es explicar por qué esto parece contingente sin serlo.

hablar de mundo posible *considerado como real* para referirse a mundo posible como, dada la simplificación, representando un contexto, y mundo posible *considerado como contrafáctico* para referirse a mundo posible como, dada la simplificación, constituyendo un índice.) Dado un enunciado y un contexto, quedan por tanto determinados *dos* contenidos distintos o proposiciones. Por un lado, la proposición *horizontal* es aquella que se aseveraría diciendo el enunciado en el contexto —salvo que actúe algún mecanismo conversacional. Por otro lado, la proposición *diagonal* es aquella que para cada contexto recibe el valor que el enunciado dicho en ese contexto recibe cuando se evalúa con respecto al índice de ese contexto.[8]

Supongamos, siguiendo a Kripke, que lo convencionalmente asociado a la expresión 'Héspero,' el material "fijador de la referencia," es algo así como que *es el cuerpo celeste en tal y cual posición en el cielo por la mañana*. Con respecto a un cierto contexto (real), la proposición horizontal del enunciado 'Héspero estaba precioso ayer' es *que Venus —el cuerpo celeste que en el mundo del contexto está en tal y cual posición en el cielo por la mañana— tenía tales y cuales rasgos —que constituyen la preciosidad, dados los estándares estéticos del contexto— el martes —día anterior al del tiempo presente del contexto*. Esta proposición, verdadera de hecho, resulta falsa con respecto a un índice similar al del contexto original salvo porque la coordenada de mundo posible la constituye un mundo contrafáctico en el que ese mismo objeto en ese mismo momento no tiene esos rasgos que constituyen la preciosidad dados esos mismos estándares estéticos, debido por ejemplo a una posible perturbación atmosférica. La proposición horizontal del enunciado en el contexto es por tanto contingente.

Supongamos que consideramos ahora un contexto cuyo mundo es éste que acabamos de considerar, en el que por cierto el cuerpo celeste que está en tal y cual posición en el cielo por la mañana no es Venus sino Marte. La proposición horizontal de ese mismo enunciado con respecto a ese otro contexto *que Marte —el cuerpo celeste que en el mundo del contexto está en tal y cual posición en el cielo por la mañana— tenía tales y cuales rasgos —que constituyen la preciosidad, dados los estándares estéticos del contexto— el martes— día anterior al del tiempo presente del contexto*. Supongamos que, debido a la misma perturbación atmosférica, esta proposición también resulta falsa con respecto al índice de este otro contexto. La proposición diagonal del enunciado en cuestión es por tanto contingente.

La proposición diagonal encapsula el componente epistémico de los enunciados, dado que modela aquello lingüístico convencionalmente asociado al enunciado y es por tanto responsable de la aprioridad o aposterioridad de los mismos. Los enunciados de identidad 'agua es H_2O' o 'Héspero es Fósforo' son necesarios, puesto que sus proposiciones horizontales son en efecto necesarias. La explicación kripkeana de la

[8] *Horizontal* y *vertical* son las denominaciones clásicas, véanse Stalnaker (1979) y Lewis (1980). Chalmers (1996) las llama *secundaria* y *primaria*, respectivamente, y éstas son las denominaciones que adopta Pérez Otero (2002).

ilusión de contingencia con respecto a ellos consiste en mantener que sus proposiciones diagonales son, sin embargo, realmente contingentes.[9]

En contra de (4)

De acuerdo con Pérez Otero, se puede explicar la ilusión de contingencia de ciertos enunciados necesarios en términos de ser aposteriori —según Kripke, y en efecto—, y de hecho la explicación kripkeana descansa en último término en esto. Ya he argumentado por qué son incorrectas estas tres ideas en lo que precede. Me gustaría para acabar considerar la cuestión del alcance de casos a los que la explicación kripkeana se aplica. Según Pérez Otero, la explicación kripkeana

> no se aplica cuando están involucrados nombres no descriptivos. Consideremos por un momento la posibilidad de que —contrariamente a lo que defiende Kripke— todos los designadotes rígidos sean descriptivos. Ni siquiera en ese caso es claro que *E2* [i.e., la explicación kripkeana — DLdS] pueda dar cuenta de todas las verdades necesarias aparentemente contingentes; pues es muy plausible que algunas de esas verdades no contengan ningún designador rígido. (2002, p. 472)

Esta última sugerencia —que hay casos de necesidades aposteriori distintos a los que involucran consciencia en los que la ilusión de contingencia no pueda explicarse kripkeanamente— es ciertamente interesante. De haberlos uno tendría razones para pensar que, aunque no pueda ser la considerada en términos de la aposterioridad, por los motivos que hemos visto, debe haber una explicación alternativa, o en caso contrario habría razones para rechazar las ideas kripkeanas sobre la naturaleza de la modalidad que en la formulación del argumento propuesta está contenida en la tercera premisa, que repito:

> Un enunciado que parece contingente es contingente a menos que podamos explicar la ilusión de contingencia.

Lamentablemente Pérez Otero no menciona ninguno —esto es, ninguno que no sea del tipo de 'Cicerón es Tulio,' que consideraré a continuación—, y los candidatos que conozco me parecen todos menos que convincentes.[10]

La explicación particular de la ilusión de contingencia con respecto a 'Héspero es Fósforo' que propone Kripke explota esencialmente el hecho de que hay material descriptivo conceptualmente asociado (de modo que no da el significado de pero sí fija la referencia de) tanto a 'Héspero' como a 'Fósofo,' que involucran (distintas) propiedades meramente contingentes del planeta al que refieren. Ahora bien, no es claro que

---

[9] En la primera parte de Pérez Otero (2002), éste discute la que llama *ecuación bidimensionalista básica*, que podemos formular en general como aquella de acuerdo con la cual un enunciado es aposteriori si y sólo si su proposición diagonal es contingente, y objeta a la tesis de que dicha ecuación proporcione "un análisis del concepto de conocimiento apriori utilizando nociones modales" (Pérez Otero 2002, p. 463). Lamentablemente no menciona ningún autor que haya mantenido dicha tesis reductiva. Véase la discusión de Chalmers (2006).

[10] Véanse, para los mismos, Block & Stalnaker (1999), Yablo (1999), y Wright (2002).

Kripke crea que esto sea en general el caso con respecto a todos los nombres propios. Aunque a veces habla acerca de 'Cicerón es Tulio' bajo el supuesto de que sí hay semejante material descriptivo asociado a los nombres (del tipo de 'hombre cuyos trabajos se estudiaban en "Latín" de tercero de instituto,' y 'orador romano que denunció a Catalina'), pudiera bien ser que esto fuese sólo una licencia expositiva y que creyese que en realidad no hay tal material con respecto a 'Cicerón' y 'Tulio.' Ahora bien, para que 'Cicerón es Tulio' fuese como 'Héspero es Fósforo' respecto a su carácter necesario aposteriori, debería pasar que 'Cicerón' y 'Tulio' fuesen como 'Héspero' y 'Fósoforo' respecto al hecho de tener conceptualmente asociado material descriptivo del tipo pertinente. Pero entonces parece que ocurre algo como lo siguiente. *O bien* 'Cicerón es Tulio' no es como 'Héspero es Fósforo' respecto a su carácter necesario aposteriori, y en particular es apriori. Razonable o no, yo no he logrado encontrar nada en los textos de Kripke incompatible con esta opción (aunque tampoco he encontrado nada que la sugiera) y de hecho ésta es la línea que siguen algunos de los filósofos "millianos" en este sentido, como Martin Davies & IL Humberstone (1980) o Scott Soames (2003). *O bien* resulta al cabo que el carácter necesario aposteriori de 'Cicerón es Tulio' motiva un tratamiento más "fregeano" de la semántica de los nombres involucrados. Ésta es la línea que siguen los filosófos bidimensionalistas que he mencionado.

Quizás esta disyunción no sea exhaustiva. Después de todo, quizás existe alguna explicación alternativa a la explicación de la ilusión de contingencia con respecto a 'Héspero es Fósforo' que pudiese aplicarse a 'Cicerón es Tulio' compatiblemente con la ausencia de cualquier material descriptivo asociado con los nombres propios involucrados. Es más, quizás ésta explicación alternativa explique también la ilusión de contingencia con respecto a 'Dolor es excitación de las fibras C'. Supongo que hay un sentido de 'quizás' en el que sí, quizás sí. Después de todo, como ya hemos visto, el argumento a favor de la última premisa que da Kripke se basa meramente en que es difícil ver qué tipo de explicación alternativa pudiera haber, y así quedaría sin base en cuanto alguien ofreciese una explicación alternativa tal. Pero, hasta donde yo sé, nadie ha hecho convincentemente hasta la fecha algo así.

REFERENCIAS

Block, N. & R. Stalnaker (1999). "Conceptual Analysis, Dualism, and the Explanatory Gap", *Philosophical Review* 108, 1-46.

Chalmers, D. (1996). *The Conscious Mind*. Oxford: Oxford University Press.

——— (2006). "The Interpretation of Two-Dimensional Semantics", en M. García-Carpintero and J. Macià (eds.), *Two-Dimensional Semantics*. Oxford: Oxford University Press.

Hill, C. (1997). "Imaginability, Conceivability, Possibility and the Mind-Body Problem", *Philosophical Studies* 87, 61-85.

Jackson, F. (1994). "Armchair Metaphysics", en M. Michael and J. Hawthorne (eds.), *Philosophy of Mind*. Dordrecht: Kluwer.

Kaplan, D. (1989). "Demostratives", en J. Almog, J. Perry and H. Westtein (eds.), *Themes from Kaplan*. Oxford: Oxford University Press.

Kripke, S. (1971). "Identity and Necessity", en S.P. Schwartz (ed.), *Naming, Necessity, and Natural Kinds*. Ithaca NY: Cornell University Press.

Kripke, S. (1980). *Naming and Necessity*. Oxford: Blackwell.
Lewis, D. (1980). "Index, Context, and Content", en S. Kanger and S. Öhman (eds.), *Philosophy and Grammar*. Boston: Reidel. Reimpreso en su *Papers in Philosophical Logic*. Cambridge: Cambridge University Press, 1998.
Loar, B. (1997). "Phenomenal States, Second Version", en N. Block, O. Flanagan and G. Guzeldere (eds.), *The Nature of Consicousness*. Cambridge MA: MIT Press.
——— (2003). "Qualia, Properties, Modality", *Philosophical Issues* 13, 113-29.
Pérez Otero, M. (2002). "Aplicaciones filosóficas del bi-dimensionalismo: modalidad y contenido epistémico", *Theoria* 17, 457-77.
Soames, S. (2002). *Beyond Rigidity*. Oxford: Oxford University Press.
Stalnaker, R. (1978). "Assertion", *Syntax and Pragmatics* 9, 315-32. Reimpreso en su *Context and Content*. Oxford: Oxford University Press, 1999.
Yablo, S. (1999). "Concepts and Consciousness", *Philosophy and Phenomenological Research* 59, 455-63.
Wright, C. (2002). "The Conceivability of Naturalism", en T. Gendler and J. Hawthorne (eds.), *Conceivability and Possibility*. Oxford: Oxford University Press. Reimpreso en su *Saving the Differences*. Cambridge MA: Harvard University Press, 2003.

Dan López de Sa es Postdoctoral Research Fellow en Arché (University of St Andrews) desde 2004, y miembro de LOGOS (Universitat de Barcelona) desde 1997. Está particularmente interesado en la teoría de la vaguedad como indecisión semántica, la noción de rigidez para predicados, la relación entre la dependencia de respuesta y el realismo, la naturaleza de los valores, y la caracterización y discusión de las diferentes formas de relativismo. Ha publicado artículos de investigación en *Analysis*, *Dialectica*, *Grazer Philosophische Studien*, *International Journal of Philosophical Studies*, *Proceedings of the Aristotelian Society*, *Teorema*, y *Theoria*.

Dirección: Arché–The AHRC Research Centre for the Philosophy of Logic, Language, Mathematics and Mind, University of St Andrews, 17 College Street, St Andrews KY16 9AL, Scotland. LOGOS–Grup de Recerca en Lògica, Llenguatge i Cognició, Universitat de Barcelona. E-mail: dlds@st-andrews.ac.uk.

Superveniencia, propiedades maximales y teoría de modelos

(Supervenience, Maximal Properties, and Model Theory)

Xabier DE DONATO RODRÍGUEZ y Marek POLANSKI

Recibido: 21.11.2005
Versión Final: 06.05.2006

BIBLID [0495-4548 (2006) 21: 57; pp. 257-276]

RESUMEN: En el presente artículo, se examinan y discuten dos argumentos con consecuencias reduccionistas debidos a Jaegwon Kim y a Theodore Sider respectivamente. De acuerdo con el argumento de Kim, la superveniencia fuerte implicaría la coexistencia necesaria de propiedades (es decir, tal y como normalmente se interpreta, la reducción). De acuerdo con el de Sider, ocurriría lo mismo con la superveniencia global. Uno y otro hacen un uso esencial de sendas nociones de propiedad maximal, las cuales son discutidas aquí a la luz de una interpretación natural e interesante de la teoría de las propiedades implícita en sus argumentos. Bajo esta nueva interpretación, en términos modelo-teóricos (véase apartado 4), obtenemos diversas posibilidades de relaciones formales entre las tesis de superveniencia y la reducción, según la lógica utilizada. Al menos bajo una interpretación interesante, los argumentos de Kim y Sider no son correctos, quedando demostrado así que dichos argumentos no son válidos en general.

Descriptores: superveniencia, propiedades maximales, teoría de modelos, reducción, fisicismo no reductivo

ABSTRACT: *We discuss and analyze two reductive arguments due to Jaegwon Kim and Theodore Sider respectively. According to the first one, strong supervenience would imply necessary coextension of properties (i.e., reduction). According to the second, this would be also the case of global supervenience. Kim and Sider make essential use of their respective notions of maximal properties, which we analyze here in the light of a natural and interesting interpretation of the underlying theory of properties. Under this interpretation, in terms of model theory (see § 4), we obtain different possibilities of formal relations between the superveniencie theses and reduction, depending on the logic we use. Under at least one interesting interpretation, the arguments of Kim and Sider are not correct and we become the conclusion that these arguments are not valid in general.*

Keywords: *supervenience, maximal properties, model theory, reduction, non-reductive physicalism*

1. Introducción

Desde los años ochenta, ha tenido lugar un debate en filosofía de la mente en torno a la adecuación de la superveniencia global y la fuerte (que abreviamos SG y SF respectivamente) como tesis fisicistas no reduccionistas. Algunos argumentos en relación con esta cuestión han sido propuestos con el fin de mostrar (o pueden interpretarse en el sentido de que muestran) que SG y SF son demasiado fuertes para servir a propósitos no reduccionistas, porque bajo supuestos adicionales aceptables, tanto SG como SF tendrían consecuencias reduccionistas. En el presente artículo, discutiremos dos conocidos argumentos en esta dirección, debidos respectivamente a Kim y Sider. Ambos argumentos (y otros semejantes) están basados en lo que podríamos llamar una "teoría naïve de las propiedades" (TNP), la cual está constituida por un cuerpo informal de tesis más o menos caracterizadas (explícita o implícitamente) sobre propiedades y sus extensiones en mundos posibles. Tanto Sider como Kim hacen un uso esencial de sendas nociones de propiedad maximal. Para poder considerar la corrección

(formal y material) de sus argumentos, es necesario disponer de una reconstrucción formal precisa de sus (diferentes) caracterizaciones de propiedad maximal. En consecuencia, comenzaremos ofreciendo una reconstrucción de TNP como base conceptual subyacente y, especialmente, de las nociones de propiedad maximal de Kim y Sider. El hecho de que ambos argumentos revelen una estructura similar nos ha conducido a analizarlos aquí conjuntamente. Como veremos, ambos argumentos son reconstruibles en una forma bajo la cual son formalmente correctos, pero en ambos casos hay involucrada una cuestión crucial y es la de la existencia de propiedades maximales, la cual afecta a la corrección material de los argumentos. Bajo nuestra reconstrucción esta cuestión crucial queda iluminada de un modo que lleva a interesantes resultados, así como a distintas interpretaciones posibles de las relaciones lógicas entre SF y SG, por un lado, y la reducción por otro. De acuerdo con esta interpretación, que resulta totalmente natural, a las propiedades se les puede hacer corresponder fórmulas, mientras que los mundos posibles se pueden interpretar como modelos. El aparato lógico que emplearemos es, por tanto, el de la teoría de modelos y los sistemas considerados serán $\mathcal{L}_{\infty\omega}$, $\mathcal{L}_{\omega\omega}$, y $\mathcal{L}_{\omega_1\omega}$. Obtendremos la conclusión de que, en $\mathcal{L}_{\omega\omega}$ y con estructuras finitas, los argumentos de Kim y Sider (y las tesis que se supone apoyan) no son materialmente correctos(as): para poder ser concluyentes, incluso restringiéndonos a estructuras finitas, ambos argumentos necesitan fórmulas infinitas. Para obtener argumentos materialmente correctos, además de restringirnos a estructuras finitas, basta el uso de fórmulas expresables en $\mathcal{L}_{\omega_1\omega}$. Ni Kim ni Sider hacen ningún supuesto en relación con el número de individuos en un mundo posible, sino que más bien pretenden la validez general de sus argumentos. El uso de $\mathcal{L}_{\infty\omega}$ parece reflejar adecuadamente las intenciones de Kim, pero desafortunadamente algunos pasos de su argumento no pueden ser formalizados en esta lógica. Por otra parte, la invalidez del teorema de definibilidad de Beth en $\mathcal{L}_{\infty\omega}$ nos revela que la tesis de Sider es falsa y que, por tanto, su argumento no puede ser materialmente correcto. Todas estas consideraciones muestran que, bajo una interpretación natural como la aquí presentada, las estrategias reduccionistas de Kim y Sider no resultan concluyentes en general.

2. Teoría naïve de las propiedades

Las distintas tesis de superveniencia están usualmente formuladas de acuerdo con una base conceptual que podemos llamar "teoría naïve de las propiedades" (o TNP). De hecho, tal teoría no se ha formulado con precisión en ninguna parte. Lo que podemos encontrar en la bibliografía sobre el tema es más bien un cierto cuerpo de conceptos y supuestos que, a pesar de carecer de precisión formal, está formulado de manera suficientemente clara como para realizar una reconstrucción sin esenciales correcciones. En lo que sigue vamos a proponer una reconstrucción de los principales elementos de TNP, sus supuestos ontológicos y principios generales, en la medida en que sean relevantes para el estudio de las relaciones entre superveniencia y reducibilidad.

2.1. Reconstrucción de TNP

La ontología de TNP consiste en tres categorías disjuntas de entidades:

- mundos posibles
- individuos y tuplas finitas de individuos
- propiedades unarias y relacionales de individuos

Los individuos son entidades no analizables que existen en mundos posibles. Un individuo puede existir en más de un mundo posible. Las propiedades están instanciadas por individuos (o tuplas de ellos) en los mundos posibles en que estos individuos existen. Necesitamos tres tipos de variables. Usaremos las variables u, v, w para mundos posibles. Para referirnos a individuos y tuplas de ellos usaremos las letras a, b, c (con subíndices si es necesario). Para las propiedades, usaremos P, Q, R. La aserción "*a* tiene la propiedad P en el mundo w" es simbolizada por "$P_w(a)$", que expresa la relación ternaria entre propiedades, mundos y (tuplas de) individuos. Las propiedades tienen sus extensiones en mundos posibles. La extensión de la propiedad P en w es simbolizada mediante "P^w". A cada mundo w asociaremos un conjunto no vacío de individuos $D(w)$ que comprende todos y sólo aquellos individuos que existen en w. Si P es una propiedad n-aria (o relacional), entonces P^w es un subconjunto (posiblemente vacío) del conjunto de todas las n-tuplas de $D(w)$. Dos propiedades son idénticas si sus extensiones en cada mundo son las mismas. Podemos identificar en consecuencia cada propiedad P con una función definida sobre el conjunto de los mundos posibles que asigna a cada mundo w el conjunto P^w. Cada mundo w puede verse como una entidad compleja de la forma $(D(w), P^w, Q^w, ...)$ consistente en el universo de w seguido por la secuencia de extensiones de todas las propiedades que los individuos tienen en ese mundo. De este modo, los criterios de identidad para mundos y propiedades quedan claramente definidos. La propiedad P *implica* la propiedad Q ($P \supset Q$) si para cada mundo w: $P^w \subseteq Q^w$. Claramente, P y Q son idénticas syss $P \supset Q$ y $Q \supset P$. De una propiedad n-aria P se dice que es *consistente* si hay un mundo w tal que $P^w \neq \varnothing$. Ahora podemos definir las operaciones booleanas para propiedades. Para cualesquiera propiedades n-arias P y Q, Q es la negación de P si para cada mundo w, $Q^w = D(w)^n \setminus P^w$. Para cada familia de propiedades n-arias Θ, la extensión de la conjunción de Θ en el mundo w es la intersección del conjunto de las extensiones de las propiedades de la familia Θ en w. Análogamente, definimos la disyunción de un conjunto de propiedades. Asimismo podemos definir operaciones cilíndricas para propiedades que correspondan a los cuantificadores existencial y universal.

2.2. Propiedades maximales

La noción de *propiedad maximal* juega un papel crucial en el debate sobre superveniencia y reducción. Existen varias explicaciones diferentes de esta noción. Aquí reconstruimos dos de ellas: una debida a Jaegwon Kim, la otra a Theodore Sider. Merecen

especial atención, puesto que tanto Kim como Sider hacen un uso esencial de dicho concepto en sus respectivos argumentos, los cuales muestran aparentemente que interesantes tesis de superveniencia, como son SF y SG, tienen, contrariamente a las motivaciones que había detrás de ellas, consecuencias reduccionistas.

2.2.1. Kim sobre maximalidad

Podemos comenzar con las palabras con las que Kim presenta su noción de propiedad maximal (Kim 1993, p. 58-59):

> (…) consider what we may call *B-maximal properties*: these are the strongest consistent properties constructible in B (…). These properties are mutually exclusive, and every object must have just one of these. Clearly, two objects are indiscernible in B just in the case they have the same B-maximal property.

En el texto de Kim, B denota una cierta familia de propiedades. Una propiedad es consistente si es instanciada por (al menos) un objeto en un mundo posible. La noción de propiedad maximal se define relativamente a esta familia. Si explicamos "*P* es al menos tan fuerte como *Q*" como "$P \supset Q$", podemos extraer de la caracterización kimeana de propiedad B-maximal los siguientes postulados:

[KMP1] *P* es una propiedad *B*-maximal syss *P* es una *B*-propiedad consistente y no existe ninguna *B*-propiedad *Q* que sea consistente y tal que *Q* implique *P* pero *P* no implique *Q*.

[KMP2] Para cada mundo posible *w* y cada individuo *a* en *w*, existe una propiedad *B*-maximal *P* tal que *a* tiene *P* en *w*.

[KMP3] Si *P* y *Q* son dos propiedades *B*-maximales distintas, entonces no hay ningún individuo que tenga *P* y *Q* en el mismo mundo posible.

[KMP4] Dos individuos son *B*-indiscernibles syss tienen la misma propiedad *B*-maximal.

Donde decimos "individuo", podemos decir también "tupla de individuos". [KMP4] hace referencia a la noción de *B*-indiscernibilidad, entendida por Kim en los siguientes términos: un individuo *a* en un mundo *w* es *B*-indiscernible del individuo *b* en un mundo *u* si y sólo si para cada propiedad *P* en *B*: *a* tiene *P* en *w* syss *b* tiene *P* en *u*. Emplearemos el símbolo "$\approx_B$" para denotar la relación de *B*-indiscernibilidad, que es una relación binaria entre pares de la forma (*w*, *a*) donde *w* es un mundo posible y *a* es un individuo (o tupla de ellos) de *D*(*w*). Usando el lenguaje de TNP, podemos definir la *B*-indiscernibilidad en el sentido de Kim de la siguiente manera:

[K-B-INDISC] $(w, a) \approx_B (u, b) \Leftrightarrow_{df} \forall P \in B\,(P_w(a) \leftrightarrow P_u(b))$

Introduzcamos la expresión Max_B para predicar la maximalidad, de forma que $Max_B(P)$ se lea como "*P* es una propiedad *B*-maximal". Los anteriores postulados pueden reescribirse entonces de la siguiente forma:

[*KMP1*] $Max_B(P) \leftrightarrow P \in B \ \wedge \exists w \exists a\, P_w(a) \wedge \forall Q \in B\, (\exists w \exists a\, Q_w(a) \wedge Q \supset P \rightarrow P = Q)$

[*KMP2*] $\forall w\, \forall a \in D(w)\, \exists P \in B\, (Max_B(P) \wedge P_w(a))$

[*KMP3*] $\forall P\, \forall Q\, \forall w\, \forall a \in D(w)\, (Max_B(P) \wedge Max_B(Q) \wedge P_w(a) \wedge Q_w(a) \rightarrow P = Q)$

[*KMP4*] $(w, a) \approx_B (u, b) \leftrightarrow \forall P\, (Max_B(P) \rightarrow (P_w(a) \leftrightarrow P_u(b)))$

[*KMP1*] puede servir como una definición de *B*-maximalidad. Denote $\Delta^B_{w,a}$ la propiedad *B*-maximal de un objeto *a* en un mundo *w*, si tal propiedad existe. En caso negativo, $\Delta^B_{w,a}$ queda indefinida. La expresión $\Delta^B_{w,a} \cong \Delta^B_{u,b}$ significará: " $\Delta^B_{w,a}$ y $\Delta^B_{u,b}$ o ambas están definidas y entonces son idénticas o ambas están indefinidas". Resulta obvio que para cualquier mundo *w* y todo objeto *a* en *w*, tenemos:

$(\Delta^B_{w,a})_w(a)$ $\Leftrightarrow$ el objeto *a* tiene en *w* la propiedad *B*-maximal

$\Leftrightarrow$ $\Delta^B_{w,a}$ está definida

Los postulados [*KMP2*] y [*KMP4*] pueden reescribirse entonces en una forma equivalente y más abreviada:

[*KMP2*]´ $\forall w\, \forall a \in D(w)\, (\Delta^B_{w,a})_w(a)$

[*KMP4*]´ $(w, a) \approx_B (u, b) \leftrightarrow \Delta^B_{w,a} \cong \Delta^B_{u,b}$

Si asumimos que la familia *B* está cerrada bajo *operaciones booleanas finitas*, entonces los postulados [*KMP1*] - [*KMP4*] no son independientes entre sí:

Proposición 1

Bajo clausura booleana finita de *B*:

(i) [*KMP1*] implica [*KMP3*]

(ii) [*KMP1*] y [*KMP2*] implican [*KMP4*]

Prueba

(i) Asumamos [*KMP1*]. Sean *P* y *Q* dos propiedades *B*-maximales tales que $P_w(a)$ y $Q_w(a)$. Entonces $[P \wedge Q]$ es consistente y, obviamente, $[P \wedge Q] \supset P$ y $[P \wedge Q] \supset Q$. Por la *B*-maximalidad de *P* y *Q* y por [*KMP1*], tenemos que $P \supset [P \wedge Q]$ y que $Q \supset [P \wedge Q]$. Por tanto, $P = Q$.

(ii) Asumamos [*KMP1*] y [*KMP2*]. La implicación $\rightarrow$ en [*KMP4*] es obvia. Mostraremos la otra dirección. Sean *w*, *a*, *u*, *b* tales que $(w, a) \approx_B (u, b)$ no es el caso. Entonces para algún $Q \in B$ tenemos $Q_w(a)$ y $\div Q_u(b)$. Por [*KMP2*], existen $\Delta^B_{w,a}$ y

$\Delta^B_{u,b}$, y por tanto $[\ \Delta^B_{w,a} \wedge Q\ ]_w(a)$ y $[\ \Delta^B_{u,b} \wedge \div Q\ ]_u(b)$. Por [*KMP1*], tenemos $[\Delta^B_{w,a} \wedge Q\,] = \Delta^B_{w,a}$ y $[\Delta^B_{u,b} \wedge \div Q\,] = \Delta^B_{u,b}$. Por tanto, $\Delta^B_{w,a}$ y $\Delta^B_{u,b}$ no pueden ser idénticas. *q.e.d.*

[*KMP4*] motiva una definición alternativa de propiedades *B*-maximales: pueden ser definidas como propiedades asociadas con clases de equivalencia de $\approx_B$. Considérese, en efecto, la siguiente condición:

[*KMP5*] $Max_B(P) \leftrightarrow P \in B \wedge \exists w\, \exists a \in D(w)\ \forall u\, \forall b \in D(u)\ (P_u(b) \leftrightarrow (w, a) \approx_B (u, b))$

Pero tal definición alternativa resulta, en realidad, equivalente a [*KMP1*]. Veámoslo:

Proposición 2

[*KMP5*] es equivalente a [*KMP1*]

Prueba

Mostramos que las siguientes condiciones son equivalentes para cada $P \in B$:

(1) $\exists w \exists a\ P_w(a) \wedge \forall Q \in B\ (\exists w \exists a\ Q_w(a) \wedge Q \supset P \rightarrow P = Q)$

(2) $\exists w\, \exists a \in D(w)\ \forall u\, \forall b \in D(u)\ (P_u(b) \leftrightarrow (w, a) \approx_B (u, b))$

(1) $\Rightarrow$ (2)

Por (1) tenemos que para algún w y algún a: $P_w(a)$. Considérese un mundo posible arbitrario u y un objeto b en ese mundo. Si $(w, a) \approx_B (u, b)$, entonces obviamente $P_u(b)$. Supongamos ahora que $(w, a) \approx_B (u, b)$ no es el caso. Entonces para algún $Q \in B$ tenemos que $Q_w(a)$ y $\div Q_u(b)$. Mostraremos que $P_u(b)$ no es el caso. Para ello supongamos lo contrario, esto es que $P_u(b)$. Por consiguiente, $[P \wedge \div Q\,]_u(b)$ y claramente $[P \wedge \div Q\,] \supset P$. Por (1), tenemos $[P \wedge \div Q\,] = P$. Puesto que $P_w(a)$, tenemos también que $\div Q_w(a)$, con lo que llegamos a una contradicción.

(2) $\Rightarrow$ (1)

Sean w y a tales que: $\forall u\, \forall b \in D(u)\ (P_u(b) \leftrightarrow (w, a) \approx_B (u, b))$. Entonces también tenemos que $P_w(a)$. Sea Q una *B*-propiedad tal que $Q \supset P$ y $Q_u(b)$, para algún u, b. Esto implica que $P_u(b)$ y, por tanto, $(w, a) \approx_B (u, b)$. Ahora mostremos que $P \supset Q$. Si $P_v(c)$, entonces $(w, a) \approx_B (v, c)$. En consecuencia, $(u, b) \approx_B (v, c)$. De lo que se sigue que $Q_v(c)$. *q.e.d.*

Obsérvese que $(w, a) \approx_B (u, b)$ es equivalente a $\forall Q \in B\ (Q_w(a) \rightarrow Q_u(b))$. De ello resulta el siguiente:

Corolario 1

Bajo [*KMP1*] o [*KMP5*], las siguientes dos condiciones resultan equivalentes para todo $P \in B$:

(i) $Max_B(P)$

(ii) $\exists w\, \exists a\, \forall u\, \forall b \in D(u)\ (P_u(b) \leftrightarrow \forall Q \in B\ (Q_w(a) \rightarrow Q_u(b)))$

Se sigue de este corolario que una propiedad P es B-maximal syss P es la conjunción de todas las B-propiedades de algún objeto a en un mundo w. Esto significa que [*KMP2*] se sigue si B está cerrado bajo operaciones booleanas infinitas no restringidas. Según Kim, este hecho parece ser constitutivo. Kim escribe (Kim 1993, p. 152): "I don't see any special problem with an infinite procedure here, any more than in the case of forming infinite unions of sets or the addition of infinite series of numbers."

Veremos luego que, en algunos casos interesantes, esta asunción resulta problemática.

2.2.2. Sider sobre maximalidad

La explicación de la noción de B-maximalidad propuesta por Theodore Sider en Sider (1999) es diferente de la de Kim. Su punto de partida es la noción de B-indiscernibilidad, que es más restrictiva que $\approx_B$. Sider la llama B-indiscernibilidad *global*. Simbolicemos esta relación con $\tilde{\approx}_B$. La B-indiscernibilidad *global* de Sider queda definida por:

[S-B-INDISC] $(w, a)\ \tilde{\approx}_B\ (u, b) \Leftrightarrow_{df}$ hay un B-isomorfismo de w en u que transforma a en b.

Un B-isomorfismo de w en u se define como una biyección f del universo del mundo w en el de u tal que para cada B-propiedad Q y cualquier a del universo de w se cumple la siguiente condición: $Q_w(a)$ syss $Q_u(f(a))$.

Considere el lector el siguiente pasaje de Sider (cfr. Sider 1999, p. 919):

> The relation of global A-indiscernibility (…) is clearly an equivalence relation; with each of its equivalence classes there is an associated property: the property had by all and only those members of the equivalence class. I call these properties *maximal A-properties*. (…) The recipe for coming up with a maximal property is this: select some possible object and describe all its features with respect to A, both intrinsic and relational. *All* relational properties must be mentioned, and so along the way it will be necessary to completely describe the distribution of A-properties and relations throughout the entirety of that object's possible world.

La noción sideriana de B-propiedad maximal es abstraída de la relación $\tilde{\approx}_B$. Simbolizaremos el predicado de maximalidad sideriana mediante $SMax_B$. De acuerdo con Sider, su definición queda establecida mediante la siguiente fórmula:

[*SMP1*] $SMax_B(P) \leftrightarrow P \in B \wedge \exists w\, \exists a \in D(w)\, \forall u\, \forall b \in D(u)\, (P_u(b) \leftrightarrow (w, a)\ \tilde{\approx}_B (u, b))$

Como mera consecuencia de [*SMP1*], tenemos:

Proposición 3

[*SMP1*] y [*KMP1*] (o bien [*KMP5*]) implican conjuntamente que para cualesquiera dos mundos posibles *w*, *u*, y cualesquier objetos *a*, *b*, y toda *B*-propiedad *P:*

(i) si $(w, a)\ \tilde{\approx}_B (u, b)$, entonces $(w, a) \approx_B (u, b)$,

(ii) si $SMax_B(P)$ entonces $Max_B(P)$.

Prueba (inmediata).

Para esta clase de propiedades maximales, la unicidad resulta directamente de la proposición anterior más la Proposición 1(i):

Corolario 2

[*SMP1*] implica: $\forall P, Q\, \forall w\, \forall a\, (SMax_B(P) \wedge SMax_B(Q) \wedge P_w(a) \wedge Q_w(a) \rightarrow P = Q)$

Denote $\Theta^B_{w,a}$ la propiedad *B*-maximal de un objeto *a* en un mundo *w*, si tal propiedad existe. En caso contrario, $\Theta^B_{w,a}$ queda indefinida. Incluso dando por garantizada la condición de clausura booleana infinitaria irrestricta de *B*, no está claro si esta condición garantiza a su vez la existencia de $\Theta^B_{w,a}$ para todo *w* y todo *a* en *w*. La asunción de que $\Theta^B_{w,a}$ está siempre definida resultará crucial para el argumento de Sider. En consecuencia, añadiremos el siguiente postulado:

[*SMP2*] $\forall w\, \forall a \in D(w)\, \exists P \in B\ \ (SMax_B(P) \wedge P_w(a))$

3. Superveniencia y reducibilidad

Consideremos algunas variantes de la tesis de la superveniencia como relación entre dos familias de propiedades *A* y *B*. Asumamos que la familia *B* (de propiedades *básicas*, también llamadas *subvenientes*) está cerrada al menos bajo operaciones booleanas finitas. La principal motivación existente tras los distintos conceptos de superveniencia es la idea de una relación de determinación (o dependencia) no reductiva. Las propiedades *supervenientes* están determinadas por las básicas sin dejarse en cambio reducir por éstas.

3.1. Superveniencia local, superveniencia global y reducibilidad en el marco de TNP.

Tenemos una variante local y otra global de superveniencia de *A* sobre *B*. En la literatura especializada se pueden distinguir dos formas de superveniencia local. La llamada *superveniencia fuerte* de *A* sobre *B* se define así:

(*SF*) $\forall w, u \in W\ \forall a\ \forall b\ ((w, a) \approx_B (u, b) \rightarrow (w, a) \approx_A (u, b))$

La versión *débil* es:

(*SD*) $\forall w \in W\ \forall a\ \forall b\ ((w, a) \approx_B (w, b) \rightarrow (w, a) \approx_A (w, b))$

Por otra parte, una explicación comúnmente aceptada de *superveniencia global* tiene la siguiente forma:

(*SG*) $\forall w, u \in W\ \forall f\ (w \cong_f^B u \rightarrow w \cong_f^A u)$

$w \cong_f^B u$ significa que f es una biyección del universo de w en el universo de u tal que para cada Q de B y cualquier a del universo de w se cumple: $Q_w(a)$ syss $Q_u(f(a))$.

La reducción en términos de *B*-propiedades usualmente se explica como *coexistencia necesaria*:

(*CN*) $\forall P \in A\ \exists Q \in B\ \forall w \in W\ \forall a \in w\ (P_w(a) \leftrightarrow Q_w(a))$.

Una forma más débil de reducción es como sigue:

(*CND*) $\forall P \in A\ \forall w \in W\ \exists Q \in B\ \forall a \in w\ (P_w(a) \leftrightarrow Q_w(a))$.

3.2. Los argumentos reductivos de Kim y Sider

En una serie de artículos (cfr. Kim 1993), Jaegwon Kim arguyó que si la familia B de propiedades subvenientes está cerrada bajo operaciones infinitas irrestrictas, entonces (SF) implica (CN) y, análogamente, (SD) implica (CND). Theodore Sider ofrece en Sider 1999 un argumento en favor de la tesis de que (SG) implica (CN). Como dijimos, tanto Kim como Sider hacen un uso esencial de sus respectivas nociones de propiedad maximal (de un objeto en un mundo), las cuales hemos visto arriba (2.2.1 y 2.2.2.). Con ayuda de sendas nociones, ambos definen para cada A-propiedad P su correspondiente B-sustituto Δ^P (Kim) o, respecivamente, Θ^P (Sider) en los términos, muy similares, que vemos a continuación:

$$\Delta^P =_{df} \{\Delta^B_{w,a} : P_w(a) \wedge w \in W \wedge a \in D(w)\}$$

$$\Theta^P =_{df} \{\Theta^B_{w,a} : P_w(a) \wedge w \in W \wedge a \in D(w)\}$$

Kim y Sider ofrecen sendos argumentos con el objeto de mostrar que, bajo (SF), en el caso de Kim, y (SG), en el de Sider, las arriba definidas Δ^P respectivamente Θ^P son *B*-propiedades necesariamente equivalentes a *P*. Ambos argumentos reflejan pareja estructura.

Si en el argumento que veremos a continuación sustituimos $\Delta^B_{w,a}$ por $\Gamma_{w,a}$, Δ^P por Γ^P así como $\approx_A$ y $\approx_B$ por $\equiv_A$ y $\equiv_B$ respectivamente, lo que obtenemos es una recons-

trucción del argumento de Kim. Y si sustituimos $\Theta^{B}_{w,a}$ por $\Gamma_{w,a}$, Θ^{P} por Γ^{P} e igualmente $\vec{\approx}_{A}$ y $\vec{\approx}_{B}$ por $\equiv_{A}$ y $\equiv_{B}$ respectivamente, el argumento resultante es esencialmente el dado por Sider (cfr. Sider 1999, p. 920-921) (excepto en algún detalle poco importante que no viene al caso):

(i) asumamos $P_{w}(a)$

(ii) entonces $(\Gamma_{w,a})_{w}(a)$

(iii) y, de ahí, $(\Gamma^{P})_{w}\, a)$

(iv) asumamos $(\Gamma^{P})_{w}(a)$

(v) entonces, para algún *u* y algún *b:* $P_{u}(b)$ and $(\Gamma_{u,b})_{w}(a)$

(vi) así $(w, a) \equiv_{B} (u, b)$

(vii) por tanto $(w, a) \equiv_{A} (u, b)$

(viii) de modo que $P_{w}(a)$

El paso (iii) se sigue de (ii) y la definición de Δ^{P}, respectivamente Θ^{P}. Similarmente, (v) se sigue de (iv) más la definición de Δ^{P}, respectivamente Θ^{P}. El paso (vi) es una consecuencia del paso (v), la definición de $\Delta^{B}_{u,b}$, respectivamente de $\Theta^{B}_{u,b}$ y [*KMP5*], respectivamente [*SMP1*]. El paso (vii) se sigue de (vi) y de la asunción de la superveniencia, (SF) o respectivamente (SG). El paso (viii) se sigue a su vez de (vii) y (iv). El paso (ii) requiere en ambos casos algunas asunciones sobre la existencia de ciertas propiedades. Bajo nuestra reconstrucción, $(\Delta^{B}_{w,a})_{w}(a)$ respectivamente $(\Theta^{B}_{w,a})_{w}(a)$ se cumple syss $\Delta^{B}_{w,a}$ respectivamente $\Theta^{B}_{w,a}$ existen. Este paso requiere claramente [*KMP2*], respectivamente [*SMP2*]. Tanto Kim como Sider parecen concebir esta asunción existencial como no problemática. Kim afirma que está garantizada por la condición de clausura infinita irrestricta de *B*. En el caso de Sider, la justificación de [*SMP2*] resulta menos transparente. Mostraremos que, al menos bajo una interpretación interesante del formalismo de TNP, estos supuestos cruciales son falsos y que, en consecuencia, de acuerdo con esta interpretación, los argumentos apoyados en ellos no son correctos.

4. Superveniencia en términos de modelos

4.1. Paráfrasis modelo-teórica

TNP es un marco abstracto (preteórico) para el estudio de la superveniencia y reducción de propiedades. Esta teoría es neutral con respecto a muchas cuestiones concernientes a la naturaleza de las propiedades y de los mundos posibles. En nuestra opinión, las tesis generales sobre las relaciones entre la superveniencia y la reducibilidad

de propiedades también son aplicables a sistemas de propiedades definidas por fórmulas. Los argumentos reductivos reconstruidos más arriba serían de poco interés si tales argumentos resultaran ser incorrectos en relación con tales familias de propiedades. Pero justamente ocurre que, bajo la paráfrasis modelo-teórica, los supuestos cruciales de TNP resultan ser verdaderos sólo en relación con un conjunto restringido de condiciones.

La teoría de modelos proporciona herramientas conceptuales suficientes para explicar todas las nociones involucradas en el debate de la superveniencia. La mayoría de ellas tienen contrapartes modelo-teóricas que resultan bastante naturales. Considere el lector un conjunto de letras predicativas (un vocabulario) L junto con dos subvocabularios disjuntos L_B y L_A. Sea **K** una clase de estructuras para L que representa la colección de mundos posibles. Cada estructura M contiene un universo de objetos *dom*(M) y para cada predicado P de L, su denotación P^M. Tal estructura puede ser vista (o al menos ser representada) como un mundo posible. En consecuencia, una propiedad n-aria asociada con **K** es una función que asigna a cada estructura M de **K** un conjunto de n-tuplas de objetos de *dom*(M). Funciones de este tipo son las que en la bibliografía especializada reciben a menudo en inglés el nombre de *queries*. Si restringimos nuestras consideraciones a propiedades que sean expresables con ayuda de fórmulas de L, entonces algunas de estas *queries* (de hecho la mayoría) no cuentan como propiedades.

4.2. Indiscernibilidad y maximalidad

Las relaciones de indiscernibilidad $\approx$ y $\vec{\approx}$ definidas en el marco de TNP (véase 2.2.) tienen como contrapartes modelo-teóricas dos relaciones definidas para clases de fórmulas en lugar de para clases de propiedades. Sean M y N estructuras para L y sea Φ una colección de fórmulas en un subvocabulario de L. Sean $\bar{a}$ y $\bar{b}$ tuplas de elementos de la misma longitud de M y N respectivamente. Bajo Φ-isomorfismo de M en N entendemos una biyección f entre *dom*(M) y *dom*(N) tales que para toda fórmula $\varphi(x_1, \ldots, x_n) \in \Phi$ y cualesquier elementos $a_1, \ldots, a_n$ en M: $M \quad \varphi[a_1, \ldots, a_n]$ syss $N \quad \varphi[f(a_1), \ldots, f(a_n)]$.

Diremos que $\bar{a}$ es en M *indiscernible localmente* de $\bar{b}$ en N *con respecto a* Φ (en símbolos $(M, \bar{a}) \approx_\Phi (N, \bar{b})$) syss $\bar{a}$ satisface en M exactamente las mismas fórmulas Φ que satisface $\bar{b}$ en N. Una tupla $\bar{a}$ en M es *globalmente indiscernible* de la tupla $\bar{b}$ en N *con respecto a* Φ (en símbolos $(M, \bar{a}) \vec{\approx}_\Phi (N, \bar{b})$) syss existe un Φ-isomorfismo f de M en N que transforma $\bar{a}$ en $\bar{b}$. Claramente, $(M, \bar{a}) \vec{\approx}_\Phi (N, \bar{b})$ implica $(M, \bar{a}) \approx_\Phi (N, \bar{b})$. Con ayuda de las relaciones $\approx_\Phi$ y $\vec{\approx}_\Phi$, podemos ahora definir contrapartes modelo-teóricas de las nociones de maximalidad reconstruidas en 2.2. Nuevamente, en lugar de maximalidad de propiedades, hablaremos de maximalidad de fórmulas.

Diremos que una fórmula $\varphi(\bar{x}) \in \Phi$ es *localmente maximal* con respecto a Φ en una clase **K** syss se cumple:

[L-Max] hay M en **K** y una tupla $\bar{a}$ en M tales que para toda N en **K** y toda $\bar{b}$ en N:

$N \quad \varphi(\bar{x})\,[\bar{b}]$ syss $(M, \bar{a}) \approx_{\Phi} (N, \bar{b})$

Diremos que una fórmula $\varphi(\bar{x}) \in \Phi$ es *globalmente maximal* en una clase **K** syss se cumple:

[G-Max] hay M en **K** y una tupla $\bar{a}$ en M tales que para toda N en **K** y toda $\bar{b}$ en N:

$N \quad \varphi(\bar{x})\,[\bar{b}]$ syss $(M, \bar{a}) \tilde{\approx}_{\Phi} (N, \bar{b})$

4.3. Superveniencia modelo-teórica

Sean $\Phi(\mathcal{A})$ y $\Phi(B)$ dos colecciones de fórmulas de los vocabularios $L_{\mathcal{A}}$ y L_B respectivamente, las cuales contienen todas las fórmulas atómicas de estos dos lenguajes. $\Phi(\mathcal{A})$ y $\Phi(B)$ representan respectivamente las familias de propiedades supervenientes y subvenientes. Ahora podemos formular paráfrasis modelo-teóricas de los distintos conceptos de superveniencia relativamente a una clase **K** de modelos para un vocabulario L que incluya ambos $L_{\mathcal{A}}$ y L_B. Comencemos por la SF. $\Phi(\mathcal{A})$ *superviene fuertemente sobre* $\Phi(B)$ con respecto a **K** si se cumple que:

[SF] Para cada M, N en **K** y cualesquiera tuplas $\bar{a}$ y $\bar{b}$: $(M, \bar{a}) \approx_{\Phi(B)} (N, \bar{b})$ implica $(M, \bar{a}) \approx_{\Phi(\mathcal{A})} (N, \bar{b})$.

Por otro lado, podemos expresar la *superveniencia débil* de $\Phi(\mathcal{A})$ *sobre* $\Phi(B)$ con respecto a **K** mediante la siguiente condición:

[SD] Para cada M en **K** y cualesquiera tuplas $\bar{a}$ y $\bar{b}$: $(M, \bar{a}) \approx_{\Phi(B)} (M, \bar{b})$ implica $(M, \bar{a}) \approx_{\Phi(\mathcal{A})} (M, \bar{b})$.

Finalmente, diremos que $\Phi(\mathcal{A})$ *superviene globalmente* sobre $\Phi(B)$ con respecto a **K** si se cumple lo siguiente:

[SG] Cada $\Phi(B)$-isomorfismo de una estructura M de la clase **K** en la estructura N de **K** es también un $\Phi(\mathcal{A})$-isomorfismo de M en N.

En nuestro marco modelo-teórico, podemos explicar igualmente sendas nociones, débil y fuerte, de reducibilidad de $\Phi(\mathcal{A})$ a $\Phi(B)$ con respecto a **K**. Estas versiones de reducibilidad entre familias de fórmulas corresponden a las condiciones de coextensión débil y coextensión fuerte definidas con respecto a una colección de mundos posibles.

Diremos que $\Phi(\mathcal{A})$ es *reducible* a $\Phi(B)$ en **K** si se cumple que:

[CN] Para cada fórmula φ de $\Phi(\mathcal{A})$ existe una fórmula ψ en tal que para cada M en $\mathbf{K}$ φ y ψ están satisfechas en M por exactamente las mismas tuplas de elementos.

Más adelante, para cada L_A-fórmula φ, hablaremos de la L_B-fórmula correspondiente según [CN] como de su "L_B-sustituto".

Por otro lado, la *reducción débil* de $\Phi(\mathcal{A})$ a $\Phi(B)$ en $\mathbf{K}$ puede ser expresada en los términos siguientes:

[CND] Para cada fórmula φ de $\Phi(\mathcal{A})$ y cada M en $\mathbf{K}$ hay una fórmula ψ de $\Phi(B)$ tal que φ y ψ son satisfechas en M por exactamente las mismas tuplas de elementos.

Las siguientes relaciones de implicación se cumplen cualquiera que sea la elección particular de la clase $\mathbf{K}$:

$$\begin{array}{ccccc} [\text{CND}] & \leftarrow & [\text{CN}] & & \\ \downarrow & & \downarrow & & \\ [\text{SD}] & \leftarrow & [\text{SF}] & \rightarrow & [\text{SG}] \end{array}$$

Nótese que si $\mathbf{K}$ es la clase de todos los modelos de una teoría T (de primer orden) en un vocabulario L y las colecciones $\Phi(\mathcal{A})$ y $\Phi(B)$ son, respectivamente, el conjunto de todas las fórmulas finitarias en algún subvocabulario L_A y L_B de L, entonces [CN], [SF] y [SG] coinciden. Para poder verlo, obsérvese que la condición de SG de todas las fórmulas finitarias en L_A sobre todas las fórmulas finitarias en L_B no es otra cosa que la definibilidad implícita de todos los predicados de L_A por parte de las fórmulas finitarias en el vocabulario L_B con respecto a la teoría T. Por otra parte, la condición de reducibilidad de todas las fórmulas finitarias en L_A a las fórmulas finitarias en L_B es equivalente a la definabilidad explícita de todos los predicados en L_A por parte de fórmulas finitarias en L_B con respecto a la teoría T. En otras palabras, la implicación [SG] $\rightarrow$ [CN] es equivalente al Teorema de definibilidad de Beth. Esta observación en realidad no es nueva, sino que fue hecha ya en su momento por Hellman y Thompson (cfr. Hellman, G. y F.W. Thompson 1975). Conviene notar al mismo tiempo que la estrucutra de la prueba del Teorema de Beth no guarda semejanza con el patrón argumental común a los argumentos reductivos de Kim y Sider. En particular, no descansa en suposición alguna acerca de la existencia de fórmulas local o globalmente maximales. De hecho, la prueba es más bien compleja comparada con los argumentos de Kim y Sider. La corrección de estos dos argumentos no puede mostrarse en el marco de una lógica finitaria. La razón no es simplemente que ambos argumentos hacen uso de operaciones infinitas. Téngase en cuenta que la mera existencia de fórmulas maximales es ya altamente problemática. Todo indica que, restringiéndonos a lenguajes finitos e

identificando mundos posibles con clases de modelos, no necesitamos invertir más tiempo en el estudio de dichos argumentos. O bien restringimos nuestras consideraciones a clases elementales de modelos y entonces obtenemos directamente (sin ulterior argumentación) los resultados reduccionistas, o bien incluimos clases de modelos no elementales, lo que sería tanto como no poder esperar ya ningún resultado interesante en relación con nuestro debate. Pero, como vamos a ver en lo que queda del presente artículo, la teoría de modelos puede enseñarnos todavía muchas más cosas de interés que lo que a primera vista parecía.

4.4. Superveniencia y modelos finitos

¿Qué sucedería si impusiéramos restricciones a la cardinalidad de la clase K? Consideremos primeramente la situación que se nos presenta cuando K es una clase arbitraria de estructuras *finitas* en un vocabulario finito L con subvocabularios disjuntos L_A y L_B. Es un hecho bien conocido que para un vocabulario dado cada estructura finita puede ser descrita hasta la isomorfía por una única fórmula *finitaria* en ese vocubulario. Más aún, si M es una estructura finita L y L' es un subvocabulario de L, entonces para cada tupla $\bar{a}$ del universo de M podemos encontrar, de un modo canónico, una fórmula $\delta_{M,\bar{a}}$ en L' tal que para cada estructura N para L y cada tupla $\bar{b}$ (de la misma longitud que $\bar{a}$) del universo de N se cumple lo siguiente:

[#] $N \quad \delta_{M,\bar{a}}[\bar{b}]$ syss hay un L'-isomorfismo entre M y N que transforma $\bar{a}$ en $\bar{b}$.

Tal fórmula $\delta_{M,\bar{a}}$ es maximal tanto global como localmente. De ello se sigue claramente que si Φ es el conjunto de todas las fórmulas finitas L', entonces las relaciones de indiscernibilidad $\approx_\Phi$ y $\tilde{\approx}_\Phi$ coinciden si se definen con respecto a una clase de estructuras finitas. Obsérvese que cualquier fórmula $\delta(\bar{x})$ del tipo mencionado no puede ser definida como la conjunción de todas las fórmulas finitas satisfechas por $\bar{a}$ en M. Tal conjunción no es una fórmula finita. Las paráfrasis modelo-teóricas de [*KMP2*] y [*SMP2*] se cumplen a pesar de que no podemos construir conjunciones infinitas. Hay distintas maneras de construir fórmulas maximales. Sin embargo, hay que notar que la construcción de $\delta_{M,\bar{a}}$ supone cuantificación (finita) e identidad. Una manera sencilla de construirla podría ser mediante el siguiente ejemplo:

sea $M = (dom(M), R)$, donde $dom(M) = \{a_1, a_2, a_3\}$ y R es la siguiente relación binaria: $\{(a_1, a_2), (a_1, a_3), (a_3, a_3)\}$. Sea $\bar{a} = (a_1, a_2)$ y sea L un lenguaje con un símbolo relacional binario P (que representa R). Como $\delta_{M,\bar{a}}$ puede tomarse la siguiente fórmula de L:

$\exists x_3 \big(\{\neg(x_i = x_j): i < j \leq 3\} \wedge P(x_1, x_2) \wedge P(x_1, x_3) \wedge P(x_3, x_3) \wedge \neg P(x_1, x_1) \wedge$
$\wedge \neg P(x_2, x_1) \wedge \neg P(x_2, x_2) \wedge \neg P(x_2, x_3) \wedge \neg P(x_3, x_1) \wedge \neg P(x_3, x_2) \wedge \forall x_4 \quad \{x_4 = x_i:$
$i < j \leq 3\}\big)$

Es fácil ver que para cada estructura N de la forma ($dom(N)$, S), donde S es una relación binaria y cada par ordenado (b_1, b_2) de elementos de $dom(N)$ se cumple:

$N \quad \delta_{M,\bar{a}}[x_1/b_1, x_2/b_2]$ syss existe un isomorfismo f de M en N tal que $f(a_1) = b_1$ y $f(a_2) = b_2$.

La existencia de fórmulas maximales tiene, entre otras, la siguiente consecuencia:

Proposición 5

Sean $\Phi(A)$ y $\Phi(B)$ los conjuntos de todas las *fórmulas infinitas pero contables* en los dos subvocabularios L_A y L_B de L respectivamente. Para cada clase **K** de estructuras finitas para L, las siguientes condiciones resultan equivalentes entre sí:

(1) $\Phi(A)$ superviene fuertemente sobre $\Phi(B)$ con respecto a **K**.

(2) $\Phi(A)$ superviene globalmente sobre $\Phi(B)$ con respecto a **K**.

(3) $\Phi(A)$ es reducible a $\Phi(B)$ en **K**.

Prueba

Basta probar la dirección (1) → (3). Sea $\varphi(\bar{x})$ cualquier fórmula infinitaria contable en L_A. Y ahora, para cada M en **K** y cada tupla $\bar{a}$, sea $\delta_{M,\bar{a}}$ una fórmula de L_B que satisfaga la condición de maximalidad. Podemos definir una fórmula $\psi(\bar{x})$ en la clase **K** esencialmente del mismo modo en que Kim define Δ^P para P, es decir, mediante la disyunción:

$$\{\delta_{M,\bar{a}}: M \quad \varphi(\bar{x})[\bar{a}], M \in \mathrm{K}\}.$$

Nótese que, aunque **K** es una *clase arbitraria* de estructuras finitas para L, hay sólo un número contable de fórmulas de la forma $\delta_{M,\bar{a}}$ y, así, la expresión de arriba resulta ser una disyunción contable, la cual, en virtud de (1), es obviamente equivalente a $\varphi(\bar{x})$ en cada estructura de la clase **K**. *q.e.d.*

Así pues, en lenguajes infinitarios y disponiendo de clases arbitrarias de estructuras finitas, tenemos que se cumple el siguiente diagrama de relaciones de implicación:

$$\begin{array}{ccccc} [CN] & \leftrightarrow & [SF] & \leftrightarrow & [SG] \\ \downarrow & & & & \\ [CND] & \leftrightarrow & [SD] & & \end{array}$$

4.5. Dificultades con los argumentos reductivos en $\mathcal{L}_{\infty\omega}$

La asunción de que todos los mundos posibles son finitos puede resultar en algunos contextos bastante natural, pero no es precisamente la que adoptan Kim y Sider. En el presente apartado, veremos que cuando adoptamos una lógica con conyunciones y disyunciones arbitrariamente largas y sin restricciones sobre la cardinalidad de los modelos, entonces los argumentos de Kim y Sider topan con serias dificultades. La lógica que parece representar mejor tales irrestricciones es $\mathcal{L}_{\infty\omega}$.

Comencemos con el argumento de Sider. Podemos usar el hecho de que el Teorema de Beth no es válido en $\mathcal{L}_{\infty\omega}$ (cfr. Gregory 1974, p. 26) para mostrar que la afirmación de Sider de que [SG] implica [CN] resulta falsa. Sin embargo, un contraejemplo debido a Gregory puede servirnos, tras una ligera modificación, para mostrar que [SG] ni siquiera implica [SD] en general (cfr.: Polanski 2005). De este modo, el argumento de Sider no puede ser correcto. Nótese, sin embargo, que al contrario que en el caso de los modelos finitos, la paráfrasis modelo-teórica del axioma [*SMP2*], que es uno de los supuestos implícitos de Sider, resulta falsa cuando tratamos con una clase arbitraria de modelos infinitos incluso si el lenguaje del que disponemos permite conyunciones y disyunciones infinitas. Para verlo, consideremos las estructuras $\mathcal{Q}$ y R consistentes en los números racionales y los números reales con sus respectivos órdenes usuales. Valiéndonos de la técnica standard *back-and-forth* y de un conocido teorema debido a Karp, se puede mostrar que para cada número racional q las estructuras ($\mathcal{Q}$, q) y (R, q) son equivalentes con respecto a todas las sentencias infinitarias. Por tanto, q tiene en $\mathcal{Q}$ y en R las mismas propiedades expresables en $\mathcal{L}_{\infty\omega}$. Pero, claramente, no existe un isomorfismo entre $\mathcal{Q}$ y R.

¿Qué hay de la prueba de Kim? En principio, parece que puede formalizarse en $\mathcal{L}_{\infty\omega}$ sin mayores problemas. Comencemos, por razones heurísticas, con el argumento de Kim en favor de la implicación [SD] → [CND]. Asumamos que la colección $\Phi(\mathcal{A})$ de todas las fórmulas infinitarias en $L_{\mathcal{A}}$ superviene débilmente en K sobre la colección $\Phi(\mathcal{B})$ de todas las fórmulas infinitarias en $L_{\mathcal{B}}$. Supongamos ahora que pudiéramos construir, para cada modelo M de K y cualquier n-tupla $\bar{a}$ en M, una fórmula $\psi_{M,\bar{a}}$ en $L_{\mathcal{B}}$ tal que para todas las tuplas $\bar{b}$ en M de la misma longitud que $\bar{a}$ se cumpla: $M \quad \psi^{\bar{a}}_{M}[\bar{b}]$ syss $\bar{a}$ y $\bar{b}$ satisfacen en M las mismas fórmulas infinitarias en $L_{\mathcal{B}}$. En tal caso podríamos sin duda proporcionar un $L_{\mathcal{B}}$-sustituto de cualquier fórmula $\varphi(\bar{x})$ en $L_{\mathcal{A}}$. La siguiente disyunción infinita hace las veces de dicha expresión:

$\{\psi^{\bar{a}}_{M}: M \quad \varphi(\bar{x})[\bar{a}],\ \bar{a} \in M\}$.

Obsérvese que para cada *M* la cardinalidad del conjunto de todas las *n*-tuplas de *M* es o bien finita o bien igual a la cardinalidad de *M*. El aparato de $\mathcal{L}_{\infty\omega}$ nos capacita para construir una disyunción del tipo que acabamos de ver. ¿Cómo construir una fórmula del tipo de $\psi^{\bar{a}}_{M}$? Sea $\bar{a}$ una tupla en *M*. Para cada tupla $\bar{b}$ en *M* de la misma longitud que $\bar{a}$, si existe alguna fórmula infinita en L_B que sea satisfecha en *M* por $\bar{a}$ pero no por $\bar{b}$, entonces elegimos una fórmula de este tipo (si no existe, entonces vamos a la tupla siguiente). Al final del proceso, cuando hemos examinado todas las tuplas $\bar{b}$ en *M*, lo que obtenemos es un *conjunto* de fórmulas. (La efectividad o no de este proceso, su realizabilidad o no por parte de una mente finita, no desempeñan ningún papel en el presente contexto). Sea $\psi_{M,\bar{a}}$ la conjunción de este conjunto de fórmulas. Obviamente, $\psi^{\bar{a}}_{M}$ es satisfecha por $\bar{a}$ en *M*. Considérese cualquier tupla $\bar{b}$ en *M*. Si $\bar{a}$ y $\bar{b}$ satisfacen en *M* exactamente las mismas fórmulas infinitas en L_B, en tal caso tenemos $M \quad \psi^{\bar{a}}_{M}[\bar{b}]$ (de otro modo, un miembro de la conjunción $\psi^{\bar{a}}_{M}$ sería falso de $\bar{b}$ en *M*). Si $\bar{a}$ y $\bar{b}$ no satisfaciesen en *M* exactamente las misma fórmulas infinitarias en L_B, entonces habría (a partir de la citada construcción) al menos un miembro de la conjunción $\psi^{\bar{a}}_{M}$ que no es satisfecho por $\bar{b}$ en *M*, y consiguientemente $M \quad \psi^{\bar{a}}_{M}[\bar{b}]$. Obsérvese que $\psi^{\bar{a}}_{M}$ no necesita ser localmente maximal, de modo que el argumento de Kim referente a la reducción débil no requiere en realidad [*KPM2*]. Sin embargo, en el planteamiento original de Kim parece ser esencial el uso de propiedades (fórmulas) maximales.

Las anteriores consideraciones nos llevan a la siguiente:

Proposición 6

Sean $\Phi(\mathcal{A})$ y $\Phi(B)$ los conjuntos de todas las fórmulas infinitarias en los vocabularios $L_{\mathcal{A}}$ y L_B respectivamente y sea **K** cualquier clase de modelos para un vocabulario L, extensión que incluye $L_{\mathcal{A}}$ y L_B. Entonces $\Phi(\mathcal{A})$ superviene débilmente sobre $\Phi(B)$ con respecto a **K** syss $\Phi(\mathcal{A})$ es débilmente reducible a $\Phi(B)$ en **K**.

Procederemos ahora de manera análoga para mostrar la implicación [SF] → [CN] de acuerdo con las pautas propuestas por Kim. Nótese para empezar que, para llevar a cabo la estrategia kimeana en este caso general, no necesitamos fórmulas localmente maximales. Desafortunadamente, no podemos seguir la estrategia naïve seguida por Kim y definir una fórmula localmente maximal $\delta_{M,\bar{a}}$ de $\bar{a}$ en *M* como la siguiente conjunción:

$\{\psi(\bar{x}): M \quad \psi(\bar{x})[\bar{a}],\ \psi(\bar{x})$ es una fórmula infinita en $L_B\}$.

Trabajando con $\mathcal{L}_{\infty\omega}$ no necesitamos preocuparnos por limitaciones de cardinalidad, puesto que la construcción de esta lógica nos garantiza que para cualquier *conjunto* de fórmulas existe su conjunción. El problema no es por tanto la cardinalidad, cuanto el hecho de que la colección que viene detrás del conyuntor no forma conjunto alguno. La razón es que la colección de todas las $\mathcal{L}_{\infty\omega}$-fórmulas en un vocabulario es una *clase propia*. Es fácil ver que, como consecuencia directa de esto, la familia de fórmulas presentada en la conyunción de arriba también debe ser una clase propia. Nótese asimismo que el camino elegido en el caso infinitario es completamente inútil aquí, puesto que no podemos construir fórmulas con infinitos cuantificadores. Necesitamos un medio de superar estas dificultades. Afortunadamente, en este caso podemos encontrar de una forma canónica una L_B-fórmula $\delta_{M,\bar{a}}$ en $\mathcal{L}_{\infty\omega}$ que es equivalente a la conjunción virtual de todas las fórmulas infinitarias en L_B satisfechas por esta tupla en M. Procederemos de la manera siguiente: comenzaremos con una tupla $\bar{a} = (a_1, \ldots, a_n)$ en M a ser caracterizada y consideraremos el modelo expandido $(M, \bar{a})$. Para este modelo, construimos una (así llamada) *sentencia de Scott* $\beta_{(M,\bar{a})}$ en un vocabulario apropiadamente expandido $L_B(\bar{c})$ (donde $\bar{c} = (c_1, \ldots, c_n)$ es tupla de n nuevas constantes diferentes dos a dos) con la siguiente propiedad:

[##] para cada N de la clase K y cada tupla $\bar{b}$ (de la misma longitud que $\bar{a}$) en N: $(N, \bar{b})$ $\beta_{(M,\bar{a})}$ syss $(M, \bar{a})$ y $(N, \bar{b})$ hacen verdaderas las mismas sentencias en $L_B(\bar{c})$.

Una sentencia con esta propiedad es llamada una *sentencia de Scott* para la estructura en cuestión. Entonces definimos $\delta_{M,\bar{a}}$ como la fórmula en L_B que obtenemos cuando sustituimos en $\beta_{(M,\bar{a})}$ cada constante c_i por la variable x_i. Por [##], $\delta_{M,\bar{a}}$ es una fórmula localmente maximal satisfecha por $\bar{a}$ en M. La sentencia $\beta_{(M,\bar{a})}$ puede ser definida como la conjunción de una familia $\{\theta_{\bar{d}} : \bar{d} \text{ en } (M, \bar{a})\}$ donde $\theta_{\bar{d}}$ es la siguiente fórmula:

$$\forall \bar{x}\ \left[\psi^{\bar{d}}_{(M,\bar{a})} \rightarrow \left(\forall y \quad \{\psi^{\bar{d}e}_{(M,\bar{a})} : e \text{ en } (M, \bar{a})\} \wedge \quad \{\exists y\ \psi^{\bar{d}e}_{(M,\bar{a})} : e \text{ en } (M, \bar{a})\}\right)\right]$$

De este modo, hemos producido un análogo lingüístico de una propiedad localmente maximal. Ahora resulta tentador definir un L_B-sustituto de una fórmula $\varphi(\bar{x})$ en L_A tal y como lo hace Kim, esto es, mediante la disyunción:

$$\{\delta_{M,\bar{a}} : M \quad \varphi(\bar{x})[\bar{a}] \ \&\ \bar{a} \text{ en } M \ \&\ M \text{ en K}\}.$$

Si esta disyunción existe, es claro que es equivalente a $\varphi(\bar{x})$ en K. ¿Pero por qué debería existir esta disyunción, si K es una clase propia? Obsérvese que hay clases de estructuras K tal que la colección de sentencias de Scott de todos los elementos de K no

forma un conjunto. En tales casos no podemos construir la correspondiente disyunción. Más aún, hay también casos en los que ninguna fórmula bien formada es equivalente a cada virtual disyunción. Ilustraremos esto por medio de un ejemplo. Sea K la clase de todos los buenos órdenes. Para cada *M* de K sea β_M una sentencia de Scott para *M*. Podemos construirla en una forma canónica tal que dos elementos de K que satisfagan las mismas sentencias infinitarias tengan la misma sentencia de Scott. De acuerdo con un conocido resultado, dos buenos órdenes equivalentes con respecto a las sentencias formuladas en $\mathcal{L}_{\infty\omega}$ son ya isomorfos. De esto y del hecho de que la colección de todos los ordinales es una clase propia, concluimos que la colección de sentencias de Scott de todos los elementos de K no forma un conjunto. Por tanto, no podemos construir la correspondiente disyunción. Pero, ¿podemos acaso encontrar una fórmula que se comporte igual que la disyunción? También se puede demostrar que esto es falso. Una fórmula tal axiomatizaría la clase de todos los buenos órdenes, lo que es imposible. Estas dificultades, sin embargo, no excluyen de manera automática la posibilidad de mostrar que [SF] implique [CN]. Lo que en cualquier caso muestran es que el argumento de Kim, al menos en su forma original, no puede ser formalizado en $\mathcal{L}_{\infty\omega}$. No hay, como acabamos de ver, un principio general correcto que justifique uno de los pasos cruciales en la argumentación de Kim. Conjeturamos que la tesis de Kim es asimismo falsa. Desgraciadamente, no podemos (al menos por el momento) proporcionar un contraejemplo.

Nuestras conclusiones con respecto al uso de $\mathcal{L}_{\infty\omega}$ quedan recogidas en el siguiente esquema:

$$\begin{array}{ccccc} [\text{CN}] & \rightarrow & [\text{SF}] & \rightarrow & [\text{SG}] \\ & & \downarrow & & \\ & & [\text{SD}] & \leftrightarrow & [\text{CND}] \end{array}$$

5. Significación filosófica de los presentes resultados

Argumentos reductivos basados en operaciones infinitas, tales como el formulado por Kim, han sido aceptados por muchos autores sin cuestionar la legitimidad de asumir la existencia de propiedades maximales. La explicación modelo-teórica de los diferentes conceptos de superveniencia, así como del de reducción, muestran, sin embargo, lo interesante y fecunda que resulta la cuestión acerca de la lógica subyacente que escogemos con el fin de reconstruir dichos argumentos. Como hemos visto, es necesario el uso de la lógica infinitaria para que dichos argumentos puedan ser correctos, pero al mismo tiempo son necesarias ciertas restricciones. Los argumentos funcionan siempre y cuando trabajemos con estructuras finitas, pero, bajo una lógica infinitaria irrestricta (esto es, conyunciones y disyunciones arbitrariamente largas) *con clases arbitrarias de estructuras infinitas*, la afirmación de Sider resulta falsa, mientras que el argumento de Kim no se puede formalizar. ¿Qué consecuencia filosófica podemos extraer de esto?

Glanzberg, el único autor —al menos que sepamos nosotros— que se ha ocupado de la relevancia de la lógica infinitaria para el debate de la superveniencia, subraya algo en lo que estamos perfectamente de acuerdo, aunque valga decir que por razones distintas (Glanzberg 2001, p. 427): "The moral of the discussion above is not that infinitary logic is bad, but rather that (...) one must decide how much infinitary logic to allow".

No queremos ser categóricos, sino que dejamos abierta la puerta a las tesis antirreduccionistas. Aunque no disponemos de un contraejemplo de la tesis de Kim, el hecho de que su argumento no pueda ser formalizado en $\mathcal{L}_{\infty\omega}$ sin restricciones muestra que al menos generalmente no funciona y de que probablemente haya algo erróneo en su tesis (no sólo bajo nuestra interpretación, sino de un modo fundamental). La interpretación modelo-teórica aquí propuesta muestra cuán decisivos son, para el debate filosófico que nos ocupa, los supuestos sobre propiedades/fórmulas maximales, así como las restricciones impuestas en el sistema lógico elegido. En la presente contribución, no hemos querido defender tesis antirreduccionistas, sino sólo mostrar que ciertos argumentos/tesis reduccionistas no funcionan en general, dejando así abierta la posibilidad de dar con una tesis antirreduccionista coherente (en filosofía de la mente o en otros campos) basada en el concepto de superveniencia fuerte/global.

REFERENCIAS

Ebinghaus, H.-D. y J.Flum (1995). *Finite Model Theory*. Berlin: Springer.

Glanzberg, M. (2001). "Supervenience and Infinitary Logic", *Nous* 35, 419-439.

Gregory, J. (1974). "Beth Definability and Infinitary Logic", *Journal of Symbolic Logic* 39, 22-26.

Hellman, G., y F.W. Thompson (1975). "Physicalism: Ontology, Determination, and Reduction", *Journal of Philosophy* 72, 551-64.

Kim, J. (1993). *Supervenience and Mind*. Cambridge (Mass.): Cambridge University Press.

Polanski, M. (2005). "Stalnaker on Strong and Global Supervenience", manuscrito no publicado.

Sider, T. (1999). "Global Supervenience and Identity across Times and Worlds", *Philosophy and Phenomenological Research* 59, 913-937.

Xavier DE DONATO RODRÍGUEZ es doctor en lógica y filosofía de la ciencia por la Universidad Ludwig-Maximilian de Munich. Actualmente es investigador en estancia posdoctoral en el Instituto de Investigaciones Filosóficas de la UNAM (México). Su área principal de investigación es la filosofía de la ciencia y, en particular, la aplicación de métodos formales al estudio de las relaciones interteóricas.

DIRECCIÓN: Instituto de Investigaciones Filosóficas, Universidad Nacional Autónoma de México. Circuito Maestro Mario de la Cueva, s/n. Ciudad de la Investigación en Humanidades. Ciudad Universitaria, 04510, Coyoacán. México, D.F. E-mail: xdonato@minerva.filosoficas.unam.mx.

Marek POLANSKI es doctor en lógica y filosofía de la ciencia por la Universidad Ludwig-Maximilian de Munich. Actualmente imparte seminarios en esta misma universidad. Marek Polanski es autor de *Zur logischen Analyse von Theorienreduktion und Theorienäquivalenz*, München: Centrum für Informations- und Sprachverarbeitung, 2002. Su área principal de investigación es la lógica, así como la aplicación de métodos formales, en particular de la teoría de modelos, en filosofía de la ciencia y filosofía analítica.

DIRECCIÓN: Seminar für Philosophie, Logik, und Wissenschaftstheorie, Ludwig-Maximilians-Universität München Ludwigstr. 31, D-80539 München. E-mail: marek.polanski@web.de.

The Case against Evaluative Realism*

Dan LÓPEZ DE SA

Received: 2005.10.19
Final Version: 2006.04.12

BIBLID [0495-4548 (2006) 21: 57; pp. 277-294]

ABSTRACT: In this paper I offer a characterization of evaluative realism, present the intuitive case against it, and offer two considerations to support it further: one concerning the internalist connection between values and motivation, and the other concerning the intuitive causal inefficacy of evaluative properties. The considerations ultimately rely on the former intuitions themselves, but are not devoid of interest, as they might make one revise what one took to be his own realistic supporting intuitions, if such one had.

Key words: evaluative realism, flexibility, metaethics, internalism, causal efficacy.

In this paper I want to present a case against evaluative realism. The considerations I will submit will not constitute a refutation of it, given that in the central points they dwell on intuitions that, if sound, would support rejecting realism quite directly, with the result that as arguments they might be accused of begging the question. For better or for worse, I think that no stronger case against (nor for) evaluative realism is forthcoming. But this does not make the considerations worthless, I hope, for they make explicit some of the consequences of the realist approach. To the extent to which one regards them as counterintuitive, the considerations may eventually make one revise one's judgments about what one took to be one's own relevant intuitions.

The paper is divided into six sections. In the first section I focus on the target: taking some earlier work of mine as my starting-point, I propose to characterize evaluative realism as rejecting what I call the flexibility of values, in contrast to other proposals that make evaluative realism either too easy or too hard. In the second section I present the particular flexible account of values I would favor, which is mainly due to David Lewis, and the scenarios whose intuitive description (I take it) strongly favor this flexibility, which are variants of the "Moral Twin Earth" submitted to related aims by Terence Horgan and Mark Timmons. People quite often claim, nonetheless, that

* Earlier versions were presented at the LOGOS GRG (Barcelona, 2003) and at the *VI Taller d'Investigació en Filosofia* (Tarragona, 2004). Thanks to the audiences in both occasions, and especially to Oscar Cabaco, Néstor Casado, Gemma Celestino, José A. Díez, Sisco Gris, Pablo Rychter, Achim Spelten and Mike Wilms, and also to Agustín Arrieta, Josep Corbí, Esa Díaz-León, Manuel García-Carpintero, Kevin Mulligan, Ekai Txapartegi, Agustín Vicente and an anonymous referee, for very stimulating discussion and objections. Research has been partially funded by the research projects BFF2002-10164 (European Science Foundation EUROCORES programme "The Origin of Man, Language and Languages") and BFF2003-08335-C03-03 (MCyT, Spanish Government), and the research group 2001SGR00018 (Dursi, Catalan Government). Thanks also to Mike Maudsley for his linguistic revision.

they do not share the relevant flexibility supporting intuitions. The two main considerations I will offer aim to urge them to revise what they take to be their own intuitions. In section three I will claim that evaluative realism, including the dispositional variety of it, cannot account for internalism about values. In section four I will also consider the somewhat trickier case of internalism about value-judgments. In section five I will present the Missing Explanation Argument, due to Mark Johnston, to the effect that flexible properties cannot be involved in causally explaining general dispositions of subjects to respond in certain ways, and I will claim that evaluative properties intuitively do not appear in such explanations. And finally in section six I will consider why, on the face of it, this is compatible with the views about so-called moral explanations of philosophers like Nicholas Sturgeon.

1. Evaluative Realism vs. the Flexibility of Values

The diversity of views intended under the label of "realism" is in my view particularly acute with regard to realism about evaluative properties. Before presenting the one I will use, I want to briefly mention some alternatives that, in my view, make evaluative realism either too easy or too hard.

Consider for instance what is offered by Geoffrey Sayre-McCord:

> Realism involves embracing just two theses: (1) that claims in question, when literally construed, are literally true or false (cognitivism), and (2) some are literally true. Nothing more. (1988b, p. 5)

I think it should be clear that realism so conceived will be a quite uncontroversial position. To illustrate, consider a caricature-like subjectivist account of values, having it that something like the following defines being good:

> *x* is good iff we value *x*.

False as it might be for other reasons, the proposal does satisfy (1) and (2) and hence would be a realist proposal conceived in this way. But if this counts as realist, almost any possible view would as well.[1] Adding an epistemic element of the sort

> It is possible to find out about some moral sentences that they are true. (Thomson 1998a, p. 171)

does not seem to change the situation, since on occasions we can clearly find out what we value.

So it seems that one might have a non-realist approach to evaluative properties that respects that instantiations of them make simple predications of predicates signifying them straightforwardly, and sometimes knowably, true. On the other side, and, I

[1] I don't mean to suggest that Sayre-McCord is unaware of this, quite the contrary: he explicitly considers various possible "subjectivist" positions as varieties of realism, see (Sayre-McCord 1988b, pp. 16-9). It is only in this "cheap" sense, I take it, that Lewis himself describes his position as a realist one: values as he conceives them "do exist", see (Lewis 1989, p. 93). For a similar view, consider Jackson: "*Realists* [are] cognitivists [who have it that the statements in question are semantically truth-apt] who take the extra step of holding that the ethical properties are instantiated" (Jackson 1998, p. 128).

take it, motivated by considerations like those just submitted, David Brink says the following:

> A moral realist thinks that moral claims should be construed literally; there are moral facts and true moral propositions. Ethics is objective, then, insofar as it concerns matters of fact and insofar as moral claims can be true or false (and some of them are true). But moral realism claims that ethics is objective in another sense, which is not always distinguished, from this first kind of objectivity. Not only does ethics concerns matter of fact, it concerns facts that hold independently of anyone's belief about what is right or wrong. This first kind of objectivity distinguishes moral realist and other cognitivist theories form nihilism and noncognitivism; the second kind of objectivity distinguishes moral realism from constructivist version of cognitivism. (1989, p. 20)

As this is worded, though, it might seem that it also allows our caricature-like subjectivist above to count as a realist, as according to her the relevant responses on which goodness depends were not anyone's "belief about what is right or wrong" but rather a given connative attitude: *valuing*. But let us interpret Brink more liberally, as holding that evaluative realism requires that evaluative properties have essences that are independent of relevant subjective mental responses, regardless of whether they are doxastic or not. So understood, it would certainly exclude our subjectivist. But the problem now is that arguably it would exclude too much. There is a sense in which dispositions have natures that are not independent on their manifestations: dispositions can be possessed when the manifestation does not occur, to be sure, but their relation to them is part of their essences, of what makes them the properties they are.[2] Take, for instance, dispositionalism about colors. According to the view, colors are dispositions to produce in certain subjects, say, normal human perceivers as they actually are, certain mental responses, say, the experience of a certain color being instantiated, under certain conditions, say, normal viewing conditions as they actually are. Hence colors have natures that involve mental responses. This, one may say, makes a difference with respect to the alternative so-called *primary view* about colors: according to dispositionalism colors are less than fully objective properties, but this is not so according to the primary view. But both views arguably are, and are certainly taken to be, varieties of *realism* about colors. *Mutatis mutandis*, one should expect, for the case of values: a view according to which values are fully objective properties, whose natures are independent of any mental subjective response, should certainly count as a form of evaluative realism. I will refer to such a view as *(evaluative) objectivism*. But realism should not require objectivism by definition.

It is worth noticing that arguably both for primary, fully objective properties, and for secondary, real but dispositional, properties, broadly conceived Fregean considerations require that there should be some descriptive material that fixes that they are signified by certain expressions and concepts. And in the case of colors, they arguably involve precisely the relevant chromatic subjective responses. For reasons that are familiar from Kripke (1980), this suffices for the following to be not only true but also *a priori*

---

[2] See Fine (1994) for an elaboration of the view on essences I am relying on here, and García-Carpintero (2002) for the application to the distinction dispositional vs. categorical.

> *x* is red iff *x* is disposed to produce in normal human perceivers an experience as of red in normal viewing conditions

even if only contingently true. According to these realist views about colors, you only get something that holds *necessarily* by rigidifying on the relevant expressions, as in

> *x* is red iff *x* is disposed to produce in normal human perceivers *as they actually are* an experience as of red in normal viewing conditions *as they actually are*.

That is something that both "primarists" and dispositionalists can, and do, hold. As suggested, the difference between them seems to lie in whether they hold that the former holds *in virtue of the nature of the color* or not, see García-Carpintero (2002) and Wedgwood (1998) —and in my view, to settle this question, *a posteriori* considerations provided by the specialist are required.

I think that something like this is precisely characteristic of *realism* about colors, and this is what I propose to generalize. Let me say then that if **F** is a property, an *rd biconditional* for (a predicate signifying) it is a substantial biconditional of the form:

> *x* is *f* iff *x* has the disposition to produce in subjects *S* the mental response *R* under conditions *C*

or the form

> *x* is *f* iff subjects *S* have the disposition to issue the *x*-directed mental response *R* under conditions *C*

where 'is *f*' signifies **F**, and 'substantial' is there to avoid "whatever-it-takes" specifications of either *S*, *R* or *C*.[3]

Let me also say that a specification of the subjects in an rd biconditional is *rigid* iff the relevant predicate involved in the specification is rigid,[4] and *flexible* otherwise. Take for instance 'normal human perceiver.' This is not, as it stands, a rigid specification. For take the relevant predicate 'is a normal human perceiver' and suppose that in the actual world, it is true (even if knowable only *a posteriori*) that being such is being a

---

[3] One such "whatever-it-takes" specification of, say, subjects S would be "those subjects, however they be, such that something is disposed to produce in them responses R under conditions C iff it is F." *Mutatis mutandis* for the responses and the conditions.

[4] I am assuming, with Kripke (1980), and a lot of people in discussions on philosophy of mind, philosophy of science or metaethics, that the notion of rigidity might be extended to be applicable to predicates, roughly along the lines of: a predicate is rigid iff it signifies the same property in all relevant worlds. Proposals like this have recently received criticisms, among which: that it would trivialize, making *all* predicates trivially rigid (see for instance Soames 2002), and that in any case it would overgeneralize, counting as rigid predicates some that do not signify natural properties/kinds (see for instance Schwartz 2002). I try to respond to these criticisms, respectively, in my unpublished 'Rigidity for Predicates and the Trivialization Problem' and 'The Over-Generalization Problem: Predicates Rigidly Signifying the "Unnatural."' In the latter I also argue that the relevant simple predicates like those that will concern us here, 'is red,' 'is funny,' 'is good' and the like are, nonetheless, rigid. Given this I will speak of them *signifying properties*, without relativizing such talk to worlds.

human with a perceptual apparatus meeting condition *ABC*. Now consider a counterfactual situation in which, for whatever reason you might think of, the human perceivers that are *normal there* are those with a perceptual apparatus meeting the different condition *DEF*. Now intuitively, it is this other property of being a human with a perceptual apparatus meeting condition *DEF* which would be relevant for evaluating sentences containing 'is a normal human perceiver' with respect to this other world. But then 'is a normal human perceiver' is not a rigid predicate, but a flexible one. Its relevant rigidification, which can be put as something like 'is a normal human perceiver *as they actually are*' leads nonetheless to a rigid specification of the subjects, of the sort 'normal human perceivers *as they actually are*.'

An rd biconditional is *rigid* iff it involves a rigid specification of the subjects, and is *flexible* otherwise. Finally, a given property is *flexible* iff there is a flexible rd biconditional for (a predicate signifying) it which holds (*a priori* and) in virtue of its nature and hence necessarily.[5]

With all these stipulations I can state my proposal about realism thus:

> A property is *real* iff it is not a flexible property.

What considerations would be relevant for the issue as to whether a given predicate signifies a real vs. a flexible property?

Suppose that 'is *f*' signifies[6] property **F**, and suppose that *S* and *C* are the relevant flexible specifications of subjects and conditions, and $S_@$ and $C_@$ their relevant rigidifications, and that the only relevant rd biconditionals are

(R) *x* is *f* iff *x* is disposed to produce in $S_@$ the response *R* under conditions $C_@$.

(F) *x* is *f* iff *x* is disposed to produce in *S* the response *R* under conditions *C*.

Both are, we may suppose, true with respect to the actual world and, we may also suppose, *a priori* knowably so. But the following asymmetry arises; abstracting now from issues about essence vs. necessity, their metaphysical status covaries with the nature of **F** as stated in

F is real	iff (R) is necessary	iff (F) is contingent
F is flexible	iff (R) is contingent	iff (F) is necessary.

This provides a way of testing whether 'is *f*' signifies a real or a flexible property, and based just on *a priori* considerations. The recipe is, very abstractly put, this: consider what could be a counterexample of the necessity of the relevant statement on the assumption that the predicate signifies one particular kind of property. I will refer to them as *target situations*. Then check how these should be intuitively described (with respect to the relevant predicate) and conclude accordingly.

---

[5] This is the notion labeled *flexible response-dependence* in López de Sa (2003). I am abstracting here from issues related to response-dependence.

[6] See footnote 4 above.

I will instantiate this sort of relevant consideration in the next section. But let me end this one with the following remark about words. The question of the appropriateness of labels *per se*, when philosophical terms are at issue, does not appear to be particularly interesting philosophically, *once* the relevant distinctions are clear and attended to. There certainly seems to be a contrast between entirely objective properties *and* dispositional properties, *on the one side*, and flexible properties, *on the other*, as issued in the question of how the relevant target situations should be intuitively described. My aim here is to present a case against the view that evaluative properties are *of the former kind*, whatever they are called. As I said, though, I will call them *real properties*.

2. A Flexible Lewisian Theory of Values and the Intuitions about Evaluative Twin Earth

That some evaluative properties are intuitively flexible is, I take it, quite uncontroversial. Consider the case of 'is funny.' Suppose that the following are the relevant flexible and rigidified rd biconditionals

> *x* is funny iff *x* is disposed to amuse us under appropriately attentive conditions.
>
> *x* is funny iff *x* is disposed to amuse us as we actually are under appropriately attentive conditions as they actually are.

Now take something funny, even something, as I am ready and willing to grant, *really really* funny, like *The Simpsons*. Gerald Lang suggests that we would not take very seriously the suggestion that it "would continue to be funny even if a comprehensive alteration in our comic sensibilities took place" (Lang 2001, p. 201). That is, in a very compressed form, an instance of the relevant consideration we have just considered, to the effect that being funny is flexible and not real. As there is no doubt that *The Simpsons is* actually funny, there is no doubt that it is disposed to amuse us as we actually are under appropriately attentive conditions as they actually are. Consider now a relevant counterfactual target situation, *w*, in which this alteration of our sensibilities takes place, but which, apart from this, resembles the actual world as much as possible. *The Simpsons* is *not* disposed to amuse us *as we would be in w* under appropriate attentive conditions *as they would be in w*.

So far we have the relevant target situation, appropriately neutrally described, as no hypothesis about the extension of 'is funny' *with respect to w* is introduced. Hence, that it is a possibility is something agreeable by *both* defenders of the view that 'is funny' signifies a real property and defenders of the view that it signifies a flexible one. The crucial question is now: how should it be intuitively described with respect to 'is funny'? In particular, is it true or false, intuitively, that '*The Simpsons* is funny' *when evaluated with respect to w*? Lang says that we would not even take seriously the suggestion that it might be true. But now, if '*The Simpsons* is funny' is false with respect to *w* and '*The Simpsons* is disposed to amuse us as we actually are under appropriately attentive conditions, as they actually are' is true with respect to *w*, the rigidified biconditional is only contingently true with respect to the actual world. Hence, 'is funny' signifies a flexible property, and not a real one.

One might say: "But we do say, at least sometimes, that *The Simpsons is* funny, in the objective mood, as it were, rather that *we find them funny*. Furthermore, we say those things even acknowledging that they may not amuse some people, for after all some days, although funny, they don't even amuse us. Why couldn't we say then that *The Simpsons* are really funny even in the target situation, only that those unlucky people fail to be disposed to be amused by them?" The straight answer is that we *could* definitely say this: it's only that intuitively we, or at least most of us, don't want to. Remember that the crucial issue is how a given target situation should be *intuitively described*. In the submitted consideration, there is also another important element which is worth stressing to avoid possible misunderstandings. The fact that we have simple predicates like 'is funny' signifying the property of being funny arguably entails that there should be a "real"/appearance distinction concerning what is funny, that being funny should be distinct from seeming funny or actually amusing. But that of course is also the case *even if funny is a flexible property*, and hence in particular does not entail anything about what the proper intuitive description of target situations should be. There are things which seem funny even though they are not really funny at all (see (Wright 1992, p. 101) for a dozen examples of this) and conversely, as submitted, *The Simpsons* are funny even if they sometimes fail to seem so. But that is indeed entailed by the use of the *dispositional* idiom in the rd biconditionals. Dispositions can be possessed without issuing their characteristic manifestations. And conversely, their manifestation could occur without being the manifestation of a possessed disposition. Flexible properties are not dispositions, true enough. But with respect to each world, the things that have a given flexible property in this world are those that are disposed to produce the relevant response in the subjects as they are in that world under the conditions as they are in that world. Hence, in each world, having the property, being funny, is not the same as issuing the relevant response, seeming funny.

The same situation occurs, I claim, for a number of similar *soft* evaluative predicates: 'is tasty', 'is disgusting', 'is comfortable', not to mention 'is sexy,' 'is fashionable', or 'is cool.' With respect to any of these, it seems, hardly anyone would claim to have the intuitions supporting their signifying real properties. Does it generalize with respect to *all* evaluative predicates, including the *hard* cases of moral and some aesthetic predicates? Consider the following general rd biconditional, adapted from the proposal by David Lewis in his 'Dispositional Theories of Value' (1989):

x is good iff we are disposed to value x in appropriate reflective conditions.

Some remarks are in order. First, *valuing* is the favorable attitude of desiring to desire. That valuing is a desiderative rather than a doxastic attitude is arguably entailed by its being a *favorable* attitude. But "first order" desiring would certainly *not* do: we, unfortunately quite often, desire things we do not value at all. Weakness of will is, of course, a case at hand. Take "unwilling smokers," as one might call them, like myself. I desire to smoke a *Ducados* quite often, I actually *love* smoking. But I find some uneasiness even in *reporting* it as I have just done. It is not, or at least not only, that I have a contrasting desire *not to* smoke: that would be a case of conflicting desires —which by the way

could eventually issue in conflicting valuings or even in moral dilemmas. But phenomenologically, my case of smoking is not, or at least is not only, constituted by what I experience when for instance I have contrasting desires about enjoying a good film this afternoon or remaining in my office finishing this paper. In this case I do not want to be rid of either desire: I would prefer the world to be so that they could both be satisfied, but unfortunately I will have to act upon only one of them. My smoking is different: I *do* want to be rid of my desirings to smoke: even if I desire to smoke, I desire not to desire to smoke at all. So that, when I light I cigarette, I'd say that my will is weak, given that I desire not to have the desire that makes me do so. So failing to desire as one values is failing to desire as one desires to desire. Hence, it seems, valuing is desiring to desire.[7]

Second, *we* are, according to the proposal, those that are disposed to value, with respect to the relevant particular issue at stake, exactly like the speaker. It is important to stress that, so understood, 'we' turns out to be a *flexible* characterization of a group of subjects. The relevant predicate signifies with respect to the actual world the property of being relevantly the way I am *actually*. But I could be otherwise, and in particular my disposition to value could be very different from what it actually is. But then, with respect to those worlds in which I am suitably different, it will signify the property of being relevantly the way I *would be* in those situations.

Third and finally, *appropriate reflective conditions* are rather schematic. In Lewis' original paper, he submits that the relevant conditions are the conditions of fullest possible imaginative acquaintance with the thing in question, *possible* for the subjects in question and relatively to the thing in question (Lewis 1989, pp. 77-9). This element has met with some resistance in the literature (see for instance Johnston (1989) and Smith (1994)). Some would claim that further elements should be included, notably awareness of all (non-evaluative) relevant facts. I tend to agree with Lewis that this further element, crucial for the question of *balancing* different probably conflicting values, should not be included in the conditions determining the values to be balanced in the first place (see Lewis 1989, pp. 79-82). But the issue is delicate, and I would rather not go into it here. My proposal is then to characterize the relevant conditions as *the appropriately reflective conditions*, having in mind these Lewisian conditions of fullest possible imaginative acquaintance, but perhaps also some others like the ones considered if further thought renders them appropriate.

Let me come back to the issue at hand. It would seem that if 'is good' signifies a flexible property then arguably all evaluative predicates do so as well.[8] Does it? As we

---

[7] In the meantime, and fortunately, I quit. Fur further discussion of the objections against the sufficiency of desiring to desire for valuing, and of its necessity, see my unpublished 'What is Valuing?,' where I try to show how the Lewisian proposal should be properly understood, or otherwise amended: valuing is desired desiring, or perhaps merely desiring one does not desire against.

[8] This is straightforward if one characterizes, as I am inclined to do, *evaluative* predicates as those that suffice for 'is good' (or 'is bad'). One should expect the claim also to hold, I imagine, in some alternative formulation is adopted.

have seen, settling this depends on the status of the relevant flexible and rigid biconditionals

> *x* is good iff we are disposed to value *x* in appropriate reflective conditions.
>
> *x* is good iff we, as we actually are, are disposed to value *x* in appropriate reflective conditions, as they actually are.

which in turn depends on what turns out to be the intuitively proper description of suitably neutrally described counterfactual target situations. I claim that one of these is the generalized version of the Moral Twin Earth submitted in related contexts and for related aims by Terence Horgan and Mark Timmons,[9] which I will call *Evaluative Twin Earth* or *ETE* for short.

Take something that I am —and hence *we* are— *actually* disposed to value under appropriate reflective conditions: (say) Santi's lying to me on some particular occasion. We, as we actually are, are disposed to value Santi's lying to me under appropriate reflective conditions, as they actually are, and hence it is —actually— good. But I could be different. In particular my dispositions to value this particular lie under those relevant conditions could be suitably more "deontologist," as it were. I could be such that I am not disposed to value it under appropriate reflective conditions. So let us consider a situation in which I *am* like that, but agrees with the actual situation in as much as it's possible compatibly with this difference, and call it ETE. We, as we are in ETE, are not disposed to value Santi's lying to me under appropriate reflective conditions.

So far, again, we have the relevant target situation, appropriately neutrally described, as no hypothesis about the extension of 'is good' *with respect to ETE* is introduced. Hence, that it is a possibility is something agreeable by *both* defenders of the view that 'is good' signifies a real property and defenders of the view that it signifies a flexible one.[10] The crucial question is again: how should it be intuitively described with respect to 'is good'? In particular, is it true or false, intuitively, 'Santi's lying to me is good' *when evaluated with respect to ETE*? My own intuitions, and as I understand him, Lewis' also, are that it should be *false* with respect to ETE. But then, if 'Santi's lying to me is good' is false with respect to *ETE* and 'We, as we actually are, are disposed to value Santi's lying to me under appropriate reflective conditions, as they actually are' is true with respect to *ETE*, the rigidified biconditional is only contingently true with respect to the actual world. Hence, 'is good' signifies a flexible property, and not a real one. Hence, arguably all evaluative predicates, and not only soft ones, do so as well:

---

[9] See Horgan & Timmons (1991), (1992a), (1992b), (1996) & (2000a). Their own positive views might differ from mine, though. For their views see *inter alia* Horgan & Timmons (2000b).

[10] It is worth emphasizing that the availability of the Evaluate Twin Earth *per se* merely depends the relevant psychological facts being contingent and hence that it is a possibility is something that all the disputants have reason to accept. As we are about to see, how is it to be intuitively described in terms of the predicate 'is good' is what would settle the question as to whether the predicate signifies a property of one or the other kind. I am indebt here to an anonymous referee for this journal.

the intuitive flexibility of values is then vindicated and thereby evaluative realism is rendered unintuitive.[11]

As I said at the beginning, people quite often claim, nonetheless, that they do not share the relevant flexibility supporting intuitions. As I tend to think that they *do* have them after all, the aim of this paper is to offer two considerations in the light of which some might revise what they took to be their own realist supporting intuitions when the proper description of the ETE is concerned. These considerations exploit what I take to be counterintuitive consequences of the realist alternative. If people initially claiming that they do not share the relevant flexibility supporting intuitions also find those consequences counterintuitive, that would provide them with reasons for revising what they took to be their own realist supporting intuitions when the proper description of the ETE is concerned.[12] Of course some realist would be ready to bit the bullets, and hence the considerations cannot constitute a refutation of the alternative, realist, approach to values.

3. Internalism vs. Evaluative Realism

John Mackie famously once developed an argument from queerness against there actually being objective goods, where:

> An objective good would be sought by anyone who was acquainted with it, not because of any contingent fact of this person, or every person, is so constituted that he desires this end, but just because the end has to-be-pursuedness somehow built into it. (Mackie 1977, p. 112)

Here he is pointing to what it is sometimes called the *practicality* of the evaluative or *internalism*, roughly: values, whatever they are, have a to-be-pursuedness somehow built into them. That certainly seems something *constitutive* of values as we conceive them. So it is according to Lewis:

> If something is a value, and if someone is of the appropriate 'we', and if he is in ideal conditions, then it follows that he will value it. And if he values it, and if he desires as he desire to desire, then he will desire it. And if he desires it, and his desire is not outweighed by other conflicting desires, and if he has instrumental rationality to do what serves his desires according to his beliefs, then he will pursue it. And if the relevant beliefs are near enough true, then he will pursue it as effectively as possible. A conceptual connection between value and motivation. But a multifariously iffy connection. Nothing less iffy would be credible. But still less it is credible that there is no connection at all. (Lewis 1989, p. 72)

I propose to state this internalist claim about values thus:

(I) It is necessary and *a priori* that: If something is good, we would desire it under appropriate reflective conditions (weakness of will and the like aside).

---

[11] For further details and discussion see López de Sa (2003).

[12] Notice that, for the strategy to be successful, some of those claiming that do not share the flexibility supporting intuitions should find the consequences counterintuitive without being antecedently ready to reject realists views on the matter. And, in my own experience, some *do* so find them. Hence the considerations are not worthless. I am indebt here to an anonymous referee for this journal.

Evaluative realism cannot account for (I). The reason is straightforward: realism entails that the relevant flexible rd biconditional would be at most contingently true. But any counterexample to its necessity is such that the embedded conditional in (I) would be false with respect to it. Hence it would *not* be necessary with respect to the actual world, and hence (I) is false.

Evaluative objectivists, who hold that evaluative properties are fully objective, typically agree and even emphasize this, but then give (I) up and go externalist. They usually claim that that is indeed a virtue of their position, given that the externalist component is independently motivated. Some of them stress what is an undeniable fact: that sometimes people fail to be moved by what is good, even by what they know is good. That would challenge a strengthened version of (I) having it that values *directly* motivate the relevant subjects by directly issuing in them the relevant desire. That would be, I agree, as a matter of fact *not true*, let alone necessarily and *a priori* so: we have already considered cases of weakness of will in which we fail to desire as we value. These by itself would refute the strengthened version of (I). But (I) is suitably weaker, not only on that score, but importantly in requiring that one should value the good *only* under certain, appropriately reflective, conditions. So in order to refute (I) you will need a case of something which is good but such that the appropriate valuers don't desire it at all, *not even* under the appropriate reflective conditions and when their will is strong enough to desire as they desire to desire. But this seems quite a hard thing to have. This case, one is inclined to say with Lewis, is simply not credible.

Someone like David Brink would agree with a lot of this, although he would put it the other way round, as it were:

> [T]he internalist cannot rest content with the extensional claim that everyone is in fact motivated [by what is morally good]. Any externalist could claim that. The internalist about motives claims that it is true in virtue of the concept of morality that [moral goodness] necessarily motivate. According to the internalist, then, it must be conceptually impossible for someone to [know that something is morally good] and remain unmoved. This fact raises a problem for internalism: internalism makes the amoralist conceptually impossible. (Brink 1986, pp. 29-30)[13]

The dialectical situation is weird enough, though, for the conceptual impossibility of such an amoralist, who is not at all disposed to desire something that is good, even under appropriate reflective conditions and with a strong enough will, far from raising a problem for internalism is precisely what motivates it. The reason for (I) can be put by the thought that such an amoralist *is* conceptually *im*possible.

Do we have here an irremovable clash of intuitions? This could be the case, of course. But I take it to be dialectically fruitful enough, for as I said some realists *do* indeed seem to appeal to (I) in rejecting objectivism and to claim instead that evaluative (and moral) properties are, though real, somehow more *subjective* by being essentially tied to (evaluative) responses, in the same way as colors are according to the dispositionalist. But this move is unsuccessful.

---

[13] In the original passage, instead of the inserted claims about moral *values* Brink has claims about moral *considerations* and judgments, but I take it that he would certainly, even readily, concur with what I say about the properties and facts. I'll consider internalism concerning judgments in the next section 14.

In his response to Mackie, John McDowell took an "analogist" line of this kind, by arguing that the model for real evaluative properties should not be looked for in the case of primary qualities, as Mackie did, but in the case of *secondary* qualities:

> [I]t seems impossible —at least on reflection— to take seriously the idea of something that is like a primary quality in being simple *there*, independently of human sensibility, but is nevertheless (not conditionally on contingencies about human sensibility) such as to elicit some 'attitude' or state of will from someone who becomes aware of it. (McDowell 1985, p. 111)
>
> Shifting to a secondary-quality analogy renders irrelevant any worry about how something that is brutely *there* could nevertheless stand in an internal relation to some exercise of human sensibilities. Values are not brutely there —not there independently of our sensibility— any more than colours are: thought, as were colours, this does not stop us supposing that they are there independently of any particular apparent experience of them. (McDowell 1985, p. 120)

I don't want to go here into McDowell's specific views concerning values —nor colors, for that matter. Rather, I want to claim that to the extent that one tries to accommodate (I) by claiming that values are real even if not fully objective properties, but rather dispositions to produce certain evaluative response in (rigidly) specified subjects under (rigidly specified) conditions; to that extent the attempt fails. For dispositionalists about value do not deal with (I) any more effectively than objectivists did (as has been also explicitly emphasized with respect to the original Moral Twin Earth by Holland (2001)). And this is so given that the previous remark about the incompatibility of realism and (I) did not appeal to any specific view about the nature of being good besides the assumption that it was a real property and, hence, applies in particular to the relevant, secondary, dispositions.

The dispositionalist about values can of course at this point simply deny that (I) is true, and try to be comforted (say) with the *a priori* component of it, as we have seen evaluative objectivists do. The issue as to whether internalism about values is right or not depends on exactly the same intuitions that would support more directly the reality of the flexibility of values. Hence this is not an independent consideration for settling the issue. But given that, as we have seen, some people mistakenly think that they can accommodate internalism about values within a realist framework, the consideration is worth making, as it is capable of making some revise what they took to be their own realist supporting intuitions.[14]

4. Evaluative Judgment and Motivation

Internalism in meta-ethics is sometimes intended as a related, though distinct, claim asserting an *a priori* and necessary connection between evaluative (moral) *judgment* and motivation. According to the flexible account of section 2, there is such a connection. Lewis says of it that

> it is even iffier that the connection between value itself and motivation; and again I say that if it were less iffy, it would be less credible. If someone believes that something is a value, and if he

[14] I elaborate on the dilemma against McDowell in López de Sa (2006), and he responds in McDowell (2006). I hope to discuss the issue further elsewhere.

> has come to this belief by the canonical method [of putting himself in ideal conditions and finding whether he values it], and if he has remained in ideal conditions afterward or else retained the desire to desire that he had when in ideal conditions, then it follows that he values that thing. And if he desires as he desires to desire, then he desires that thing; and so on as before. (Lewis 1989, p. 73)

One could here wonder whether it is really true that were it less iffy, it would be less credible. For the belief one reaches by the canonical method, if it includes *succeeding* in achieving the relevant conditions, would indeed constitute evaluative (moral) *knowledge*. But as it is sometimes stressed, the conceptual connection between evaluative judgment and motivation seems to be independent of whether the judgment is *in fact* true: false beliefs about what is good could motivate just as much as true ones (see for instance Dreier (1990). But the canonical method might not be interpreted as necessarily successful: it is sufficient that one reaches what one *takes to be* the relevant conditions. So we have the following:

> If someone believes that something is a value, and if he has come to this belief by the canonical method of putting himself in *what he takes to be* ideal conditions and finding whether he values it, and if he has remained in *what he takes to be* ideal conditions afterward or else retained the desire to desire that he had when in *what he takes to be* ideal conditions, then it follows that he values that thing. And if he desires as he desires to desire, then he desires that thing.

That is so even if he is not right in what he takes to be the relevant conditions, and hence, even if one's belief is in fact not true. Now, to the extent that one *typically* forms one's evaluative judgment by trying to approximate the canonical method, one's judgment typically entails that one is disposed to desire it, under appropriate reflective conditions (weakness of will aside).[15] But even if one typically does it, one need not:

> If someone reached the same judgement in some non-canonical way —as he might— that would imply nothing about his valuing or desiring or pursuing. (Lewis 1989, p. 73)

But this, it seems to me, accords pretty well with the common-sense view.

Can the realist account at least for this internalism about evaluative judgement? Brink thinks not. I tend to think he is right, although arguing for such a further incompatibility would involve some complications.[16] In any case, the consideration that I wanted to offer was the previous one.

5. The Intuitively Properly Missing Evaluative Explanations

Mark Johnston has recently argued against the view that colors and other manifest properties are *response-dependent*, when a property is response-dependent in his terms iff

---

[15] If I understand them right, this is close to what is argued in Jackson & Pettit (1995), see also Jackson (1998).

[16] One in my view plausible sufficient condition would be what some philosophers have argued was right in verificationism: for a family of properties like evaluative ones it should be possible to determine sometimes that some of them are instantiated.

it is a flexible property in mine.[17] Abstracting from the details, it runs more or less thus: the idea that some properties are perceptible requires "receptivity", that there should be causal explanations of the general dispositions of the subjects to elicit the responses under the conditions *in terms of those properties*. But those explanations would go missing if the properties were flexible: satisfying receptivity entails that the relevant flexible rd biconditionals are *contingent*. Hence the label *Missing Explanation Argument* or *MEA* for short.

This provides in my view a further consideration that could make one revise what one took to be one's own realist intuitions in the evaluative case. According to the *MEA*, if a property is flexible, there will certainly be "deep" causal explanations of the general dispositions of subjects, an explanation that will appeal to certain real properties that unify the relevant instances in the actual world, but those will not appeal to the flexible properties themselves. But it is precisely this that intuitively seems to occur with respect to evaluative properties. Take a soft case. Our general disposition, as we actually are, to be amused by some things in appropriately attentive conditions, as they actually are, will certainly have casual explanations in terms of real properties: perhaps we are actually disposed to be amused by some things *because* they make us expect a connection between ideas that we know are not so connected. But intuitively we would not offer *as a causal explanation* of our dispositions to be amused that things *are funny*. And *mutatis mutandis* for the general case: one should expect there to be a complicated causal explanation of why it is that we are actually disposed to value certain things and not others in the conditions, but intuitively it would not do as a causal explanation *that they are good*.

As before, this consideration again falls short of constituting a full-blooded *argument* against evaluative realism. As an argument it could be seen as presupposing that

---

[17] His characterization of response-dependence is:

> [A] property, Being F, is response-dependent if there is some predicate 'is f' which expresses the property (i.e., whose extension across possible worlds is just the things which have the property) such that some substantial way of filling out 'R', 'S' and 'C' makes
>
> > x is f if and only if x is disposed to produce x-directed response R in all actual and possible subjects S under conditions C
>
> *a priori* and necessary; (Johnston 1998, p. 9)

once it is required

> that the canonical biconditionals are not merely superficial necessities produced by "rigidifying" on a relation that is itself contingent. The equivalence "x=Neptune if and only if x = the planet in the actual world which causes perturbations in the orbit of Uranus" is superficially necessary in this way. (Johnston 1998, p. 10)

That the proper target of the argument are response-dependent properties so understood and hence not the views that most people submit under the label of response-dependent accounts of colors — dispositionalist theories of colors— is something I stress in my unpublished 'The Explanations that are Missed according to the Missing Explanation Argument.' This, acknowledged by Johnston himself (1998, 37), is rightly emphasized by Haukioja (2000, 109), but apparently has escaped other critics, like López de Sa (2000) and Miller (2001).

the relevant causal explanatoriness of the property in question vis-à-vis the relevant responses is a *necessary* condition for its reality, in the sense I am using the notion. Now, the evaluative realist could complain, one could grant that it would be a sufficient condition, and one could even grant, as occurred in the premises of the MEAs, that concerning colors, or any other kind of *perceptible* property, it *is* a necessary condition. But why should it be in general? In particular, why should it be the case that for evaluative properties to be real they must be causally explanatory vis-à-vis the relevant responses in the way envisaged in which they intuitively aren't? That is, as I understand it, the content of Nagel's complaint (quoted in Sturgeon 1985, p. 235):

> it begs the question to assume that *explanatory* necessity is the test of reality in this area. (Nagel 1980, p. 114)[18]

Fair enough, I'm inclined to acknowledge. But as with the previous issue concerning internalism, I take it that inasmuch as reflection upon the tension between evaluative realism and internalism about values could make one revise what one took to be realist supporting intuitions concerning the proper intuitive descriptions of the evaluative target situations, reflection upon the present issue about causal explanatoriness could oblige one to make a similar revision. This being so, and even if it falls short of constituting an argument against evaluative realism, I hope the consideration is not devoid of interest.

6. Revisiting Evaluative Explanations

I have just suggested that some evaluative realists explicitly endorse the —counter-intuitive, as I take them to be— consequences of their views: externalism and causal explanatory impotence. The latter may come as something of a surprise, in that some other evaluative realists, notably Nicholas Sturgeon, are usually seen precisely as defending that there *are* moral explanations of the sort that I claim evaluative realists and anti-realists alike acknowledge that intuitively there aren't. In this section I want to defend that this impression concerning Sturgeon does not stand up to a closer analysis, and that his arguments are not incompatible with what I have been claiming so far.

In his classic paper 'Moral Explanations' (1985), Sturgeon aims to rebut a claim he attributes to Harman,[19] according to which "*even if* we assumed the existence of moral facts they would still appear explanatorily irrelevant" (Sturgeon 1985, p. 237), for discussing which, as he observes, he is free to, and does, "*assume*, for the sake of the argument, that there are moral facts" (Sturgeon 1985, p. 237). One could think at this point that that is not a substantive assumption, amounting to something like "there are true simple modal statements." Not so: as he himself makes explicit, his assumption has a much richer content —and, as we are going to see, essentially so— that those moral facts involve moral properties that are, or supervene upon more basic,

---

[18] As I understand her, something like this is also the view of Judith J. Thomson, see Harman & Thomson (1996), Thomson (1998a) and (1998b).

[19] See footnote 21.

natural properties (Sturgeon 1985, p. 247), so that for anything that has them, "could not have differed in its moral quality without differing in those other [more basic features that makes it have it] as well. (Sturgeon 1985, p. 249).

Let me say a few words on supervenience. There are good (in part a posteriori) reasons for holding that everything supervenes upon the way the world naturally is. How to characterize exactly the content of this rough claim is, of course, by no means easy. But it will be clear that evaluative properties such as the Lewisian, flexible, approach conceives them, *do* clearly supervene on the natural in this sense. (At least, they do so on the assumption that psychological entities, to which evaluative ones flexibly reduce, do.) Furthermore, in the literature there is also a claim sometimes intended as a supervenience claim such that anyone is committed (at any moment) to evaluate similarly things that she judges not to differ naturally. That also holds, again obviously, for flexible response-dependent values. What is *not* true according to the flexible proposal is that evaluative properties supervene *locally* on natural entities, and more in general, on entities which are independent of the relevant valuers. Indeed for any target situation, if its proper description favors a flexible account, then it constitutes a counterexample of the relevant local supervenience claim. And conversely, the relevant realist alternatives could indeed be alternatively characterized by holding the relevant local supervenience claims.[20]

It is then clear that the content of Sturgeon's assumption is, in my terms, that moral properties are *real* properties. This is OK for evaluating Sturgeon's target: that even if moral properties were real properties, they would be explanatorily irrelevant. And his argument is straightforward:

> [C]onsider Harman's own example in which you see the children igniting a cat and react immediately with the thought that it is wrong. Is it true, as Harman claims, that the assumption that the children are really doing something wrong is "totally irrelevant" to any reasonable explanation of your making that judgment? Would you, for example, have reacted in just the same way, with the thought that the action is wrong, even if what they were doing hadn't been wrong, and could we explain your reaction equally well on that assumption? … [I]f what they are actually doing is wrong, and *if moral properties are, as many writers have held, supervenient on natural ones*, then in order to imagine them not doing something wrong we are going to have to suppose their action different form the actual one in some of its natural properties as well. So our question becomes: Even if the children have been doing something else, something just different enough *not to be wrong*, would you have taken them even so to be doing something wrong? (Sturgeon 1985, p. 247, my emphasis)

And the answer, I am ready to grant, could be 'no.' Suitably generalized, and in our terms: we have granted that there will be deep empirical explanations of our issuing the relevant evaluative responses when confronted with instances of evaluative properties, and our general capacity of so issuing them. If it is *assumed* that the relevant explanatory properties *are* the evaluative properties, then evaluative properties wouldn't be explanatorily irrelevant. That is something that a defender of the flexibility of val-

[20] For further discussion, see López de Sa 2005.

ues could, and I think should, accept.[21] But that is *not* an "argument to the best explanation" for evaluative realism, though it is sometimes seen in this way.[22]

Conclusion

Evaluative realism, I have claimed, should be characterized as denying the flexibility of values. But, I have also claimed, the intuitive description of the relevant counterfactual target situations, like the Evaluative Twin Earth, *does* support flexibility. As some people often claim they disagree with this, I have tried to make explicit some of the consequences of the relevant alternative realist descriptions, in the hope that some will acknowledge their counterintuitive character and revise thereby what they took to be their own realist supporting intuition. According to some evaluative realists, the flexibility supporting intuitions about the proper description of the relevant counterfactual target situations would have their own counterintuitive consequences. To the best of my knowledge, it is usually claimed that flexibility would have unacceptable relativistic consequences with respect to the evaluative domain, for instance being incapable of accounting for the fact that people *disagree* in normal conversations about evaluative issues. I do think that flexibility has relativistic consequences, but I would resist the claim that they are unacceptable. Rather, they are what intuitively seems predictable, in particular when we attend to a presuppositional element, to the effect that participants in conversations are relevantly like the speaker, which is congenial to the flexible proposal.

REFERENCES

Brink, D. (1986). "Externalist Moral Realism", *Southern Journal of Philosophy* 24, 23-41.
——— (1989). *Moral Realism and the Foundations of Ethics*. Cambridge, MA: Cambridge University Press.
Dreier, J. (1990). "Internalism and Speaker Relativism", *Ethics* 101, 6-26.
Fine, K. (1994). "Essence and Modality", *Philosophical Perspectives* 8, 1-16.
García-Carpintero, M. (2002). "A Non-modal Conception of Secondary Properties", plenary lecture at ECAP4, Lund (Sweden).
Harman, G. and J. J. Thomson (1996). *Moral Relativism and Moral Objectivity*. Oxford: Basil Blackwell.
Haukioja, J. (2000). *Rule-Following, Response-Dependence and Realism*. Turku: University of Turku.
Holland, S. (2001). "Dispositional Theories of Value Meet Moral Twin Earth", *American Philosophical Quarterly* 38, 177-96.
Honderich, T. (ed.) (1985). *Morality and Objectivity*. London: Routledge & Kegan Paul.
Hooker, B. and M.O. Little (eds.) (2000). *Moral Particularism*. Oxford: Oxford Clarendon Press.

[21] And I don't know why Sturgeon thinks Harman wouldn't: Harman has always been explicit acknowledging that "reductivists" would straightforwardly solve the question of the explanatoriness of evaluative properties.

[22] If I understand her right, Orlando (2001) reconstructs Sturgeon's argument this way, and then (rightly) accuses it of begging the question (see Orlando 2001, p. 339). As for her own "abductive argument" for evaluative realism, she also seems to grant that evaluative properties are explanatorily irrelevant *vis-à-vis the relevant responses*, but adds that moral facts could explain other moral facts. Provided that the explanation here is not empirical, this is fully compatible with the flexible account I favor.

Horgan, T. and M. Timmons (1991). "New Wave moral Realisms Meets Moral Twin Earth", *Journal of Philosophical Research* 16, 447-65.
——— and M. Timmons (1992a). "Troubles on moral Twin Earth: Moral Queerness Revived", *Synthèse* 92, 221-60.
——— and M. Timmons (1992b). "Troubles for New Wave Moral Semantics: The Open Question Argument Revived", *Philosophical Papers* 21, 153-75.
——— and M. Timmons (1996). "From Moral Realism to Moral Relativism in One Easy Step", *Crítica* 28, 3-39
——— and M. Timmons (2000a). "Copping Out on Moral Twin Earth", *Synthèse* 124, 139-52.
——— and M. Timmons (2000b). "Nondescriptivist cognitivism: Framework for a New Metaethic", *Philosophical Papers* 29, 121-53
Jackson, F. (1998). *From Metaphysics to Ethics*. Oxford: Clarendon Press.
——— and P. Pettit (1995). "Moral Functionalism and Moral Motivation", *Philosophical Quarterly* 45, 20-40.
———, P. Pettit and M. Smith (2000). "Ethical Particularism and Patterns", in Hooker and Little (2000).
Johnston, M. (1989). "Dispositional Theories of Value," *Proceedings of the Aristotelian Society* 63 (suppl.), 139-74.
——— (1998). "Are Manifest Properties Response-Dependent Properties?", *The Monist* 81, 3-43.
Kripke, S. (1980). *Naming and Necessity*, Oxford, Blackwell.
Lang, G. (2001). "The Rule-Following Considerations and Metaethics. Some False Moves", *European Journal of Philosophy* 9, 190-209.
Lewis, D. (1989). "Dispositional Theories of Value", *Proceedings of the Aristotelian Society*, 63 (suppl.), 113-38, reprinted in Lewis (2000), from where I quote.
——— (2000). *Papers in Ethics and Social Philosophy*. Cambridge, MA: Cambridge University Press.
López de Sa, D. (2000). "Conceptos dependientes de respuesta. El problema de la explicación perdida", in Mary Sol de Mora *et al.* (eds.), *Actas del III Congreso de la SLFMCE*. Donostia: UPV/EHU.
——— (2003). *Response-Dependencies. Colors and Values*. Barcelona: Universitat de Barcelona.
——— (2005). "Disposiciones, motivación, y superveniencia: Réplica a Arrieta", in E. Txapartegi (ed.), *Los objetos de la ciencia*. Córdoba: Brujas.
——— (2006). "Values vs. Secondary Qualities", *Teorema* 25, 197-210.
Mackie, J.L. (1977). *Ethics. Inventing Right and Wrong*. London: Penguin Books, chapter 1 reprinted in Sayre-McCord (1989a), from where I quote.
McDowell, J. (1985). "Values and Secondary Qualities", in Honderich (1985).
——— (2006). "Response to Dan López de Sa", *Teorema* 25, 211-214.
Miller, A. (2001). "The missing-explanation argument revisited", *Analysis* 61, 76-86.
Orlando, E. (2001). "Abduction, Realism and Ethics", *Theoria* 16/41, 331-52.
Sayre-McCord, G. (1988a). *Essays on Moral Realism*. Ithaca and London: Cornell University Press.
——— (1988b). "The Many Moral Realisms," in Sayre-McCord (1988a).
Smith, M. (1994). *The Moral Problem*, Oxford: Blackwell.
Schwartz, S.P. (2002). "Kinds, General Terms, and Rigidity," *Philosophical Studies* 109, 265-77.
Soames, S. (2002). *Beyond Rigidity*. Oxford. Oxford University Press.
Sturgeon, N.L. (1985). "Moral Explanations", in D. Copp and D. Zimmerman (eds.), *Morality, Reason and Truth*, Totowa, N.J.: Rowman & Allanheld, reprinted in Sayre-McCord (1988a), from where I quote.
Thomson, J.J. (1998a). "*Précis* of Part Two", *Philosophy and Phenomenological Research* 58, 171-3.
——— (1998b). "Reply to Critics", *Philosophy and Phenomenological Research* 58, 215-22.
Wedgwood, R. (1998). "The Essence of Response-Dependence", in R. Casati and Ch. Tappolet (eds.) *European Review of Philosophy 3. Response-Dependence*, Stanford: CSLI Publications.
Wright, C. (1992). *Truth and Objectivity*. Cambridge, MA. Harvard University Press.

Dan LÓPEZ DE SA is a Postdoctoral Research Fellow in Arché (St Andrews) and member of LOGOS (Barcelona). From October 2006 is at NYU as a GenCat-Fulbright Postdoc. He was Professor Associat (2002-2004) and before a PhD Research Fellow (1998-2001) at the Universitat de Barcelona, where he got his PhD with the thesis *Response-Dependencies: Colors and Values* (2003), and a Visiting Student for a short while in 2001 at the Philosophy Program of the RSSS (ANU). His research is mainly in metaphysics, philosophy of language, and metaethics.

ADDRESS: Arché–The AHRC Research Centre for the Philosophy of Logic, Language, Mathematics and Mind, University of St Andrews, 17 College Street, St Andrews KY16 9AL, Scotland. E-mail: dlds@st-andrews.ac.uk.

The Philosophy behind Quantum Gravity

Henrik ZINKERNAGEL

Received: 2006.05.29
Final Version: 2006.10.09

BIBLID [0495-4548 (2006) 21: 57; pp. 295-312]

ABSTRACT: This paper investigates some of the philosophical and conceptual issues raised by the search for a quantum theory of gravity. It is critically discussed whether such a theory is necessary in the first place, and how much would be accomplished if it is eventually constructed. I argue that the motivations behind, and expectations to, a theory of quantum gravity are entangled with central themes in the philosophy of science, in particular unification, reductionism, and the interpretation of quantum mechanics. I further argue that there are —contrary to claims made on behalf of string theory— no good reasons to think that a quantum theory of gravity, if constructed, will provide a theory of everything, that is, a fundamental theory from which all physics in principle can be derived.

Keywords: reductionism, quantum gravity, quantum mechanics, unity of physics.

1. Introduction

One of the outstanding tasks in fundamental physics, according to many theoretical physicists, is the construction of a quantum theory of gravity. The so far unsuccessful attempt to construct such a theory is an attempt to unify Einstein's general theory of relativity with quantum theory (or quantum field theory). While quantum gravity aims to describe everything *in* the universe in terms of quantum theory, the purpose of the closely related project of quantum cosmology is to describe even the universe as a whole as a quantum system. At present, a quantum theory of gravity is mainly sought along two avenues (both of which are associated with a number of technical and conceptual problems). The first of these is canonical quantum gravity in which the classical Einstein equations are somehow quantized.[1] The second, and most popular, program for quantum gravity is that of string theory. Contrary to canonical quantum gravity, string theory aims to unite the description of gravity with those of the other forces in nature (electromagnetic, weak, and strong forces), and is in this sense the most ambitious attempt of a quantized theory of gravity. Thus, string theory not only postulates (like canonical quantum gravity) unification in the sense that all forces are quantum in nature but also that all the quantum forces can be derived from one single theory. String theory is therefore often referred to as a candidate for a theory of everything.

[1] Such a quantization might be carried out e.g. by making the space-time metric a quantum operator. In a sense this amounts to a 'discretization' of space and time insofar as one can at all speak of space and time in quantum gravity (see below). For a good popular introduction to the different approaches to quantum gravity, see Smolin (2001).

The quantum gravity project raises a number of philosophical issues, some of which I shall deal with below (in this paper the hard technical problems associated with quantum gravity will be ignored). In particular, I will critically examine the motivations behind quantum gravity and the question of whether such a theory is, if not strictly necessary, then at least desirable. Furthermore I will address the question of whether a quantum theory of gravity, if constructed, can fit the bill of being a kind of ultimate theory which could in principle account for all physical phenomena.

The outline of the paper is as follows. I first (section 2) discuss how the motivations behind a quantum theory of gravity are related to the ideas of unity and reductionism in physics. In this connection, I review Bohr's idea of unity without reductionism, and discuss how the enterprise of quantum gravity is related to the interpretation of quantum mechanics. I then (section 3) briefly review an argument for the necessity of a quantized theory of gravity, and argue that such a theory is necessary neither for consistency reasons nor (at least so far) on experimental grounds. In a broad sense quantum gravity may be conceived of as any theory which couples general relativity (and thus a classical description of gravity) with quantum theory. I argue that the expectation —which serves as a motivation for quantum gravity in the broad sense— that general relativity and quantum theory must be connected in the high energy regime might be questioned (in particular due to the so-called cosmological constant problem). In section 4, I put forward an argument which suggests that the eventual construction of a quantum theory of gravity is not likely to be a fundamental theory in the sense often advocated (i.e. a theory from which all other theories of, and phenomena in, physics could be derived). I point out that whatever formal relations can be established between quantum gravity and the supposedly less fundamental (classical) theories, the latter are in any case needed to specify the field of application of the former. This raises doubts concerning the sense in which quantum gravity could be *the* fundamental theory.

2. Reductionism and the Unity of Physics

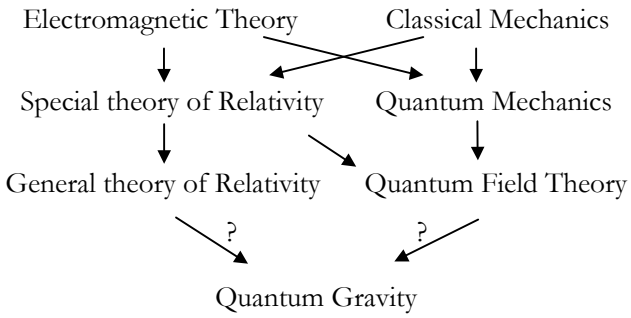
The quest for unification is a major drive behind the search for quantum gravity. For instance, Kiefer (2004, p. 2) writes in the introduction to his recent book on quantum gravity concerning the main motivations for this theory:

> The first motivation is unification. The history of science shows that a reductionist viewpoint has been very fruitful in physics (Weinberg 1993). The standard model of particle physics is a quantum field theory that has united in a certain sense all non-gravitational interactions. [...] The universal coupling of gravity to all forms of energy would make it plausible that gravity has to be implemented in a quantum framework too.

As discussed below, it is not always the case that unification coincides with reductionism. In any case, it is true that the idea of unification between the different natural phenomena, and the theories that describe them, has been a guiding principle in physics at least since the days of Galileo. Indeed, the "success" history of physics can, at least partly, be portrayed as the history of unification. Think for example of Newton's unification of heavenly and terrestrial phenomena by the universality of gravitation; or

Ørsted, Faraday, and Maxwell's unification of electric and magnetic forces embedded in Maxwell's equations. A more modern example is provided by Glashow, Salam and Weinberg's electroweak theory of elementary particles which couples electromagnetic and weak forces (the latter being responsible for certain types of radioactivity). In order to see where quantum gravity fits into the unification picture, it is helpful to briefly review the relations between some of the central theories of physics.

Quantum field theory and general relativity stand as two of the greatest achievements in 20th century physics.[2] Both theories are already unified in the sense that they combine various former theories in a common framework. Thus, the special theory of relativity is a combination of electromagnetism and the non-gravitational part of classical mechanics; and general relativity is a generalization of the special theory in which gravitation is also included. In a similar manner, quantum field theory is a combination of special relativity and quantum mechanics. The situation can be schematically represented as follows (note that only the physical theories relevant for this paper are included):



The arrows in this scheme represent the direction towards deeper or more general layers in our description of nature (see below). The question marks represent that a theory of quantum gravity and —more generally— the connection between the quantum (right-hand) side and general relativity, is still a speculation only.

In various ways the scheme is an expression of reductionism. On the one hand, the arrows indicate the direction towards something smaller (right hand side). On the other hand, the arrows indicate the direction towards something more general (both left and right hand side). A theory of quantum gravity (in particular string theory

[2] Quantum field theory is here and in the following understood as the common framework for the theory of light and electrons as fields (quantum electrodynamics), the theory of weak nuclear forces, and the theory of quarks and gluons. The combination of these theories —known as the standard model of particle physics— describes the inner structure of atoms via quantum fields.

which aims to unify all forces known in nature) combines these trends by describing something both smaller and more general than what is found on the higher levels.[3]

In accordance with these reductionist trends the higher levels are often seen as merely useful special cases of the deeper levels. Quantitatively, this thought is backed by the fact that at least some of the mathematical expressions of the deeper levels are identical with those of the higher levels in certain limiting cases.[4] When such mathematical identity can be established, reductionism contains the possibility of reconstructing the higher levels from the deeper ones. Indeed, an important motivation behind the most ambitious quantum gravity program, string theory, is precisely to reverse the arrows of the above scheme and derive all known physics from a few basic principles of this theory. This idea is in accordance with Einstein's declaration from 1918:

> The supreme task of the physicist is to arrive at those universal laws from which the cosmos can be built up by pure deduction.[5]

The Einsteinian ambition is thus not just a reduction to more fundamental theories — that is, either to dissect objects into smaller and smaller parts or show that theories on a higher level are special cases of those of a deeper level (or both in the case of string theory). Also, and more explicitly in the quote, the idea is *reconstructing* the universe from scratch. That is, if we have the universal laws described by a fundamental theory, and we have identified the fundamental constituents of matter, then we can derive — at least in principle— all phenomena in the universe (modulo the indeterminism stemming from quantum theory), as well as the theories describing these phenomena.[6] A contemporary expression of such an ambitious reductionism/reconstructivism can be found in Tegmark and Wheeler (2001). These authors include a much more general

---

[3] A referee points out that one ought to distinguish between theoretical and ontological reductionism since the former is much more difficult and limited than the latter. However, this distinction is not without problems in the quantum context. For instance, while modern physics asserts that matter is made up of atoms, any adequate description of these objects (and the precise sense in which they constitute matter) requires quantum theory. Moreover, recent studies (of decoherence) have revealed that atoms and molecules can behave as quantum objects in one context, and as classical objects in another (see e.g. Arndt *et al* 1999).

[4] As noted e.g. in Weinstein (2005, p. 5), none of the programs for quantum gravity has as yet succeeded in showing that the less fundamental theories (general relativity in the case of canonical quantum gravity or general relativity + the standard model of particle physics in the case of string theory) can be obtained in some limiting case.

[5] The quote is from a conference entitled 'Principles of research' delivered in Berlin in connection with the 60th birthday of Max Planck.

[6] It should be noted that Einstein was unsatisfied with quantum theory as a final theory and would therefore, presumably, not have agreed with quantum gravity being a candidate for such a unified theory. Indeed, Einstein did not engage in quantum gravity debates and instead worked, until his death in 1955, on a classical unified theory of physics (attempting to combine electromagnetism and gravity), see Stachel (1999).

scheme than the one above, in which subjects such as chemistry, biology, psychology and sociology are all seen as derived from fundamental physics.

Unity without reductionism?

Various doubts can be raised against reductionism. Often the debate is focused on the notion of emergence —the question of whether new and irreducible phenomena exist at the higher levels of description (irreducible in the sense that the emergent phenomena cannot be explained by the deeper level).[7] For the subsequent discussion on how much can be expected from a theory of quantum gravity, however, it will be more useful to briefly review a different anti-reductionistic argument which can be associated with Bohr's insistence on the necessity of classical physics in the description of quantum phenomena. If correct, this argument demonstrates not only that phenomena of (some of) the higher levels cannot be reduced to (or reconstructed from) the deeper ones, but also that the phenomena of the deeper levels cannot be defined, and are therefore dependent on, (some of) the higher levels.

Bohr contended that we cannot account for (or understand) the quantum phenomena —for instance the motion of an atom, or the interference pattern in the famous double-slit experiment— unless reference is made to a specific experimental context in which the measurement apparatus must be described by the concepts of classical physics, see e.g. Bohr (1958, p. 4). The idea of being described by classical physics concepts implies that the measurement apparatus —in contrast to quantum systems— has well-defined values of both position and momentum, and thus it is not subject to any quantum uncertainties or superpositions. According to Bohr a main reason for the necessity of this distinction between the quantum objects and the measuring instrument is that the interaction between the object and the apparatus is a defining feature of the quantum phenomena (Bohr 1958, p. 4).

Of course, Bohr's view is just one of a number of proposed alternative interpretations of quantum theory. Most of these alternatives follow the line of von Neumann and attempt (in a reductionist spirit) to treat the measurement apparatus itself as a quantum system. As is well known, however, such approaches run into the notorious measurement problem. Stated briefly, the problem is that *if* everything, including measurement apparatuses, is quantum (and thus correctly described by Schrödinger's equation), then we ought to see superpositions in the measurement outcomes, e.g. apparatus pointers being in various positions at the same time —and that clearly contradicts what we in fact do see. Responding to this problem involves invoking assumptions —such as many worlds, hidden variables, or modified dynamics— which go beyond the quantum formalism itself to somehow 'explain away' why no quantum

[7] Note that some proponents of emergent phenomena still agree with reductionism but reject reconstructionism. For instance, Anderson (1972) holds that "The ability to reduce everything to simple fundamental laws does not imply the ability to start from those laws and reconstruct the universe". Precisely how the relation between different levels should then be understood, however, is unclear; see e.g. Cat (1998) for a critical discussion of Anderson's view and for a detailed discussion of the relationship between unity, emergence, and (different notions of) reductionism in modern physics.

strangeness is seen in the measurement results (for an overview of proposed responses to the measurement problem, and their problems, see e.g. Albert 1992).[8]

Bohr actually agreed that the measurement apparatus can also be described by quantum theory. However, he writes (1939, p. 104):

> ... in each case some ultimate measuring instruments, like the scales and clocks which determine the frame of space-time coordination —on which, in the last resort, even the definitions of momentum and energy quantities rest— must always be described entirely on classical lines, and consequently kept outside the system subject to quantum mechanical treatment.

The point is that we can treat a measuring apparatus (or part of this) as a quantum system, but only when some other system is then treated classically. This requirement guarantees, in consistency with what we observe, that measurements do indeed have definite outcomes.[9] Thus Bohr can effectively be taken to argue that *any* system, at least in principle, can be treated quantum mechanically, but that not all systems can be treated that way at the same time. This means that to those who hold that all objects *are* quantum in some ontological sense, Bohr might well have responded that although the existence of both measurement apparatuses and, say, atoms are beyond doubt — we cannot say exactly what these objects are like. All which can be inferred is that in some circumstances objects can be described *as if* they were quantum and in other circumstances *as if* they were classical.[10]

These brief remarks cannot, of course, constitute a comprehensive analysis of Bohr's view —and much less a satisfactory defence of it. But they indicate one way to have unity without reductionism. For if quantum physics (and quantum phenomena) cannot be understood without classical physics (and classical phenomena, e.g. objects with well-defined values of both position and momentum), it is altogether unclear

---

[8] Note that even if some of these 'quantum reductionist' approaches (in which all systems are treated as being quantum) are taken as adequate responses to the measurement problem, the general reductionist program in physics is not automatically vindicated. For instance, hidden variable theories like Bohm's and the spontaneous collapse models of Ghiradi, Rimini and Weber have been charged of being incompatible with special relativity (see e.g. Barrett 2000 and 2003).

[9] See Howard (1994) for an interesting suggestion of how Bohr's ideas on this point might be reconstructed and understood in terms of entanglement between the measurement apparatus and the quantum object under investigation. Howard hints (1994, p. 204) that Bohr's insistence on classical descriptions can be understood without assuming any "...fundamental ontological or epistemological distinction [between the classical and the quantum]". I do not agree with this claim but I cannot argue the point here.

[10] Of course, many scholars have been dissatisfied with Bohr's 'dissolution' of the measurement problem (by always having part of the system being described classically). Among the problems with Bohr's account are that there is no clear prescription of how the borderline between the classical and the quantum should be made (rather it is context dependent from case to case), see e.g. Bell (2004, p.171); and that is does not give an account of (but rather black-boxes) exactly what happens in a measurement situation, that is, how precisely the classical apparatus interacts with the quantum object, see e.g. Howard (1994, p. 211).

what it would mean to reduce the latter to the former.[11] And unity is not denied if this is taken to mean that entities and phenomena of two theories are interconnected but not reducible to each other. Indeed, Bohr emphasized both that, on the one hand, classical physics is necessary to define quantum phenomena, and, on the other hand, quantal laws are needed to explain the stability of classical objects (see e.g. Bohr 1958, p. 2).

This brief discussion illustrates that a motivation for quantum gravity based on an appeal to reductionism in physics can be resisted —even while the quest for unity is maintained. More importantly, as we shall see in section 4, there is a sense in which Bohr's insistence on the necessity of classical physics for understanding quantum physics might be vindicated in connection with specifying the field of application for a quantum theory of gravity.

3. Is a theory of quantum gravity necessary?

Considerations of whether or not reductionism has been a successful doctrine in physics would, of course, be largely irrelevant for motivating the project of quantum gravity if such a theory were in any case needed on experimental or logical grounds. With respect to the latter, Bryce DeWitt argued in the early 1960s that just as the electromagnetic field must be quantized to be consistent with quantum mechanics, the gravitational field should be quantized for the same consistency reason. DeWitt's argument (1962), which has since been repeated by other physicists, rests on two premises; 1) the existence of logical arguments for the quantization of the electromagnetic field; and 2) that the electromagnetic case is sufficiently analogous to the gravitational case. According to DeWitt and others, the first premise is supposed to follow from a famous analysis from 1933 in which Bohr and Rosenfeld discussed the measurability of the quantized electromagnetic field. In particular, the Bohr-Rosenfeld analysis is claimed to show that the uncertainty relations for a charged particle interacting with an electromagnetic field necessitates that the electromagnetic field is also quantized.

Rosenfeld himself, however, saw matters differently. Although he had been the first to try to construct a theory of quantum gravity (in 1930), he later expressed hesitations towards the project —in particular because there was no experimental evidence for any quantum effects of gravity (this is still true, see below). With respect to DeWitt's argument, Rosenfeld (1963) pointed out that the Bohr-Rosenfeld analysis

---

[11] Note that this is not in conflict with the formalistic fact that quantum expressions may correspond to classical expressions in certain limiting cases. One way in which classical mechanics may be said to be a limiting case of quantum mechanics is via the so-called Ehrenfest's theorem which is the quantum mechanical equivalent of Newton's second law. However, since this theorem involves mean (or expectation) values of quantum operators, and since such expectation values are bound up with the quantum mechanical measurement process, there is —in spite of formal identity of Ehrenfest's theorem and Newton's second law in certain limits— no question of deriving classical behaviour from the quantum formalism, see e.g. Joos et al (2003, p. 87). For examples of how, on a Bohrian understanding, quantum mechanics coincide with (but do not reduce) classical physics in certain limits, e.g. via the correspondence principle, see Falkenburg (1998).

showed the consistency of the electromagnetic field quantization (i.e. that it is possible to treat the electromagnetic field with quantum principles), *not* its necessity. Furthermore, Rosenfeld argued that the analogy between the gravitational and the electromagnetic field (DeWitt's second premise) is problematic due to the appearance of a definite scale for space and time intervals in the quantum theory of gravity. Such length and time scales, referred to as Planck scales, result from the combination of Newton's gravitational constant G, Planck's constant $\hbar$, and the speed of light c – the Planck length is $\sqrt{G\hbar}/c^{1,5} \approx 10^{-33}$ cm, and the Planck time is $\sqrt{G\hbar/c^5} \approx 10^{-43}$seconds. Rosenfeld notes that such small length and time scales may not be well-defined since considerations of an analogous case from quantum electrodynamics in which scales are involved (when the charge and current distributions are quantized) suggest an absolute limit to space-time localization given by the proton radius,10^{-13} cm, which is 20 orders of magnitude larger than the Planck length scale (see also Rosenfeld 1966, p. 605). Finally, Rosenfeld stressed that the eventual construction of a quantum theory of gravity could not essentially change the fundamental role of classical theory for the understanding of quantum theory. Commenting on the early Bohr-Rosenfeld analysis, he wrote (Rosenfeld 1963, p. 443):

> The ultimate necessity of quantizing the electromagnetic field (or any other field) can only be founded on experience, and all that considerations of measurability of field components can do is to illustrate the consistency of the way in which the mathematical formalism of a theory embodying such quantization is linked with the classical concepts on which its use in analysing the phenomena rests.

Thus, Rosenfeld most likely agreed with Bohr's vision of the unity of physics implying that quantum gravity could not possibly be a final theory from which classical physics (and classical phenomena) can be derived.

Recent studies have shown that the situation concerning the necessity of quantizing the gravitational field has remained essentially unchanged since Rosenfeld's remarks. Thus, Callender and Huggett (2001) and Wüthrich (2004), reviewing and evaluating arguments concerning the necessity of quantization, both argue that there are no convincing reasons to affirm that gravity must be quantized. However, all of these authors agree that a theory of quantum gravity —understood in the broad sense as any theory which couples general relativity and quantum theory— is nevertheless desirable, and that there are situations in which such a theory is needed. We turn to their arguments below after a quick look at the empirical situation.

Quantum gravity vs. observations and experiments

No observations or experiments have so far observed any quantum effects of gravity. This is, however, not surprising since quantum effects of gravity are expected to show up primarily at the above mentioned Planck scales. The most likely *observational* signa-

[13] Other effects are being contemplated within the field known as quantum gravity phenomenology. For instance, it has been suggested that quantum gravity effects could be responsible for certain puzzling observations of cosmic rays. The situation, however, is still far from settled; see e.g. Amelino-Camelia (2003).

tures of quantum gravity are to be found in the very early (small time scale) universe, or the extreme conditions (high energy scale) associated with black holes (see below) —and none of these regions are easy to access observationally.[13] Another option for probing quantum gravity effects is to try to access the Planck scales via *experimental* studies of phenomena at very small length scales. Such experimental studies, however, seem remote. For instance, Baez (2001) notes:

> To study a situation where both general relativity and quantum field theory are important, we could try to compress a cell to a size 10^{-20} times that of a proton. We know no reason why this is impossible in principle, but we have no idea how to actually accomplish such a feat.

Nevertheless, as another motivation for quantum gravity it is sometimes mentioned (e.g. Callender and Huggett 2001, p. 5) that although no effects of *quantum* gravity has been seen, experiments have established that classical gravity is indeed related to (non-relativistic) quantum theory. One such experiment used a so-called neutron interferometer to demonstrate that the gravitational field affects the behaviour of quantum systems such as a beam of neutrons (see e.g. Greenberger and Overhauser 1980). The fact remains, however, that there is a big step from the relation between classical gravity and quantum mechanics to quantum effects of gravity itself. For instance, there are no experimental signatures of a relation between quantum *field* theory (like quantum electrodynamics) and gravity.[14] On the one hand, the absence of such experimental signatures is not surprising since quantum field theory deals almost exclusively with microscopic systems (in contrast to e.g. the neutron interferometer which allows for a test of quantum mechanics at the macroscopic level) and since the gravitational force is very small in the microphysical domain. On the other hand, this situation emphasizes how difficult it is to establish whether there are any quantum effects of gravity (as a quantized field) or even any observational effects of a coupling between general relativity (gravity as a classical field) and quantum field theory.

Alternative relationships between general relativity and quantum theory?

Given that quantization of the gravitational field is not required on consistency or empirical grounds, it is natural to ask how the relationship between general relativity and quantum theory could be conceived in case the gravitational field is not quantized. For instance, Butterfield and Isham have remarked (Butterfield and Isham 2001, p. 57):

> If it is indeed wrong to quantise the gravitational field […] it becomes an urgent question how matter —which presumably *is* subject to the laws of quantum theory— should be incorporated in the overall scheme.

As we shall see, the urgency of this question depends on how relevant (experimentally and observationally) such an 'overall scheme' is —and, of course, whether there is one!

---

[14] The so-called Unruh-Davies effect and the Hawking radiation from black holes are theoretical phenomena which are predicted from a relation between quantum field theory and gravity, but so far they have not received empirical support.

In his 1963 paper Rosenfeld argued that since there is no experimental need for quantizing gravity, it is better to stick with the so-called semi-classical gravity, which combines a classical description of the gravitational field with a quantum treatment of all other force fields and matter. Technically, the left hand side of the classical Einstein equation, describing the curvature of space-time, is equated with the quantum expectation value of the so-called energy-momentum tensor which is a measure of the energy associated with (quantum) matter and radiation:

$$R_{\mu\nu} - \frac{1}{2} R g_{\mu\nu} = \frac{8\pi G}{c^4} \langle T_{\mu\nu} \rangle \qquad (1)$$

where $R_{\mu\nu}$ and R refer to the curvature of space-time, $g_{\mu\nu}$ is the metric, and $T_{\mu\nu}$ the energy-momentum tensor. However, this equation is problematic —not least because the quantum expectation value on the right hand side is calculated in a *fixed* space-time background, whereas the left hand side describes a *dynamical* space-time (this leads to difficult non-linearity and back-reaction problems; see e.g. Callender and Huggett (2001) and Rugh and Zinkernagel (2002)). Such problems are sometimes used to argue that the semi-classical approach, at least in its simplest form given by equation (1), is not likely to be the fundamental theory describing the interaction between general relativity and quantum field theory; see e.g. Kiefer (2004, p. 14 ff.).

But if there is no strict need to quantize gravity, and if semi-classical gravity is a problematic answer to the question of the relationship between general relativity and quantum theory, could one then not choose to forget about the whole business of quantum gravity? Callender and Huggett (2001, p. 4) comments on this possibility:

> Another philosophical position, which we might dub the 'disunified physics' view might in this context claim that general relativity describes certain aspects of the world, quantum mechanics other distinct aspects, and that would be that. According to this view, physics (and indeed, science) need not offer a single universal theory encompassing all physical phenomena. We shall not debate the correctness of this view here but we would like to point out that if physics aspires to provide a complete account of the world, as it traditionally has, then there must be a quantum theory of gravity [in the general sense of a connection between general relativity and quantum theory]. The simple reason is that general relativity and quantum mechanics cannot both be correct even in their domains of applicability.

As a first argument for this conclusion Callender and Huggett mention that the two theories "...cannot both be universal in scope, for the latter strictly predicts that all matter is quantum, and the former only describes the gravitational effects of classical matter...". Now, for all the impressive empirical successes of quantum theory, it does not predict that all matter *is* quantum. This conclusion only follows by adopting an ontological interpretation of quantum theory according to which, indeed, everything ultimately is quantum. I have already mentioned that, on Bohr's view, this move can be resisted insofar as objects are not either quantum or classical (even if, say, macroscopic systems are more prone to a classical description than are microscopic systems).

Whatever the plausibility of Bohr's position, however, Callender and Huggett offer a further argument for their conclusion that general relativity and quantum theory must somehow interact (this is a common argument, contained also in the quote by Kiefer in section 2): The Einstein field equation couples the gravitational field, and thus the space-time structure, to all matter and energy. And,

> [q]uantum fields carry energy and mass; therefore, if general relativity is true, quantum fields distort the curvature of spacetime and the curvature of spacetime affects the motion of the quantum fields. If these theories are to yield a complete account of physical phenomena, there will be no way to avoid those situations —involving very high energies— in which there are non-negligible interactions between the quantum and gravitational fields...

As I will discuss further below, the situations where such interactions become relevant are mostly associated with the very early universe or with speculative features of black holes. Now, it could be objected that these situations are not relevant (at least yet), since there are still no experimental or observational evidence which make them so. This could for instance be argued within the framework of Cartwright's 'Dappled World' —which might well be the implicit target for Callender and Huggett's reluctance towards 'disunified physics'. Cartwright argues (1999, p. 24 ff.) that we have no reasons to believe our theories outside the domain where the successes of these theories have been established or, at least, not outside the domain where an adequate model of the phenomena in question can be formulated —and in this sense physical theories should not be expected to give a complete account of physical phenomena. For instance, Cartwright holds that a 1000 dollar bill swept away by the wind cannot be taken to follow Newton's second law since no good-fitting molecular (or otherwise) model of the wind is available. In the absence of such a model, on Cartwright's view, the belief that the wind operates on the bank note via forces is "...another expression of fundamentalist faith" (1999, p. 28). However, as Hoefer (2003, p. 1408) reasonably complains, Cartwright's rejection to make inductive inferences from our successful theories to not easily modelled cases comes dangerously close to reject making any inductions at all!

Nevertheless, without committing myself to Cartwright's general thesis, I do think a bit of scepticism about the scope of our inductions is in order in the present case. For there are reasons to suspect that our intuitions about the relationship between general relativity and quantum (field) theory are insufficient to allow extrapolations (inductions) into the very high energy regime. I am here thinking in particular of the so-called cosmological constant problem which *might* threaten the very idea of a formal connection (like equation 1, or any other type of coupling) between quantum field theory and general relativity. In essence the problem is that whereas quantum field theory predicts an astronomically high value for $\langle T_{\mu\nu} \rangle$ —implying an extreme curvature of space-time— it is observationally known that space-time is flat or almost flat.[15] The

[15] More precisely, when the expectation value of $T_{\mu\nu}$ is evaluated in the vacuum state, it takes the same form as the cosmological constant in general relativity (this constant has, for simplicity, not been included in equation 1).

cosmological constant problem, which is so far unsolved, is conceived to be a 'veritable crisis' for fundamental physics, see Weinberg (1989). As discussed in Rugh and Zinkernagel (2002), the problem rests fundamentally on two assumptions —both of which can be questioned: (i) The quantum field theoretic vacuum energy is physically real (as in the standard interpretation of quantum field theory); and (ii) the validity of some semi-classical approach in which quantum (vacuum) energy acts as a source in Einstein's field equation. This last assumption implies a formal coupling between general relativity and quantum field theory, and the cosmological constant problem *could* therefore be argued to constitute a threat to the idea of a connection between these two theories.[16]

Thus, the apparently straightforward assumption made by Callender and Huggett that general relativity couples to all forms of (quantum) energy might be questioned. In any case, the cosmological constant problem suggests that our understanding of the connection between general relativity and quantum field theory is (still) too premature to trust extrapolations into the high energy regime. I emphasize that this does not constitute an argument to the effect that general relativity and quantum field theory are *not* connected. It is rather that, given the present unclear situation surrounding the cosmological constant problem, all cards should be left on the table.

So much by way of sketching how some of the motivations behind the search for a quantum theory of gravity —either as a theory of quantized gravity or in the broad sense of a theory which couples general relativity and quantum theory— could perhaps be resisted. In the following section, I will disregard possible scepticism as concerns the motivations for quantum gravity, and instead ask how much would be accomplished if such a theory is eventually constructed.

4. Could quantum gravity be the *fundamental theory?*

Although a quantum theory of gravity in which gravity is quantized has still not been found, the eventual construction of such a theory is often associated with at least two conjectures:

1. Quantum gravity will imply that our usual classical notions of space and time are only approximate valid concepts (valid at our length and time scales), which somehow emerge from the 'real' quantum nature of space and time; see e.g. Butterfield and Isham (1999).
2. More generally, quantum gravity will provide the ultimate explanation of classical physics from a deeper quantum physics level (for instance from string theory); see e.g. Weinberg (1993) and Tegmark and Wheeler (2001).

---

[16] It should be mentioned that the cosmological constant problem has also been read as an argument for the necessity of quantum gravity —and solutions to the problem have been proposed within this framework (and the related idea of quantum cosmology), see also Zinkernagel (2002). Apart from the fact that no general framework for quantum gravity exists as yet, however, these solutions seem to be problematic, see Weinberg (1989, p. 20ff.) for discussion and references.

I have already indicated that the second of these (reductionist) conjectures is in conflict with Bohr's thesis that classical physics is necessary to account for the quantum phenomena, and, as I shall briefly hint below, the first conjecture is also problematic from a Bohrian perspective. Not surprisingly, therefore, many physicists working on quantum gravity (and quantum cosmology) often appeal to some version of the Everettian many-worlds interpretation of quantum mechanics; see e.g. Butterfield and Isham (1999, p. 144). Disregarding the other alternative interpretations of quantum mechanics, it may therefore seem as if one is faced with a choice between Bohr's 'one world-two theories', or Everett's 'one theory-many worlds'. While I cannot, of course, attempt to seriously adjudicate between these interpretations here, I will suggest that Bohr's idea of the necessity of classical physics cannot be as easily dismissed in discussions of quantum gravity as is sometimes suggested.[17] This suggestion can be developed by asking how the theory of quantum gravity is supposed to work without classical physics.

As we have seen above, it is usually assumed that the effects of quantum gravity should become relevant at very small (Planck) time- and length-scales, respectively $\sim 10^{-43}$ sec and $\sim 10^{-33}$ cm. There are two types of situations, both involving very high energies, where these effects should manifest themselves. The first is the very early universe where the extreme conditions near the Big Bang are expected to make Planck scale physics necessary (the high energies near the Big Bang results because the time parameter in this cosmological model is inversely proportional to the square of the temperature —and hence to energy). The second situation is connected to black holes. These exotic objects, which are supposedly common in the universe, are expected to gradually lose energy (Hawking radiation). Due to curious effects of so-called black hole thermodynamics, the final stages of a black hole is supposed to be characterized by energy loss in the form of radiation in which each photon has an energy close to the Planck energy. Consequently, it is expected that this final stage can only be understood by a full quantum theory of gravity (in which gravity is quantized); see e.g. Smolin (2001, p. 92). Moreover, quantum gravity is also expected to describe the black hole singularity at —or very near— the center of a black hole.

It is sometimes said that since general relativity predicts space-time singularities (the Big Bang and the center of black holes), this theory predicts its own demise as it is unable to describe the vicinity of these singularities (due to quantum gravity effects). As discussed below, however, theoretical (and observational) access to either the very early universe or black holes relies firmly on classical theory —namely classical general relativity. This seems to imply that one must *presuppose* classical theories in order to *define* the field of application for quantum gravity. If this is correct then it at least limits the sense in which general relativity can be reduced to quantum gravity. On the other hand, the formalism of general relativity is expected to be derivable from an eventual theory of quantum gravity in some limiting case. This situation is somewhat analogous

[17] An example of a dismissive evaluation of Bohr's position in connection with quantum gravity can be found in Butterfield and Isham (1999, p. 143).

to Bohr's view on the role of classical mechanics in quantum mechanics. For instance, Landau and Lifshitz —quoting Bohr approvingly— wrote in the introduction to their book on quantum mechanics (1981, p. 3):

> Thus quantum mechanics occupies a very unusual place among physical theories: it contains classical mechanics as a limiting case, yet at the same time it requires this limiting case for its own formulation.

A way to understand this claim is, as we have seen, that formal identities between quantum and classical expressions notwithstanding, the measurement problem implies that quantum theory by itself cannot account for any classical phenomena —such as definite measurement outcomes with well-defined space-time and energy-momentum properties. I should note that the above mentioned necessity of general relativity for quantum gravity is only somewhat analogous to the necessity of classical mechanics for quantum mechanics —for the role of the classical theory in the former case is not to account for observed phenomena but rather to specify the field of application of the quantum theory. Nevertheless, in the case of quantum gravity, it is much less obvious that one can circumvent the need for a classical theory by opting for a different interpretation of quantum mechanics.

The problems of time

In order to spell out more clearly the way in which classical physics is presupposed in defining the field of application of quantum gravity, I shall briefly consider the role of the concept of time in the very early universe. Although quantum gravity is supposed to fundamentally change our usual notion of time, it is notoriously difficult to see how the notion of time employed in the less fundamental theories could somehow emerge from quantum gravity. For instance, the central equation in canonical quantum gravity, the Wheeler-DeWitt equation, does not depend on time at all, and this obviously makes it hard to see how the equation can be relevant for theoretical descriptions of the very early (but evolving) universe. This much discussed 'problem of time' in quantum gravity is a consequence of the very different role that time plays in quantum theory (where it is a fixed background parameter) and general relativity (where time is dynamical and depends on the matter-energy distribution).[18] But apart from the problem of how time could emerge from timeless quantum gravity (which is closely related to the 'problem of time', see e.g. Butterfield and Isham 1999), it is also hard to make sense of the "reverse" transition from time in the early universe to timeless quantum gravity.

This problem, which could be called the reverse problem of time, arises as follows: As mentioned above, it is conjectured that quantum gravity (and quantum cosmology) will be particularly relevant for discussing the conditions in the very early universe where quantum effects of gravity are expected to be important. But any discussion of

[18] Kiefer (2004, p. 4) mentions the problem of time as the third main motivation behind the search of a quantum theory of gravity. However, this motivation only makes sense when it is already assumed that general relativity and quantum theory must be brought together in a unified framework (Kiefer's first motivation) or, at least, that there are situations where both of these theories are relevant (Kiefer's second motivation).

the 'early' universe obviously requires that we have a cosmic time concept which indicates that we are close (temporally) to the Big Bang singularity. Indeed, a cosmic time concept is one of the fundamental ingredients in the Big Bang model of the universe.[19] The cosmic time parameter is the proper time of a standard clock (for instance, an imagined perfect wrist watch) at rest in the so-called co-moving frame, and it is this time concept physicists and cosmologists have in mind when discussing conditions in the *early* universe. Indeed, the Planck scale is reached in cosmology by extrapolating backwards the Big Bang model in this cosmic time. Thus, the assumption that quantum gravity (or quantum cosmology) is relevant for the study of the very early universe rests on a solid classical (i.e. not described by a quantum operator) notion of time. But if it is conjectured that timeless quantum gravity is *the* fundamental theory —from which classical physics and concepts can be derived— it appears paradoxical that its central field of application (the early universe) is only defined by a concept (classical cosmic time) which is completely alien to the theory.

The above argument may be seen as a particular instance of Bohr's (and Rosenfeld's) point that classical physics and classical concepts are necessary in order to define and analyze the quantum phenomena. Thus, whether or not one agrees with Bohr that we need classical physics to relate the quantum formalism with measurements, the very definition of the central field of application for quantum gravity rests on classical concepts.[20] As already hinted, it is difficult to circumvent this argument by referring to other interpretations of quantum theory, as also such other interpretations will have to rely on a classical notion of time in order to discuss the early universe. In turn, if we cannot even discuss the central application of quantum gravity without assuming a classical time concept, it is not clear what we should understand by an assumption like 'the ultimate nature of space-time is non-classical' and, more generally, what we should understand by the assumption that 'classical physics can ultimately be explained from the deeper level of quantum gravity'.

It should finally be noted that not all researchers in quantum gravity subscribe to the reductionist conjectures associated with the theory (related to an 'all is quantum' interpretation of quantum mechanics). Thus, Rovelli mentions in the conclusion of his recent book on quantum gravity (2004, p. 370):

---

[19] In general, there is no global time parameter in classical general relativity, but such a parameter is part of the particular Big Bang solution to Einstein's field equations which is assumed to be a reasonable approximate description of our universe.

[20] A similar point may be argued for the case of quantum gravity effects related to black holes. I leave it out here however, since such effects in a sense are even more remote than those related to the very early universe: There is good empirical evidence of a universal expansion and thus of a smaller and denser state of the universe in the past. To reach the very early universe we therefore 'only' need to extrapolate backwards an empirical successful model (see however hesitations to this extrapolations in Rugh and Zinkernagel 2006). By contrast, since Hawking radiation from black holes has not yet been observed, and since nothing is known about the interior of black holes, it would seem that more than extrapolations of successful models are involved in contemplating quantum gravity effects in connection with black holes.

> I see no reason why a quantum theory of gravity should not be sought within a standard interpretation of quantum mechanics (whatever one prefers). [...] We can consistently use the Copenhagen interpretation to describe the interaction between a macroscopic classical apparatus and a quantum gravitational phenomenon happening, say, in a small region of (macroscopic) spacetime. The fact that the notion of spacetime breaks down at short scale within this region does not prevent us from having the region interacting with an external Copenhagen observer.

Now, on Bohr's version of the Copenhagen interpretation, the important point is not the observer but rather that the measurement apparatus is to be described on classical lines. In any case, and in the light of the above discussion, I think the quote expresses a highly recommendable attitude —namely that the search for quantum gravity should not *a priori* exclude specific interpretations of quantum mechanics. In fact, Rovelli's idea that space-time breaks down *within* a small macroscopic space-time region might support the idea that classical space-time concepts (e.g. the macroscopic region surrounding the 'breakdown region') are needed to formulate the 'domain of application' of quantum gravity —and can in this way hardly be seen to be derivable from this theory.[21]

I do not claim, however, that Rovelli would endorse this conclusion. For one thing, the close connection between quantum gravity and quantum cosmology (in which even the universe as a whole is described in quantum terms), and in particular the fact that quantum gravity (also on Rovelli's view) is held to be relevant for the very early universe, would seem to make it difficult to accommodate either classical apparatus or classical concepts (such as a macroscopic region of space-time) at the very early stages of the universe. In any case, I think Rovelli would have to accept that these early stages can only be addressed (and observationally accessed) via the classical time concept of standard cosmology.

5. Summary and Conclusions

A quantum theory of gravity is presently considered the holy grail of theoretical physics. In this manuscript I have tried to argue that the motivations behind the quest for this theory may be resisted, and that — in any case— there are good reasons to doubt that it can be the 'theory of everything'. More specifically:

By reviewing Bohr's conception of quantum mechanics, I have first illustrated one way to question reductionism in physics. Since reductionism serves as a motivation for quantum gravity, this shows a sense in which quantum gravity depends on interpretative issues in quantum theory. Secondly, I pointed out that there is no compelling argument to the effect that quantum gravity is necessary —neither logically nor (at least,

---

[21] Perhaps this could be taken to mean that on a Bohrian understanding of quantum theory, the quantization of the gravitational field, and therefore of space-time, can be done only 'locally' (within a classically described space-time volume). This would be in accordance with the quote by Bohr in section 2 according to which an ultimate measurement apparatus which determine a spatio-temporal framework for the quantum phenomena must be described by classical physics concepts (in particular, this would fit with a relationist account of space-time which links classical space-time to classically described rods and clocks, see e.g. Teller (1999) and Rugh and Zinkernagel (2006b)).

as yet) empirically. Furthermore, as discussed in section 3, even the more modest 'unified physics' expectation that general relativity and quantum theory must somehow come together at high energy could be resisted. Both because of the lack of empirical evidence of any interplay between general relativity and quantum theory and —in particular— due to our limited understanding of the form of such interplay (even at low energy) evidenced by the cosmological constant problem. Finally, I have argued that even if a theory of quantum gravity —in a form in which gravity is quantized— could be constructed, there are good reasons to believe that this would not remove the in principle necessity of classical physics, at least for specifying the field of application of quantum gravity. Obviously, it is hard to make predictions concerning what a quantum theory of gravity will eventually look like. But at least there are reasons to believe that quantum gravity will not supersede present physics in the sense that all other physical theories and phenomena can (even in principle) be derived from, for instance, some future form of string theory.

None of this, of course, should be taken to imply a recommendation that the quest for a quantum theory of gravity ought not to be pursued. All I suggest is that the motivations behind this quest can be resisted, and that a bit of scepticism concerning what could actually be achieved with such a theory seems appropriate.

Acknowledgements

I thank Svend E. Rugh for discussions on the topics of this paper, two referees for valuable comments, and the Spanish Ministry of Education and Science (project HUM2005-07187-C03-03) for financial support.

REFERENCES

Albert, D.Z. (1992). *Quantum Mechanics and Experience*. Cambridge: Harvard University Press.

Amelino-Camelia, G. (2003). "Quantum Gravity Phenomenology", *Physics World*, November 2003 (also available at http://arxiv.org/abs/physics/0311037).

Anderson, P. (1972). "More is different", *Science,* 177, 393-396.

Arndt, M., Nairz, O., Vos-Andreae, J., Keller, C. van der Zouw G., and Zeilinger, A. (1999). "Wave-particle duality of C_{60} molecules", *Nature*, 401, 680-682.

Baez, J. (2001). "Higher-dimensional algebra and Planck scale physics". In C. Callender and N. Huggett (cited below), 177-198.

Barrett, J.A. (2000). "The Persistence of Memory: Surreal Trajectories in Bohm's theory", *Philosophy of Science* 67, 680-703.

——— (2003). "Are Our Best Physical Theories (Probably and/or Approximately) True?", *Philosophy of Science* 70, 1206-1218.

Bell, J. (2004). *Speakable and Unspeakable in Quantum Mechanics* (second edition). Cambridge: Cambridge University Press.

Bohr, N. (1939). "The causality problem in atomic physics". Reprinted in J. Faye and H.J. Folse (1998) (eds.), *The Philosophical Writings of Niels Bohr, Vol. IV: Causality and Complementarity*, Woodbridge: Ox Bow, 94-121.

——— (1958). "Quantum Physics and Philosophy – Causality and Complementarity". In *The Philosophical Writings of Niels Bohr Vol. III, Essays 1958-1962 on Atomic physics and Human Knowledge*, Reprint 1987, Conneticut: Ox Bow (originally, Wiley 1963), 1-7.

Butterfield, J. and Isham, C. (1999). "On the Emergence of Time in Quantum Gravity". In J. Butterfield (ed.), *The arguments of time*. Oxford: Oxford University Press, 111-168.

Butterfield, J. and Isham, C. (2001). "Space-time and the philosophical challenge of quantum gravity". In C. Callender and N. Huggett (cited below), 33-89.
Callender, C. and Huggett, N. (2001). "Introduction". In C. Callender and N. Huggett (eds.), *Physics meets Philosophy at the Planck Scale*. Cambridge: Cambridge University Press, 1-32.
Cartwright, N. (1999). *The Dappled World*. Cambridge: Cambridge University Press.
Cat, J. (1998). "The physicists' debates on unification in physics at the end of the 20th century", *Historical Studies in the Physical and Biological Sciences* 28, 253-299.
DeWitt, B. (1962). "Definition of Commutators via the Uncertainty Principle", *Journal of Mathematical Physics* 3, 619-624.
Falkenburg, B. (1998). "Bohr's principles of unifying quantum disunities", *Philosophia Naturalis* 35(1), 95-120.
Greenberger, D.M. and Overhauser, A.W. (1980). "The Role of Gravity in Quantum Theory", *Scientific American* 242, 54-64.
Hoefer, C. (2003). "For Fundamentalism", *Philosophy of Science* 70, 1401-1412.
Howard, D. (1994). "What makes a classical concept classical? Toward a reconstruction of Niels Bohr's philosophy of physics". In J. Faye and H. Folse (eds.), *Niels Bohr and Contemporary philosophy*. Dordrecht: Kluwer, 33-55.
Joos, E. *et al* (2003). *Decoherence and the Appearance of a Classical World in Quantum Theory*. Berlin: Springer-Verlag.
Kiefer, C. (2004). *Quantum Gravity*. Oxford: Oxford University Press.
Landau, L.D. and Lifshitz, E.M. (1981). *Quantum Mechanics: Non-Relativistic Theory, Volume 3, Third Edition (Quantum Mechanics) (Paperback)*. Oxford: Butterworth-Heinemann.
Rosenfeld, L. (1963). "On quantization of fields". Reprinted in R.S. Cohen. and J. Stachel (eds.) (1979), *Selected Papers of Léon Rosenfeld*. Reidel: Dordrecht, 442-444.
Rosenfeld, L. (1966). "Quantum theory and gravitation". Reprinted in R.S. Cohen and J. Stachel (eds.) (1979), *Selected Papers of Léon Rosenfeld*. Reidel: Dordrecht, 599-608.
Rovelli, C. (2004). *Quantum Gravity*. Cambridge: Cambridge University Press.
Rugh, S.E and Zinkernagel, H., (2002). "The Quantum Vacuum and the Cosmological Constant Problem", *Studies in History and Philosophy of Modern Physics*, 33, 663-705.
——— and Zinkernagel, H. (2006a). "Cosmology and the meaning of time". Forthcoming.
——— and Zinkernagel, H. (2006b). "Time and the cosmic measurement problem". In preparation.
Smolin, L. (2001). *Three Roads to Quantum Gravity*. New York: Basic Books.
Stachel, J. (1999). "Introduction: Quantum field theory and space-time". In T.Y. Cao (ed.), *Conceptual Foundations of Quantum Field Theory*. Cambridge: Cambridge University Press, 166-175.
Tegmark, M. and Wheeler, J. (2001). "100 Years of the Quantum", *Scientific American*, February 2001, 68-75.
Teller, P. (1999). "The ineliminable classical face of quantum field theory". In T.Y. Cao (ed.), *Conceptual Foundations of Quantum Field Theory*. Cambridge: Cambridge University Press, 314-323.
Weinberg, S. (1989). "The cosmological constant problem", *Review of Modern Physics* 61, 1-23.
Weinberg, S. (1993). *Dreams of a final theory*. London: Vintage.
Weinstein, S. (2005). "Quantum Gravity", *The Stanford Encyclopedia of Philosophy (Spring 2006 Edition)*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2006/entries/quantum-gravity/>.
Würtrich, C. (2004). "To Quantize of Not to Quantize: Fact and Folklore in Quantum Gravity". Forthcoming in *Philosophy of Science* (available at http://philsci-archive.pitt.edu/).
Zinkernagel, H. (2002). "Cosmology, Particles, and the Unity of Science", *Studies in History and Philosophy of Modern Physics* 33, 493-516.

Henrik ZINKERNAGEL received his PhD in philosophy of physics from the Niels Bohr Institute in Copenhagen. He has held a post-doc position at the Spanish research institute CSIC and is presently a Ramón y Cajal Research Fellow at the Department of Philosophy at the University of Granada. His main research interests are philosophy of cosmology, philosophy of time and philosophy of quantum physics.

ADDRESS: Department of Philosophy, Campus de Cartuja, 18071, University of Granada, Spain. Email: zink@ugr.es.

ESTADO DE LA CUESTIÓN / *STATE OF THE ART*

3. Filosofía de la mente y de la ciencia cognitiva / *Philosophy of Mind and Philosophy of Cognitive Science*

Josep L. Prades

Filosofía de la Mente: el estado de la cuestión

(Philosophy of Mind: the State of the Art)

Josep L. PRADES

BIBLID [0495-4548 (2006) 21: 57; pp. 315-332]

Uno de los síntomas de la falta de cohesión en una tradición cultural es la dispersión: la ruptura de los supuestos compartidos que definen tanto la forma aceptable de los problemas como la relevancia de las posibles respuestas. Si la filosofía contemporánea es un ejemplo de esa situación, la filosofía de la mente es una de las áreas donde la fragmentación cultural se muestra de una forma más contundente. En ciertas zonas de la discusión filosófica (como la filosofía del lenguaje, la metafísica o la epistemología) existen desacuerdos fundamentales, pero, tales desacuerdos aparecen todavía explícitamente en la controversia contemporánea. Por citar un caso: la discusión entre las concepciones neo-fregeanas y las radicalmente anti-fregeanas es una discrepancia fundamental en filosofía del lenguaje. Sin embargo, no nos resulta extraño que la discrepancia en cuestión sea uno de los objetos explícitos de reflexión en la filosofía del lenguaje actual.

No sucede lo mismo en filosofía de la mente. Aquí, las diferencias fundamentales son el resultado de divergencias más básicas en otras áreas, como filosofía del lenguaje o metafísica. De esas discrepancias surgen percepciones muy diferentes sobre la forma misma de los problemas a tratar. Como consecuencia, en muchos casos, parece que la definición misma del problema es ininteligible para las concepciones alternativas. Por ejemplo, la idea misma de cuáles deben ser las cuestiones fundamentales en filosofía de la percepción depende de consideraciones muy generales sobre el contenido intencional, que, a su vez, están vinculadas a una toma de posición en cuestiones básicas de filosofía del lenguaje y de metafísica. La reacción típica del defensor de una concepción computacional de la percepción ante las propuestas del realismo directo neogibsoniano es la de que su adversario se niega a responder al problema básico: ¿cómo es posible que el cerebro reconstruya un rico contenido perceptivo que va más allá de los estímulos causalmente relevantes que impactan sobre el sistema nervioso? La reacción del adversario es la de insistir en que el enunciado mismo del supuesto problema es ininteligible, que depende de confusiones conceptuales. No es que no haya un territorio en que se discuta sobre esta discrepancia fundamental. Lo hay, pero pertenece a áreas de la discusión filosófica que tradicionalmente no son reconocidas como patrimonio de la filosofía de la mente.

Todo esto afecta al formato que debe tener el presente comentario. Por una parte, la panorámica presentada no puede aislarse de controversias y discusiones fundamentales que pertenecen a otras áreas más básicas de la reflexión filosófica e, incluso, a la concepción misma de la filosofía y de sus relaciones con la ciencia. Por otra parte, los

límites obvios de espacio impiden tratar de dibujar con precisión tales relaciones de dependencia. La estrategia escogida ha sido la de esbozar, en primer lugar, la que considero que es la oposición fundamental que recorre la literatura filosófica actual sobre la mente. Y, en segundo lugar, me preocuparé de apuntar ciertas conexiones entre tal divergencia básica y el tratamiento diferencial de ciertos problemas particulares: causalidad mental, percepción, intencionalidad, conciencia. Naturalmente, tendré que presuponer cierta familiaridad previa del lector con las taxonomías más generales que han rotulado las discusiones en filosofía de la mente en las últimas décadas (funcionalismo analítico, psico-funcionalismo, teorías de la identidad de tipos, teorías de la identidad de casos) para tratar de reconstruir las líneas maestras de una controversia que, muchas veces, cruza las fronteras que parecerían presuponerse en tales taxonomías.

1. La herencia del conductismo lógico

La etiqueta de "conductismo lógico" ha desaparecido prácticamente como el nombre de una posición respetable. El *conductismo lógico* pretendía la reducción de los estados mentales a disposiciones conductuales: un deseo o una creencia eran analizados como meras disposiciones a actuar de cierta forma cuando se dan ciertas condiciones. Una crítica habitual al conductismo lógico ha sido la de que olvidaba tanto el rol causal de los estados internos del organismo como el hecho de que muchas actitudes proposicionales (por ejemplo, creencias sofisticadas) no poseen manifestaciones conductuales características. El *funcionalismo* puede considerarse como un heredero natural del conductismo, con la salvedad de que trata de evitar estos dos problemas básicos. Un tipo de estado mental, para el funcionalismo, se define en términos de cierto rol funcional. Pero el rol funcional no es necesariamente una disposición a cierta forma de conducta manifiesta y puede definirse a través de multitud de relaciones con otros estados internos del organismo. Sea como fuere, es importante observar una discrepancia fundamental en el uso actual de la etiqueta "conductismo": una discrepancia que explica el hecho de que las posiciones que todavía son acusadas de ser conductistas puedan permitirse negar la acusación.

El caso arquetípico serían las *concepciones neo-wittgensteinianas de la mente*. Estas concepciones suponen que un organismo cuya conducta manifiesta y observable fuera indistinguible de la conducta de un ser humano, de un ser al que consideramos como paradigmáticamente ejemplificando estados mentales, sería un organismo que también ejemplificaría estados mentales. Este compromiso es relativamente independiente de cuál sea la organización interna del organismo en cuestión. Seres con estructura material muy diferente tendrían que contar como dotados de mentalidad en la medida en que satisficiesen los criterios ordinarios que rigen nuestras atribuciones en tercera persona, tal y como esos criterios son satisfechos paradigmáticamente por los miembros de la especie humana. ¿En qué sentido esa concepción general de la mente cuenta o no como conductismo lógico? Sus adversarios suelen insistir en que lo es. Sus defensores, lo niegan. La diferencia fundamental de interpretación estriba en el peso que se dé a la cuestión de la *reducción*: esta concepción de la mente no tiene por qué estar comprometida con una reducción de los predicados mentales a disposiciones conduc-

tuales o roles funcionales, caracterizados en términos no intencionales. Ésta es la diferencia fundamental tanto respecto al programa tradicional del conductismo lógico como a las versiones más habituales de funcionalismo. Por supuesto, esa diferencia puede ser acusada de inconsistente, o de tratar de ocupar una posición constitutivamente inestable. La discrepancia se sitúa necesariamente en un territorio mucho más general: ¿es inteligible la existencia de ciertos vínculos a priori —entre ciertas formas de comportamiento y la presencia de estados mentales— sin que los tipos mentales sean reducibles a tipos no intencionales? Hay, pues, una concepción general de la mente humana que puede ser descrita, en líneas generales, como guardando ciertas similitudes con el conductismo lógico y que, en modo alguno, ha desaparecido de la discusión contemporánea. Puede rastrearse en la producción filosófica de los seguidores de Wittgenstein, o de filósofos como Daniel Dennett, Donald Davidson, o el último Hilary Putnam. Para ellos, los criterios ordinarios de atribución de mentalidad están constitutivamente vinculados a la propiedad atribuida, de tal modo que casos paradigmáticos de satisfacción de tales criterios no dejan abierta cuestión alguna sobre si el sujeto atribuido tiene o no estados mentales. Ello, por supuesto, no implica la tesis de que la denominada "psicología popular" es una teoría científica, ni la tesis de que sus generalizaciones son asimilables a leyes que individualizan los poderes causales de las propiedades mentales ni, por supuesto, la tesis de que sea posible reducir los significados de las atribuciones de mentalidad a un vocabulario no psicológico.

No es extraño, sin embargo, que tal concepción general de la mente sea asimilada a una variante de funcionalismo que hoy en día ha pasado a ser minoritaria: el denominado "funcionalismo analítico". El *funcionalismo analítico* se caracteriza normalmente como una tesis sobre el significado de los términos mentales. Considera que nuestros predicados mentales ordinarios pueden ser analizados en términos de relaciones causales entre estímulos, respuestas y otros estados mentales. La restricción que le impone el adjetivo "analítico" es la de que tal reducción está implícita en nuestra competencia ordinaria en el uso de los predicados mentales. Es posible, por supuesto, descubrir empíricamente leyes complejas sobre nuestra psicología a las que sólo nos da acceso el conocimiento científico sofisticado. Un funcionalista analítico tiene que insistir, sin embargo, en que tal tipo de leyes pueden ser violadas por psicologías alternativas sin que eso cuente en contra de la ejemplificación de los estados mentales relevantes. Las leyes que no pueden ser violadas, en ningún mundo posible, son aquellas que forman parte de los criterios de competencia lingüística que rigen las atribuciones ordinarias de mentalidad.

Es importante, para entender muchas de las discusiones contemporáneas, tener presente qué diferencia el funcionalismo analítico de las concepciones de la mente que caracterizaré como no-reduccionistas. En esta presentación, considero que las formas habituales de funcionalismo, incluso las formas de funcionalismo a priori, cuentan como *reduccionistas* en la medida en que se comprometen con la reducción del vocabulario intencional a tipologías funcionales que, en si mismas, son inteligibles sin presuponer nociones intencionales. Se trata de una estipulación terminológica que no coincide con otro uso habitual del término "reduccionismo": muchas veces, se insiste en

que las formas más habituales de funcionalismo no son reduccionistas en la medida en que *no suponen la reductibilidad de los tipos funcionales mismos a tipologías más básicas*. Según mi estipulación terminológica, el funcionalismo analítico es normalmente reductivo: *trata de analizar los predicados mentales en términos no psicológicos*. Otra característica fundamental es su compatibilidad con una teoría materialista de la mente: el funcionalismo analítico (por ejemplo, del recientemente fallecido D. Lewis) puede identificar un estado mental de un sujeto o de una población con un tipo de estado material. Y puede hacerlo sin necesidad de renunciar a su pretensión de ser una teoría a priori: está establecido a priori que un estado (quizás desconocido) identificado por cierto rol funcional es el estado mental de un sujeto. Lo que no es a priori es cuál sea de hecho ese estado material. Su divergencia con las formas a posteriori de materialismo y de funcionalismo estriba en que, para éstas, la investigación empírica puede proporcionarnos criterios de identificación de estados mentales que ignoren y violen los criterios implícitos en las atribuciones cotidianas y accesibles al hablante ordinario. Como veremos, ésa no es una diferencia pequeña. En la aparición recalcitrante de ciertos problemas seculares (por ejemplo, el carácter supuestamente misterioso de la conciencia) es una diferencia crucial, que sitúa a los defensores de las dos formas de funcionalismo en territorios opuestos.

2. Materialismo y psico-funcionalismo

Actualmente, la mayoría de los filósofos analíticos de la mente se auto-clasificarían de *funcionalistas a posteriori*. Uno de los rasgos básicos de la discusión contemporánea no es sólo el carácter dominante de tal posición, sino el hecho de que, con ella, se ha tendido a borrar la distinción entre funcionalismo y materialismo que era común hace algunas décadas. En efecto, una vez se ha aceptado que la caracterización de las propiedades esenciales de un determinado tipo de estado mental depende de descubrimientos empíricos, que pueden imponer taxonomías y clasificaciones no isomorfas con las que son accesibles a priori a un hablante competente, parece artificioso insistir en la diferencia entre el rol causal y el realizador material. Después de todo, estados materiales idénticos en contextos idénticos son idénticos funcionalmente. Y cierta referencia mínima a un contexto causal parece necesaria si la tesis del *materialismo de la identidad* ha de ser plausible. El materialismo contemporáneo trata desesperadamente de no ser chauvinista: son concebibles seres dotados de mente con una composición material distinta a la nuestra. La identidad entre tipos de estados mentales y tipos materiales ha de ser concebida como relativa a una población determinada. Una vez aceptado este punto crucial, las diferencias con el funcionalismo a posteriori se reducen considerablemente: parece que cierta referencia al rol funcional, en sentido laxo, debe ser imprescindible para determinar si dos estados materiales diferentes tienen que contar o no como estados de tipos mentales semejantes cuando sean típicos en poblaciones muy diferentes desde el punto de vista material.

El *materialismo reductivo*, prevaleciente en la literatura filosófica actual sobre la mente, no se contrapone sólo al dualismo de sustancias: el dualismo cartesiano no es una de las opciones relevantes en la discusión contemporánea. El materialismo reductivo ha

de ser contrapuesto a formas no reductivas de materialismo. El *materialismo mínimo o no reductivo*, que comparten todas las posiciones relevantes en la filosofía de la mente actual, se caracteriza por aceptar el hecho de que las propiedades mentales sobrevienen a las propiedades básicas materiales. Dos mundos que fueran idénticos en las propiedades básicas de la materia y en sus relaciones no podrían diferir en otras propiedades: serían estética, económica y mentalmente iguales. No es este el momento de discutir el estatus de esa convicción. Una lectura atenta de la literatura nos convencerá fácilmente de que es defendida, en parte, por motivos conceptuales y, en parte, por motivos empíricos. Por motivos conceptuales: nuestra idea de lo que sea una propiedad básica es exactamente la idea de que pertenece a un tipo que genera esa relación de dependencia. Por motivos empíricos: tenemos buenas razones para creer que la metodología que sigue la física básica nos descubre propiedades cada vez más elementales. Sea como fuere, la tesis del materialismo reductivo no se reduce a la tesis del materialismo mínimo, necesita algo más: una tesis sobre la *dependencia sistemática* (un caso límite sería la identidad) entre los estados mentales y los estados materiales, más fuerte que la mera relación de *sobreveniencia global*. Aceptar, por ejemplo, la tesis de la sobreveniencia global de los hechos económicos sobre los estados físicos del universo no parece, al menos a primera vista, generar relaciones de dependencia metafísicamente interesantes entre el auge de la inflación y los hechos físicos. A primera vista, las distintas situaciones posibles que pueden describirse como situaciones de aumento desorbitado de los precios son situaciones que no tienen relaciones de semejanza interesantes desde el punto de vista de las propiedades físicas básicas involucradas. Desde este punto de vista, dos situaciones de inflación económica no tendrían que parecerse entre sí más de lo que podría parecerse cualquiera de ellas a una situación de deflación. Por supuesto, los defensores del materialismo reductivo piensan que tienen argumentos para mostrar que, al menos en el caso de los predicados mentales, sí hay razones para establecer tales relaciones de dependencia metafísica. Típicamente, identidad o, como mínimo, realización: un estado mental particular tiene los poderes causales que tiene en virtud de ejemplificar los poderes causales propios de ciertas propiedades básicas.

Entiendo, por tanto, que la división más profunda que recorre la filosofía de la mente actual no está bien caracterizada en términos de *"realismo"* versus *"anti-realismo"* ni en términos de *"conductismo lógico" (o "funcionalismo analítico")* versus *materialismo o fisicalismo a posteriori*. La primera de las contraposiciones será normalmente rechazada por la mayoría de los que tendrían que ser catalogados como "anti-realistas". La segunda de las contraposiciones olvidaría dos aspectos fundamentales de la discusión contemporánea: el carácter reductivo del conductismo lógico y el funcionalismo analítico y, por otra parte, el hecho de que éste último es perfectamente compatible con el materialismo a posteriori. La contraposición debe establecerse en términos de una cuestión crucial: *¿exige nuestra concepción científica del mundo que cualquier tipología —incluidas las mentales— que trate de atrapar propiedades genuinas, deba tener relaciones de dependencia sistemática y local con ciertos tipos de estados materiales o funcionales, que tendrían que ser determinados por la mejor ciencia posible?* Los defensores de las formas de materialismo y funcionalismo a posteriori, y también los funcionalistas analíticos que creen que su posición es compatible

con el materialismo responderían que sí. Un *eliminacionista* típico, también respondería que sí. Simplemente, rechazaría que haya genuinas propiedades mentales e insistiría en que la tipología impuesta por nuestros predicados mentales es completamente heteromorfa con la que la ciencia del cerebro nos revela. Es por ello por lo que evitaré cuidadosamente el término *"realismo intencional"* para referirme a la posición dominante en la filosofía de la mente actual y, en su lugar, utilizaré el término "reduccionismo" o "materialismo reductivo". Por supuesto, se trata de una estipulación terminológica con el único propósito de individualizar un conjunto de teorías. Debemos recordar que un funcionalista puede defender que no trata de reducir la eficacia causal de los tipos funcionales a la eficacia causal de los tipos materiales más básicos. Es ésta una discusión sobre la que tendré que volver más adelante. En cualquier caso, mi justificación para la elección de la expresión '*materialismo reductivo*' es que permite agrupar un conjunto de concepciones de la mente humana que comparten dos supuestos fundamentales:

(a) la eficacia causal de un estado mental está determinada, en cada caso particular, por la eficacia causal de un estado material al que es idéntico o que constituye su realización en ese contexto y

(b) los rasgos mentales son reductibles a propiedades materiales o funcionales.

3. Causalidad y Explicación psicológica

La sobreveniencia global de cualquier propiedad sobre las propiedades físicas no es un fenómeno decisivo a favor del reduccionismo. No es *prima facie* obvio, por ejemplo, que la sobreveniencia global de los hechos económicos, deportivos, estéticos o morales sobre los hechos físicos exija la reducción de unos a otros. Por razones expositivas y dialécticas, no voy a entrar en la forma especial en que la ausencia de reductibilidad se usa para fundamentar el eliminacionismo; pues la forma típica de reduccionismo materialista en filosofía de la mente no es eliminacionista. Por tanto, tiene razones para creer que la reducción ha de ser de hecho posible (no sólo que la reducción tendría que ser posible *si* los predicados mentales atraparan genuinas propiedades). Cualquier camino que vaya de la sobreveniencia global a la necesidad de la reducción ha de incorporar algunos supuestos expliquen por qué la opción de la reducción es más atractiva que la mera eliminación. Creo que es justo decir que los argumentos más socorridos, los argumentos que articulan el trasfondo sobre el que debe apoyarse el reduccionismo, dependen de una forma u otra de dos intuiciones fundamentales e íntimamente conectadas: *la eficacia causal de los estados mentales y el tipo particular de éxito que tienen las explicaciones psicológicas ordinarias*.

La idea de que la mente tiene eficacia causal sobre el mundo físico ha sido tradicionalmente uno de los supuestos básicos en el rechazo del dualismo y el epifenomenalismo. Pero la tendencia reduccionista que caracteriza a la filosofía contemporánea de la mente no se basa sólo en el hecho obvio de que hay relaciones causales entre mental y lo físico. No es injusto decir que se ha producido en las últimas décadas lo que puede describirse como una reacción anti-davidsoniana, combinada con la idea de

que debemos explicar satisfactoriamente el éxito de nuestras explicaciones psicológicas ordinarias. Davidson hizo famosa *una teoría de la identidad de casos* (cada suceso mental es también un suceso físico) que pretendía tener la virtud de ser compatible con una forma radical de no-reduccionismo: la identidad de casos era compatible con que no existiera ninguna relación de dependencia sistemática entre los tipos correspondientes. Las relaciones causales debían ser reducidas a relaciones entre sucesos físicos —de donde se seguía, según Davidson, que todos los sucesos causalmente efectivos, incluidos los mentales, debían ser también físicos. Por otra parte, los rasgos mentales que se mencionan en las explicaciones psicológicas ordinarias no se suponían reducibles a propiedades físicas del mundo. Es terreno común en la literatura contemporánea aceptar que una solución como la de Davidson, además de no incorporar, como vio Quine, un criterio no circular para la individualización de sucesos, comportaba la ineficacia causal de las propiedades mentales y, por lo tanto, eliminaba el rol propiamente explicativo de los rasgos mentales: un suceso mental no podría causar uno físico *en virtud de* ejemplificar algún rasgo mental. Por supuesto, desde ese punto de partida común, desde el rechazo del nominalismo davidsoniano, las rutas diferentes hacia el reduccionismo o hacia el anti-reduccionismo nacen de una diferencia fundamental. Un reduccionista pensará que el éxito de nuestras explicaciones psicológicas ordinarias necesita, dados los supuestos mínimos mencionados sobre la eficacia causal de la mente en el mundo físico, cierta relación de dependencia sistemática entre los tipos mentales y los tipos físicos. Un anti-reduccionista, lo negará. Y con ello se verá obligado —en la medida en que niegue la radical separación davidsoniana entre la metafísica de la causalidad y la epistemología de la explicación causal— a revisar algunos de los supuestos sobre la causalidad mental que a un reduccionista le parecen intocables. Por supuesto, dada la variedad de doctrinas que he agrupado bajo el genérico "reduccionismo," no todos sus defensores están igualmente comprometidos con el valor de las explicaciones psicológicas ordinarias. Es compatible con el espíritu del reduccionismo la idea de que muchas de nuestras explicaciones ordinarias no atrapan verdaderas relaciones causales y que muchos de nuestros predicados mentales ordinarios no se verían reflejados en modo alguno en una ciencia psicológica madura. No es compatible con el reduccionismo, si es que quiere evitar el eliminacionismo, insistir en que la mente carece de eficacia causal o en que no puede haber explicaciones causales psicológicas satisfactorias.

Imaginemos la siguiente posibilidad: un genio laplaceano que tiene un conocimiento completo de las leyes básicas del universo, de las propiedades básicas y de su distribución. Un defensor del reduccionismo puede aceptar que ciertas capacidades epistemológicas nuestras podrían serle, en principio, ajenas. Nuestro genio podría no tener acceso a la clasificación del mundo en las clases de semejanza que nuestros predicados mentales establecen. Podría no tener predicados que atraparan algo semejante a nuestras categorías de creencia o intención, por ejemplo. Esa no sería, sin embargo, la cuestión crucial para un reduccionista. La cuestión crucial tendría que ser la de que, si creencias e intenciones son causalmente relevantes y nuestra apelación a ellas es explicativamente relevante, los tipos introducidos por nuestras categorización mental del

mundo *no* pueden ser completamente heteromorfos desde el punto de vista de las propiedades básicas del mundo que el genio laplaceano sí podría detectar. Como Fodor ha argumentado una y otra vez, si la explicación psicológica atrapara casos particulares bajo tipologías arbitrarias desde el punto de vista del verdadero orden causal del cosmos, su éxito y estabilidad serían tan misteriosos como pudieran serlo los de una posible psicología que estableciera relaciones de semejanza entre fenómenos según, por ejemplo, el día de la semana en el que sucedieran (Fodor 2000). O como ha expresado, desde un punto de vista más general, David Lewis:

> La enorme mayoría de los aspectos sobrevenientes del mundo sólo son dados por disyunciones misceláneas infinitas de condiciones físicas infinitamente complejas. Por tanto, detectarlas, nombrarlas o pensar sobre una de ellas en un momento determinado es algo que está más allá de nuestro poder. Los rasgos mentales del mundo no están en absoluto más allá de nuestras capacidades epistémicas. Conglomerados finitos de partículas —nosotros— podemos seguir su rastro. Debe haber, por tanto, alguna suerte de simplicidad en ellos. (Lewis 1995, p. 415)

Es controvertido qué pueda seguirse de este argumento: para formas convencionales de funcionalismo, la conclusión —la simplicidad subyacente de los tipos mentales— es compatible con cierta autonomía explicativa, incluso causal, de los tipos funcionales respecto a los tipos neurofisiológicos. Para los protagonistas de la resurrección de las viejas teorías de la identidad de tipos, como, por ejemplo, Kim, la simplicidad exigible sólo puede ser garantizada por la identidad de los tipos mentales-funcionales con los tipos más básicos (J. Kim 1998, y 2005). Por otra parte, el rechazo de este argumento general es lo que caracteriza las posiciones que he descrito como anti-reduccionistas. Todas ellas han de rechazar como confuso el uso de la expresión "seguir su rastro" que aparece en la cita de D. Lewis. Parte del problema radica en que ese vocabulario parece exigir que la conducta de seguir el rastro (por ejemplo, la conducta a través de la cual se manifiesta la percepción de un estado mental en otro ser) habría de ser caracterizada de una manera neutra, independiente, del objeto que se supone rastreado. Ese es, en términos generales, el supuesto que rechazan las concepciones neo-wittgensteinianas de la mente. Posiciones a las que se ha aproximado recientemente uno de los que fue en su momento padre fundador del funcionalismo: H. Putnam. El problema con el argumento general que hemos mencionado, para estas concepciones, sería simplemente que sienta la cuestión por anticipado. El no-reduccionista dirá que el argumento de Lewis necesita de un supuesto discutible: el supuesto de que el estado mental de percibir intencionalidad en otros es un mero estado de detección, un estado caracterizable independientemente de cualquier vocabulario intencional. Por el contrario, y utilizando la terminología del propio Lewis, su adversario sostendrá que nosotros, conglomerados finitos de partículas, podemos formarnos estados mentales sobre estados mentales de otros seres (lo que Lewis denomina "seguir el rastro") sin necesidad de postular que haya orden ni simplicidad (en términos de propiedades básicas) en los estados mentales atribuidos, precisamente porque tampoco hay orden ni simplicidad en el estado de percibirlos. Por supuesto, una posición semejante necesita de una justificación mucho más complicada de la que podría describirse en un comentario de este tipo. Tendría que mostrarle al reduccionista que la defensa esbozada del anti-reduccionismo no supone un compromiso con el eliminacionismo. ¿Por qué no equi-

vale a decir que la misma (supuesta) percepción de orden y simplicidad que está en la base de nuestra competencia ordinaria es, ella misma, una ilusión? En todo caso, ¿cómo se explica el hecho de que parezca que no es completamente ilusoria? Estas son cuestiones típicas que muestran que los supuestos ocultos en muchas de las discusiones contemporáneas en filosofía de la mente se remiten a zonas mucho más generales de la discusión filosófica. Parece que la única manera en que el anti-reduccionista puede defenderse es aceptando una concepción del lenguaje y la intencionalidad cercana a la defendida por el último Wittgenstein.

Sea como fuere, la literatura contemporánea parece encontrar dificultades para elaborar una noción coherente de causalidad mental que sea capaz de mostrar que el reto que plantea la cita de Lewis puede ser respondido satisfactoriamente. Uno de los supuestos fundamentales del funcionalismo es, por supuesto, que las propiedades mentales son causalmente eficaces y que, para serlo, su eficacia causal ha de estar sistemáticamente relacionada con la eficacia causal de las propiedades básicas del mundo. Otro es que ha de explicarse satisfactoriamente cierta autonomía de los tipos mentales respecto a los tipos estrictamente neurofisiológicos. La tensión parece evidente: la aceptación de tal autonomía mínima pone en cuestión el tipo requerido de perspicua dependencia sistemática entre los procesos causales involucrados. En los últimos años, J. Kim ha sido el más destacado defensor de la tesis de que el funcionalismo que no acepte una reducción de los tipos funcionales a los tipos materiales correspondientes es una actitud incoherente. La intuición básica del argumento de Kim parece clara: la supuesta autonomía explicativa de los procesos de orden superior (mentales, funcionales) no puede ser más que un fenómeno ilusorio si atendemos a ciertos requisitos plausibles sobre el funcionamiento de toda explicación genuinamente causal. So pena de negar el cierre causal del mundo físico o aceptar que los casos de causalidad mental son casos de sobredeterminación causal, parece necesario aceptar que la causa mental de un efecto físico ha de ser ella misma una causa física. Y si las causas mentales son causas físicas, las propiedades mentales son propiedades físicas. La supuesta autonomía explicativa de las tipologías mentales y funcionales debe descansar, en contra de las pretensiones habituales de los defensores del funcionalismo, en la identidad de las propiedades mentales y las propiedades básicas del mundo. En cierto sentido, los últimos años del siglo XX pueden describirse como una época de pérdida de la inocencia respecto a la viabilidad del modelo funcionalista en tanto que esencialmente distinto al materialismo de la identidad. Los más destacados intentos de bloquear el materialismo de la identidad se caracterizan por tratar de extraer lecturas metafísicas de ciertas asimetrías explicativas entre las propiedades básicas del mundo y las propiedades de orden superior. Stephen Yablo, por ejemplo, ha insistido en la última década en la proporcionalidad de la relación causal: las causas funcionales/mentales son más proporcionales a los efectos funcionales/intencionales, mantienen relaciones contrafácticas más robustas con ellos, que las supuestas causas neurofisiológicas o físicas (Yablo 2003). Tal tipo de teoría parece encajar perfectamente en una metafísica de las propiedades de acuerdo con la cual las propiedades son conjuntos de poderes causales y los poderes causales de las propieda-

des de orden superior (o determinables) son menos específicos que los poderes causales de las propiedades de orden inferior (o determinados), en el sentido en que éstas últimas son una manera especial de ejemplificar la propiedad de orden superior (Shoemaker 2001). Nadie duda de que, en cierto contexto explicativo, apelar a una propiedad de orden superior puede ser más relevante que apelar a una de orden inferior. Lo que se nos propone es que el fenómeno no afecta sólo a la relevancia pragmática o epistémica de la explicación causal: se trataría de un fenómeno metafísicamente relevante. Es discutible que tal tipo de movimiento no pueda ser acusado por un defensor de las teorías de la identidad de tipos de cometer una petición de principio. Ello estaría conectado con las dificultades para explicar la relación de multiple-realización como un caso especial de la relación metafísica entre determinable y determinado (Ehring 1996). Por otra parte, los defensores de posiciones radicalmente anti-reduccionistas pueden considerar esta situación dialéctica como una reducción al absurdo de los supuestos fundamentales del reduccionismo. De hecho, la misma noción de realización sería para ellos sospechosa: una suerte de engendro gramatical que poseería al mismo tiempo los rasgos de los universales y de los particulares (Steward 1997). Otra manera de expresar esta reacción sería la de decir que la teoría que se nos propone entraña la ausencia de criterios metafísicos de individuación del (supuestamente) único realizador, en cada caso particular, de la propiedad funcional (Putnam, 2000, y Corbí y Prades 2000).

4. Externismo, contenido estricto, percepción e intencionalidad

Hay puntos de fricción mucho más específicos entre una concepción reducionista y una concepción anti-reduccionista de la mente. Un argumento anti-reduccionista típico consiste en apelar a ciertas intuiciones sobre la individuación del contenido que no parecen encajar fácilmente con el reduccionismo. Por ejemplo, el *externismo*, el hecho de que las relaciones efectivas con el medio sean constitutivas de la manera en la que individualizamos contenidos intencionales. Es cierto que, en principio, el reduccionismo puede combinarse con el externismo de maneras muy distintas. Un caso obvio es el de Fred Dretske, quien defiende una teoría externista incluso de los qualia perceptivos y una concepción reduccionista de la mente (Dretske 1995 y 1996). Otro caso lo constituirían las formas de funcionalismo de "largo alcance" que consideran que los procesos funcionales relevantes tienen que incluir ciertos rasgos del medio. Otra manera, típica, es la de introducir la noción de *contenido "estricto"* como la noción relevante en una explicación psicológica genuina. Lo que comparten estas concepciones es la necesidad de que esté metafísicamente determinado el componente interno del contenido. Que una ciencia psicológica madura tenga que apelar a contenidos estrictos o amplios, por ejemplo, no es tan relevante en este contexto como el supuesto de que, para explicar causalmente la conducta de un organismo, debe estar metafísicamente determinado el factor interno que es crucial en la explicación. Y no es fácil establecer los principios por los cuales tal determinación sería posible. Decir que dos individuos físicamente iguales comparten tal factor interno no es avanzar demasiado. Básicamente, porque el factor interno se supone que ha de ser compartido por individuos que no

sean exactamente iguales en todos los respectos. El problema puede ser planteado en términos muy generales: una vez que aceptamos que nuestra manera ordinaria de clasificar contenidos mentales está comprometida con el contenido "amplio", podemos clasificar conductas que son semejantes desde el punto de vista físico como formas muy distintas de acción intencional, y viceversa. Y ello hace mucho más verosímil uno de los puntos básicos de la resistencia del anti-reduccionista al argumento de Lewis que previamente citado: conductas muy diferentes desde el punto de vista físico pueden contar como el mismo tipo intencional, y conductas muy semejantes desde el punto de vista físico pueden contar como casos muy diferentes desde el punto de vista intencional. Necesitamos principios no *ad hoc* para individualizar los poderes causales que vayan más allá de la trivialidad de decir que siempre podemos conseguir bases de sobreveniencia suficientemente amplias tales que podemos asegurar que, dentro de ellas, se ejemplifica una determinada propiedad mental. Esa base acabaría siendo tan amplia que también garantizaría la ejemplificación de otras propiedades mentales independientes, por lo que no se habría establecido la base de sobreveniencia de esa propiedad mental particular que las estrategias reduccionistas requieren.

La idea de contenido estricto está vinculada a lo que tradicionalmente se ha denominado la *"Teoría Computacional de la Mente"*, la doctrina de que los poderes representacionales de la mente sólo pueden ser explicados si postulamos particulares mentales como portadores de tales poderes (representaciones) y explicamos las transacciones semánticas como el resultado de transacciones sintácticas, es decir, como interacciones causales entre diversas representaciones, en virtud de sus propiedades básicas y no representacionales. Ciertamente, un proceso computacional es a la vez semánticamente evaluable y explicable en virtud de los rasgos sintácticos, físicos, de los portadores de la representación. Pero un proceso computacional típico parece requerir la adjudicación arbitraria de contenido —en función de los intereses del diseñador y/o del usuario— a las representaciones básicas. No es de extrañar, pues, que la oposición típica a las teorías computacionales dependa de la intuición de que la apelación a procesos computacionales requeriría apelar a mecanismos básicos y no computacionales de fijación de contenido. El problema crucial es que no parece posible contar una historia creíble sin abrir la puerta a intuiciones externistas y, de ese modo, las concepciones computacionales de la percepción pierden su atractivo fundamental. Entre otras cosas porque, en ese caso, parece posible argumentar que es una ilusión suponer que la semántica de la mente *debe* estar necesariamente respaldada por clases de equivalencia desde el punto de vista de la neurofisiología. Sería perfectamente compatible con los datos empíricos el supuesto de que las clases de equivalencia intencionales fueran heteromorfas con clases de equivalencia más básica. La adjudicación de un contenido estable a partir de manifestaciones físicas bien distintas (y la adjudicación de contenidos distintos a manifestaciones físicas muy similares) podría ser explicada por el hecho de que la tipología intencional sólo es accesible desde una determinada actitud ante el mundo, ante la manera en que el agente atribuido se manifiesta en un mundo de objetos.

Por el contrario, los defensores de versiones más o menos fuertes de la teoría computacional insisten en que hay ciertos datos que serían inexplicables de otra manera: la composicionalidad y sistematicidad del lenguaje por ejemplo, o la infradeterminación del contenido perceptivo a partir de la escasez de los *inputs* causales. La reacción típica del anti-reduccionista es la de acusar al adversario de una descripción sesgada de los supuestos datos. La capacidad innata de descubrir regularidades en el lenguaje no requiere postular el tipo de representaciones internas que postula la teoría computacional. Lo mismo sucede con la supuesta infradeterminación de los estímulos perceptivos: después de todo, la teoría computacional depende de dos supuestos que están en tensión y que parecen difícilmente conciliables sin estipulaciones *previas* sobre los portadores primitivos de contenido. Por una parte, depende de la intuición galileana de que la percepción es una relación causal y de que, como en toda cadena causal, se da una dependencia contrafáctica mucho más robusta entre el contenido perceptivo y los estímulos próximos que la que se da entre el contenido perceptivo y los estímulos lejanos: el-contenido-perceptivo-de-tigre es explicado mejor por la distribución de luz en la retina que por la presencia real y verdadera del tigre. Por otra parte, la teoría computacional necesita el supuesto de que los estímulos próximos no determinan el contenido: sin tal supuesto no puede describirse el fenómeno de la pretendida infradeterminación. Una manera anti-reduccionista típica de resolver la tensión es la de defender el realismo directo y una concepción disyuntiva del contenido perceptivo, según la cual el contenido perceptivo no es el mismo cuando el tigre está delante y cuando se produce la alucinación perfecta (McDowell 1994). De hecho, el movimiento equivale a negar la intuición galileana básica sobre la (relativa) independencia contrafáctica entre el contenido perceptivo y el estímulo externo. Y, desde ese punto de vista, las teorías computacionales tratarían de postular entidades *ad hoc* (representaciones) para salvar dos principios que son incoherentes. Sea cual sea el futuro desarrollo de esta controversia fundamental, es justo decir que las defensas más sofisticadas de las teorías computacionales son conscientes de algunas de sus dificultades: por ejemplo, el carácter local de los procesos sintácticos postulados sobre representaciones particulares parece difícil de encajar con el carácter holista de muchos procesos cognitivos fundamentales. Toda la controversia suscitada en relación a la relevancia filosófica de los modelos conexionistas está vinculada a esta cuestión fundamental (Fodor 2000b y Smolensky 1993). Por supuesto, el adversario de las teorías computacionales interpreta el problema como un subproducto de una confusión fundamental: no es que niegue a priori la existencia de procesos locales y causalmente relevantes para la génesis del contenido perceptivo. Niega el argumento por el cual sólo la existencia de relaciones causales entre los rasgos sintácticos de las representaciones puede resolver ciertas perplejidades supuestamente asociadas al éxito de procesos cognitivos básicos[1]. El que los seres humanos representen el mundo no requiere que lo hagan por medio de representaciones internas. Para un reduccionista, su adversario parece condenado a negar la

[1] Una crítica exhaustiva a los supuestos fundamentales de las teorías computacionales de la mente puede encontrarse en Bennett y Hacker (2004).

intencionalidad, a afirmar que es sólo un asunto de interpretación de ciertas capacidades de nivel personal, del tipo de actitud que, contingentemente, adoptamos sobre la conducta abierta de nuestros semejantes. Para un anti-reduccionista, su adversario —al igual que los epistemólogos clásicos del XVII y del XVIII— es incapaz de ofrecer una explicación satisfactoria del hecho de que las supuestas representaciones internas adquieran el contenido intencional que se les supone. De nuevo, la discrepancia en este punto nos remite a un problema mucho más general: el del ajuste entre un sistema de conceptos y el mundo. Un reduccionista sólo puede mantener su acusación si insiste en que nuestras tipologías intencionales ordinarias —aquellas mediante las cuales "interpretamos" la conducta de nuestros semejantes— tienen una relación de ajuste con el mundo que puede ser descrita sin presuponerlas. Un anti-reduccionista niega que ese supuesto sea —en general, no sólo en relación con el vocabulario mental— inteligible y niega, además, que de su negación se siga que no hay hechos que conviertan en verdaderos o falsos nuestras atribuciones ordinarias de intencionalidad.

Para un reduccionista, la propuesta típica de su adversario —el *"realismo directo"* perceptivo— es una negativa a plantearse si siquiera las cuestiones fundamentales. Su adversario insiste en que el realismo directo no es una teoría alternativa: es un rechazo de los problemas mismos que se supone que han de ser respondidos por las teorías que defienden variantes más o menos sofisticadas de la vieja idea de los intermediarios. Como anteriormente comentábamos, el reduccionismo necesita suponer que hay una explicación razonable de la estabilidad mínima y la capacidad de predicción en la psicología de sentido común. En cierto sentido, y paradójicamente, el anti-reduccionista tiene un punto de coincidencia con él en este aspecto. Un punto sobre el que trata de alterar los términos de la discusión. El hecho de que nuestra percepción de la mentalidad de otros no sea sistemáticamente falsa es, para el anti-reduccionista, un caso particular de un fenómeno mucho más general que, según él, su adversario ignora sistemáticamente: el carácter genuinamente explicativo y exitoso ("*factive*") de la noción de percepción, que es más básica que una noción de contenido perceptivo supuestamente neutra respecto al éxito (verdad) de la percepción[2]. Y, en general, la idea de que la atribución de creencias está regida por el principio de maximización de la verdad y la racionalidad. El anti-reduccionista tenderá a considerar que ese principio, como un principio metodológico o epistémico que rige la atribución de creencias y, en general, de intencionalidad, puede servir para legitimar su negativa a entender la propuesta de su adversario. Se sigue de tal principio una asimetría epistémica fundamental entre las bases sobre las que atribuimos estados mentales y las bases sobre las que clasificamos, por ejemplo, los procesos causales físicos. Ése era el fundamento de la distinción radical propuesta por Davidson entre la metafísica de la mente y su epistemología. Las formas actuales de anti-reduccionismo, lo he mencionado, no siguen en eso a Davidson. Insisten, en cambio, en ciertas consecuencias metafísicas de tal diferencia

[2] Una defensa de la tesis de que los estados que incorporan conocimiento (conocer, percibir…) son psicológicamente más explicativos que sus correlatos "internos" y no necesariamente exitosos (creer, tener un determinado contenido perceptivo…) puede encontrarse en Williamson (2000).

epistémica. Sería injusto considerar que, al hacer tal cosa, están sentando por anticipado la cuestión contra su adversario reduccionista. Al fin y al cabo, el reduccionista parte, como hemos visto, de la necesidad de explicar el éxito (epistemológico) de nuestras atribuciones ordinarias de mentalidad. Considera, también, que el éxito epistemológico debe ser, en último término, el mejor valedor de la metafísica reduccionista que propone. Insiste, como vimos en la cita de Lewis, en que nuestra capacidad de detectar con éxito ciertos patrones en la conducta de nuestros semejantes es el mejor argumento a favor de la existencia de cierta relación de dependencia metafísica entre los tipos mentales y los tipos físicos.

De hecho, muchos de los supuestos subyacentes en las teorías computacionales de la percepción pueden considerarse una sofisticación de lo que desde Galileo se considera como la imagen científica mínimamente ortodoxa. La que se supone que se sigue del hecho de que la percepción es una relación causal ordinaria. La percepción se conceptualizó, por Galileo y Locke, básicamente en términos causales, y se consideró como un hecho obvio el que, en el último eslabón de la cadena causal, se produce un contenido que representa el mundo. Nunca nos dicen —como tampoco lo hacen los otros grandes epistemólogos del XVII y del XVIII— cómo es posible formarse contenidos perceptivos que se suponen sistemáticamente falsos (vgr. sobre colores) y que, por tanto, no pueden ser explicados en términos de las relaciones causales efectivas. Las dificultades reconocidas por todos —defensores y detractores— en las teorías computacionales de la percepción guardan un extraordinario parecido con los problemas asociados al representacionalismo clásico. La percepción es una relación causal. Por una parte, el contenido de la percepción está fijado por los últimos eslabones de la cadena causal correspondiente, como opuestos al estímulo remoto. Por otra, el efecto último de la cadena causal es, en el caso de la percepción, una representación. La tensión entre ambos supuestos resulta obvia en la manera en que se identifica el problema básico de la filosofía de la percepción: el problema de explicar cómo los últimos eslabones externos de la cadena causal —por ejemplo, los estímulos luminosos que llegan a la superficie de la retina— son procesados por el cerebro para dar lugar a un contenido que parece ir más allá de la información que está intrínsicamente en ellos. Aunque ya no se concibe que la función básica de los procesos internos de procesamiento sea la de justificar nuestra creencia perceptiva, la deuda a pagar sobre la génesis del contenido intencional en el caso de la percepción es todavía enorme. Antes de contar una historia inteligible sobre cómo es posible reconstruir la riqueza del estímulo externo a partir de las características intrínsecas del estímulo cercano (por ejemplo, la distribución de luz en la retina), necesitamos una historia creíble que nos explique cómo pueda haber contenido a procesar en las características intrínsecas de los estímulos: sin ese supuesto, no hay contraposición posible entre estímulos pobres en contenido y contenido perceptivo exuberante. No parece fácil contar una historia semejante sin

conceder al adversario algo fundamental: ciertas intuiciones externistas sobre los procesos que fijan el contenido[3].

5. La conciencia, de nuevo…

La idea de que el modelo dominante sobre la percepción convierte a la intencionalidad en un misterio puede verse reforzada por el resurgimiento de una controversia que parecía que había perdido la virulencia que tuvo hasta mediados del siglo XX: el problema de la conciencia. La percepción es crucial porque los estados perceptivos tienen a la vez contenido intencional (representan el mundo) y aspectos fenomenológicos (son un caso de sentirse de cierta manera). Locke, por ejemplo, aceptó sin ambages que su concepción de la percepción convertía la aparición de los aspectos fenomenológicos de la misma en un verdadero misterio: las leyes científicas sobre la percepción que la mejor ciencia pudiera nunca descubrir, serían compatibles con la ausencia de aspectos fenomenológicos o con su distribución completamente distinta en el mundo material. Locke negaba que pudiéramos explicar alguna vez la supuesta sobreveniencia de la fenomenología de la percepción sobre los hechos físicos. De un modo similar, en su *Naming and Necessity*, S. Kripke planteó una objeción fundamental a lo que en este comentario se ha denominado reduccionismo a posteriori (Kripke 1980). Básicamente, el argumento se basaba en la imposibilidad de describir ciertos posibles descubrimientos empíricos como descubrimientos de la esencia oculta de los estados fenomenológicos. Kripke pareció suponer que todo descubrimiento a posteriori de la esencia oculta debería presuponer una diferencia entre los rasgos epistémicamente accesibles del fenómeno —aquellos a los que accede cualquier hablante competente— y los rasgos que constituirían su esencia oculta. En el caso del agua, el carácter de ser líquido, incoloro, inodoro etc., por contraposición a la composición química (H_2O). Esa diferencia debería generar la posibilidad de algo que pareciera normalmente acuoso sin ser agua, y al revés. En el caso de los estados fenomenológicos, definidos en términos de cómo se siente el sujeto, es ésa la diferencia que sería ininteligible. La década de los 90 del siglo pasado fue testigo de un resurgimiento de la disputa. Los argumentos neo-kripkeanos tenían la consecuencia de desbaratar la posibilidad misma de una teoría coherentemente materialista y a posteriori de la conciencia: tal teoría necesitaría establecer las condiciones de identidad metafísica de los estados de conciencia a partir de ciertos descubrimientos empíricos. Por una parte, se produjeron argumentos para tratar de mostrar que existían alternativas que el argumento original de Kripke no había tenido en cuenta. Alternativas basadas, por ejemplo, en la peculiaridad de los conceptos fenoménicos. La consecuencia del argumento de Kripke debería ser, según este tipo de estrategias, no la negación de las identidades a posteriori en el caso de la conciencia, sino la relevancia de la peculiaridad de los conceptos fenoménicos. Con ellos el sujeto se pone en una relación epistémica especial con ciertos estados neurofisiológicos que son

[3] Para una visión panorámica sobre la relación entre las discusiones actuales en filosofía de la percepción y las discusiones de la epistemología tradicional, puede consultarse Noë y Thompson (2002).

—idénticos a— los estados fenomenológicos. A pesar de ello, los mismos estados que son representados por los conceptos fenoménicos pueden ser conocidos a través de otros medios de representación (los que se incluyen, por ejemplo, en una descripción neurofisiológica de los mismos) (Loar 1997, y Block y Stalnaker 1999). Este tipo de reacción ha sido considerada desde diversos puntos de vista como una maniobra *ad hoc*, para salvar los compromisos mínimos del materialismo a posteriori. Se ha argumentado, por ejemplo, que la solución propuesta viola claramente las condiciones de inteligibilidad de la noción misma de identidad a posteriori (basadas en que haya siempre una diferencia entre lo presentado y el modo de presentación) y que la apelación al carácter especial de los conceptos fenoménicos en nada ayuda a entender el vocabulario de la identidad en este caso (Chalmers 1996 y 2003, y White 2006).

La cuestión es crucial porque afecta a la inteligibilidad misma de la forma dominante de materialismo en la literatura actual. En su libro más reciente, el propio J. Kim reconoce que su modelo reduccionista debe dejar a un lado, por este tipo de motivos, los aspectos fenomenológicos de la conciencia (Kim 2005). Y este tipo de problemas están a la base de la resurrección de formas neo-cartesianas de "misterianismo": la idea de que, en el fondo, no podemos entender la emergencia de los aspectos fenomenológicos de la conciencia a partir de ciertos complicados procesos neurofisiológicos (Levine 2001). Como ha advertido recientemente N. Block (un autor que se situaría claramente en el campo del materialismo a posteriori), el problema es mucho más grave: aun suponiendo que existiera una solución aceptable al problema que respetara las premisas básicas del reduccionismo a posteriori, el precio a pagar, impuesto por la *forma* que necesariamente debería tener la supuesta solución, sería extraordinario: nos dejaría inermes frente a los problemas epistemológicos que tradicionalmente se han venido asociando a la fenomenología —espectros invertidos, otras mentes (Block 2002). Una vez realizada la correspondiente identidad a posteriori, no tendríamos ningún motivo racional para asumir que seres con neurofisiología —o con una constitución interna— radicalmente distinta a la nuestra tienen conciencia. Y tal conclusión debería ser completamente independiente de las capacidades y disposiciones funcionales que les fueran justificadamente atribuibles desde la tercera persona. El precio parecería claramente excesivo. Y, sin embargo, las teorías a las que ese problema no se plantearía (formas de funcionalismo a priori, o lo que en este artículo he denominado "anti-reduccionismo") no parecen gozar de un respaldo mayoritario en la controversia contemporánea. Todo lo contrario.

6. Significado y atribuciones de mentalidad

Terminaré trazando una última conexión entre las dos posiciones generales que he tratado de dibujar (reduccionismo *versus* anti-reduccionismo) y ciertas controversias básicas en filosofía del lenguaje. Es importante ver que el anti-reduccionista puede utilizar, y utiliza, en favor de su posición ciertas concepciones del lenguaje y la intencionalidad sobre las que se puede argumentar con independencia de su relevancia para el problema de la eficacia causal de la mente. Por ejemplo, una concepción wittgensteiniana de la estabilidad en el uso de nuestros predicados y nuestras atribuciones de mentalidad,.

O, por mencionar discusiones más recientes, una concepción contextualista y/o pragmatista del significado y de las atribuciones de intencionalidad. La conexión con este tipo de discusiones es múltiple. Mencionaré sólo su aspecto más general. Una teoría reduccionista sobre la mente necesita ser aplicable a los estados representacionales y a los contenidos intencionales. El tipo de "realismo robusto" que defiende requiere que los poderes causales de la representación atribuida sean sistemáticamente dependientes de los poderes causales de la base física, o neurofisiológica, que constituye, o es idéntica, o realiza el estado representacional atribuido al atribuir a alguien la creencia de que mañana lloverá. Si hubiera algo de verdad en los proyectos contextualistas en semántica, tendría repercusiones sobre el tipo de estado que atribuyo cuando atribuyo a alguien tal creencia. Si la atribución de tal creencia está radicalmente afectada por factores contextuales y pragmáticos, sería difícil defender la existencia de ciertos poderes causales comunes y describibles en términos no intencionales que pudieran sustentar la mínima unidad causal de la representación atribuida tal y como la concibe el proyecto reduccionista. Es parte del proyecto contextualista en semántica apelar a la pertinencia de utilizar el mismo predicado o el mismo tipo de atribución con finalidades y supuestos muy distintos. La percepción de la semejanza relevante para el uso del mismo tipo de predicado es lo que nos hace competentes lingüísticamente. Pero, si el contextualismo semántico tuviera algo de razón, sería inevitable su extensión a las atribuciones de intencionalidad. Y, con ello, parecería fuera de lugar el supuesto de que nuestras representaciones —los estados intencionales que atribuimos al atribuir intencionalidad— tienen los rasgos que un defensor del reduccionismo necesita: el hecho de que un agente ejemplifique un estado intencional de determinado tipo es relativo al contexto dialéctico en que se produce la atribución. En este punto es importante advertir que el reduccionista no puede argumentar que, aunque el proyecto contextualista en semántica tuviera básicamente razón, ello no implica que no haya verdaderos estados representacionales que no son atrapados por nuestras atribuciones ordinarias y que son los que están afectados por la tesis general del reduccionismo. El reduccionista necesita aquí, como lo necesita en el caso del externismo de lo mental, una historia creíble sobre la manera en que tales representaciones metafísicamente genuinas están relacionadas con nuestras atribuciones ordinarias; pues, recordémoslo, el tipo de reduccionismo contemporáneo que estoy discutiendo es profundamente antidavidsoniano: asume que una parte fundamental de su proyecto es el de hacer inteligible la estabilidad y el éxito de nuestras explicaciones psicológicas ordinarias[4].

REFERENCIAS

Alter, T. y S. Walter (eds.)(2006). *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*. Oxford University Press.

Bennett, M., y P. Hacker (2003). *The Philosophical Foundations of Neuroscience*. Malden, MA: Blackwell.

[4] El ataque más radical en la literatura contemporánea a la distinción tradicional entre semántica y pragmática, aplicado a la metafísica del contenido intencional, se encuentra en Travis (2000).

Block, N. (2002). "The harder problem of consciousness", *Journal of Philosophy* 99, pp. 391-425.
———, O. Flanagan y G Güzeldere (eds.)(1997). *The Nature of Consciousness: Philosophical Debates*. Cambridge MA. MIT Press.
———, y R. Stalnaker (1999). "Conceptual analysis, dualism, and the explanatory gap", *Philosophical Review* 108, pp. 1-46.
Corbí, J.E., y J.L. Prades (2000). *Minds, Causes, and Mechanisms. A Case against Physicalism*. Oxford: Blackwell Publishers.
Chalmers, D. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press.
——— (2003). "Consciousness and its place in nature", en Stich y Warfield, *Blackwell Guide to Philosophy of Mind*. Oxford: Blackwell.
Dretske, F. (1995). *Naturalizing the Mind*. Cambridge, MA. MIT Press.
——— (1996). "Phenomenal externalism or if meanings ain't in the head, where are qualia?", *Philosophical Issues* 7, pp. 143-158.
Fodor, J. (1998). *In Critical Condition*. Cambridge, MA. MIT Press.
——— (2000a). "A Science of Tuesdays", *London Review of Books* 22.
——— (2000b). *The Mind doesn't work that way*. Cambridge, MA: MIT Press.
Gillet, C., y B. Lower (2001). *Physicalism and Its Discontents*. Cambridge: Cambridge University Press.
Guttenplan, S. (1995). *A Companion to the Philosophy of Mind*. Oxford: Blackwell.
Lewis, D. (1995). "David Lewis: Reduction of Mind", en S. Guttenplan, *A Companion to the Philosophy of Mind*. Oxford: Blackwell.
Kim, J. (1998). *Mind in a Physical World : an essay on the mind-body problem and mental causation*. Cambridge, MA: MIT Press.
——— (2005). *Physicalism, or something near enough*. Princeton: Princeton University Press.
Kripke, S. (1980). *Naming and Necessity*. Cambridge, MA: Harvard University Press.
Levine, J. (2001). *Purple Haze: The Puzzle of Consciousness*. Oxford: University Press.
Loar, B. (1997). "Phenomenal states", en N. Block, O. Flanagan y G. Güzeldere (eds.), *The Nature of Consciousness: Philosophical Debates*. Cambridge MA: MIT Press.
McDowell, J. (1994). "The Content of Perceptual Experience", *Philosophical Quarterly* 44, pp. 190-205.
Noë, A., y E. Thompson (eds.)(2002). *Vision and Mind: Selected Readings in the Philosophy of Perception*. Cambridge, MA.: MIT Press.
Putnam, H. (2000). *The Threefold Cord*. New York: Columbia University Press.
Shoemaker, S. (2001). "Realization and Mental Causation", en Gillet and Lower, *Physicalism and Its Discontents*. Cambridge: Cambridge University Press.
Smolensky, P. (1993). *On the proper treatment of connectionism*. Cambridge, MA: MIT Press.
Steward, H. (1996). *The Ontology of Mind*. Clarendon Press: Oxford.
Stich, P., y T. Warfield. (2003). *Blackwell Guide to Philosophy of Mind*. Blackwell: Oxford.
Travis, C. (2000). *Unshadowed Thought: Representation in Thought and Language*. Cambridge, MA: Harvard University Press.
White, S. (2006). "Property dualism, phenomenal concepts, and the semantic premise", en T. Alter y S. Walter (eds.), *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*. Oxford University Press.
Williamson, T. (2000). *Knowledge and its Limits*. Oxford: Oxford University Press.
Yablo, S. (2003). "Causal Relevance", *Philosophical Issues* 13, pp. 316-329.

Josep L. PRADES es profesor titular de Lógica y Filosofía de la Ciencia en la Universidad de Girona. Anteriormente, fue profesor en la Universidad de Murcia. Sus publicaciones incluyen trabajos sobre la filosofía del segundo Wittgenstein, Epistemología y Filosofía de la Mente. Es autor, junto con Josep Corbí, del libro *Minds, Causes and Mechanisms,* Blackwell, Oxford, 2000. Actualmente, está interesado en problemas relacionados con la teoría de la acción y las atribuciones de intencionalidad.

DIRECCIÓN: Department de Filologia i Filosofia. Facultat de Lletres. Universitat de Girona. Pl. Ferrater Mora, 1. 17071 Girona. E-mail: josepll.prades@udg.es.

RECENSIONES / *BOOK REVIEWS*

SÁNCHEZ-MAZAS, Miguel (2002-2003): *Obras Escogidas*: Vol. I: ed. Javier de Lorenzo, Vol. II: eds. Javier de Lorenzo y Gabriel Painceyra. Donostia-San Sebastián: Servicio Editorial de la Universidad del País Vasco/Euskal Herriko Unibertsitatea.

Al fin contamos, en tan sólo dos volúmenes, con una panorámica precisa de la obra de uno de los principales y más originales filósofos españoles, el lógico Miguel Sánchez-Mazas, desaparecido en 1995 y fundador de *Theoria*. Es necesario, por tanto, agradecer a los editores y a todos los que colaboraron para reunir los textos de estos dos volúmenes. Dada la dispersión original y la diversidad de estilos de cada uno, recogerlos y unificarlos debió ser una tarea titánica. Tarea que dificultaba aún más la estructura compleja de miles de notas y cuadros que utilizaba el autor. Esta edición de gran parte de su obra, aunque mejorable en múltiples aspectos, debe ser bienvenida por la comunidad lógica, y también por todos aquellos que trabajan en cuestiones jurídicas y filosóficas en general.

El primer volumen recoge prácticamente todos los artículos que publicó el filósofo español alrededor del proyecto de la característica universal, que constituyó el eje de toda su obra. El sueño leibniziano de encontrar una transcripción aritmética para las proposiciones y así poder calcular el valor de verdad de los enunciados, que por siglos había sido mal comprendido e incluso despreciado, encuentra en Sánchez-Mazas al receptor más agudo y entusiasta.

Javier de Lorenzo, el editor de este primer volumen, lo divide en dos partes. La primera, titulada "Concepto y número", recoge los artículos de Sánchez-Mazas dedicados el estudio de la obra lógica de Leibniz, cuya principal finalidad sería el estudio de sus desarrollos técnicos, evaluando sus alcances y defectos, y buscando siempre la manera de desarrollarlos. Esta primera parte del libro se cierra con los artículos en los que Sánchez-Mazas habría encontrado un modelo *leibniziano* de la lógica, que corrige los defectos de las propuestas del filósofo alemán. La segunda parte, titulada "La característica numérica universal", recoge los artículos donde Sánchez-Mazas reelabora ese modelo para hacerlo satisfacer una serie de nuevas necesidades hasta llegar a uno que constituiría, para De Lorenzo, el aporte lógico propio del filósofo español. Esta división de su obra resulta particularmente delicada, dado que Miguel Sánchez-Mazas se preocupó siempre por subrayar la continuidad de su trabajo, no sólo internamente, sino como continuación de la obra de Leibniz. En la mayoría de sus ensayos, artículos y ponencias, buscó tanto hacer ver el recorrido que llevaba a lo que estaba presentando, como las tareas que quedaban pendientes por realizar. De Lorenzo parece querer obligar a Sánchez-Mazas a reconocer que el resultado final es suyo propio y no una mera actualización, desarrollo y perfeccionamiento de los sistemas de Leibniz, como el autor de los ensayos siempre defendió.

La partición que hace Javier de Lorenzo de la obra de Sánchez-Mazas resalta dos facetas de su trabajo: por un lado, el estudio de la lógica de Leibniz y, por otro, el desarrollo propio de un sistema de lógica intensional. Este espíritu de doble intención —de estudio filosófico-histórico y a la vez lógico-propositivo— es quizás lo que distingue por encima de todo el trabajo de Sánchez-Mazas. Aunque hay artículos que se concentran en una u otra faceta, en la mayoría de ellos ambas intenciones se entrelazan de manera muy fuerte. Es siempre, simultáneamente, historiador de la lógica y lógico-matemático. A nuestro parecer, la obra de Sánchez-Mazas puede dividirse mejor de la siguiente manera. La Primera Etapa está dedicada al *estudio cuidadoso de los sistemas lógicos leibnizianos y al desarrollo de las primeras intuiciones sobre la analogía entre concepto y número*, y abarca su trabajo desde 1950 hasta 1963 aproximadamente. Luego se sucede, a nuestro juicio, un Momento Intermedio, en el que Sánchez-Mazas se dedica a desarrollar un modelo aritmético "no-leibniziano" para la lógica, expresado en los artículos publicados entre 1968 y 1973. Finalmente, en la Segunda Etapa, trabaja expresamente en el *modelo aritmético de la lógica* de estirpe

leibniziana, que hace público por primera vez en 1977 y lo desarrolla hasta el final de su obra. Al igual que el editor del libro, dividimos el trabajo de Sánchez-Mazas en dos etapas, pero producimos el corte antes. La primera etapa corresponde más bien a la del análisis y las propuestas, y la segunda a la de la realización concreta y la presentación de resultados. Es sin duda sorprendente la coherencia de la totalidad de la obra, y la firmeza que le permite, a través de toda una vida, ir realizando lo proyectado en un comienzo.

Comencemos con la Primera Etapa. Es impresionante leer ese primer artículo: "Sobre un pasaje de Aristóteles y el cálculo lógico de Leibniz" (1951), y ver la manera cómo se encuentran allí plegados todos los elementos de su obra posterior. Luego de aprovechar un fragmento de Aristóteles para presentar la clave de su sistema, *la analogía concepto-número*, en una nota al pie al final del artículo señala lo que constituirá el proyecto de toda su vida:

> En un trabajo próximo pretendemos exponer un sistema lógicomatemático completo, basado en la analogía leibniziana de la descomposición del número y la del concepto. En él expresaremos aritméticamente la compatibilidad, la incompatibilidad y todas las relaciones extensivas y comprehensivas, así como la noción de un individuo, la definición y el silogismo. [p. 31]

Si bien Sánchez-Mazas siempre agradecerá a Couturat, editor de los manuscritos lógicos de Leibniz, el haber hecho posible acceder a ellos, nunca le perdonará la presentación que les dio en su libro *La logique de Leibniz* (1901). En su completísima lectura de los mismos, Couturat decretará que todos los ensayos de Leibniz técnico-formales fueron esencialmente incorrectos. Para él, Leibniz no pudo desarrollar un sistema completo por el hecho de partir de una perspectiva intensional de la lógica, y no extensional, como había sucedido con la lógica que estaba surgiendo entonces, a comienzos del siglo XX. Ese prejuicio de Couturat contra la lógica intensional es el que Sánchez-Mazas reconoce como un error fundamental, y gran parte de su trabajo de allí en adelante se puede describir como un intento de refutarlo por una doble ruta simultánea: haciendo lucir lo que hay de esencialmente correcto en los ensayos de Leibniz, y desarrollando hasta la completud una lógica desde el punto de vista de la comprehensión.

Otro elemento crucial que aparece desde el primer artículo es la reflexión sobre el dominio adecuado de esquematización de la lógica. Tanto Sánchez-Mazas como Couturat destacan el hecho de que Leibniz hubiese ya probado tanto representaciones aritméticas (los números característicos) como geométricas (al estilo de los diagramas de Venn). En este primer momento Sánchez-Mazas está tentado a defender que la perspectiva intensional es más propia de ser representada aritméticamente, y la extensional de manera geométrica. Si bien en un texto posterior [p. 183] se alegrará de poder mostrar que la representación geométrica de Leibniz de la lógica intensional también era plausible, Sánchez-Mazas preferirá siempre la esquematización aritmética.

En esta primera etapa de su carrera, Sánchez-Mazas se concentrará en la lectura de los textos de Leibniz, tanto filosóficos como formales. Si bien estará al tanto de los sucesos más importantes de la lógica matemática del momento, su atención se fijará sobre todo en el trabajo de los historiadores de la lógica. Un punto de choque de todos ellos está en la lectura moderna que se hace de la silogística. Ésta sigue siendo, a principios del siglo XX, el paradigma de sistema lógico, y tanto historiadores como los propios lógicos se miden siempre con esa vara. En "La teoría del silogismo desarrollada en forma de lógica" (1954), Sánchez-Mazas hace una propuesta muy original de interpretación algebraica de la misma, que tiene además la gracia de distanciarse de la perspectiva estrictamente leibniziana de sus demás ensayos.

La culminación de esta primera etapa es el libro *Fundamentos matemáticos de la lógica formal* [pp. 93-155] que publica en Venezuela en 1963 y constituye la obra que ha debido situarlo en el panorama internacional. El mismo título oculta sus principales logros, ya que lo más importante

que allí presenta no es su propuesta de una lógica comprehensiva, sino la lectura que hace de los textos de Leibniz. Después del libro de Couturat, es quizás el primer análisis detallado que se realiza en el siglo XX de los manuscritos lógicos de 1679, donde además se refutan de manera impecable las críticas injustas que había hecho el primer editor de los mismos. No sólo eso, sino que Sánchez-Mazas organiza y estructura los ensayos de Leibniz de tal manera que hace posible por primera vez intentar seguir el hilo del pensamiento del filósofo alemán en esta materia. Su propia propuesta técnica se acoge a un ensayo leibniziano (asignar un par de números a cada concepto) que nunca convenció al filósofo español y que por lo tanto intentará pronto superar.

Luego viene un momento intermedio (hasta 1972), simultáneo con su estudio de la lógica de las normas en Suiza, en el que Sánchez-Mazas hace un ensayo de característica numérica para la lógica con un esquema de asignación muy distinto del propuesto por Leibniz. Esto merecería un estudio aparte, conjuntamente con sus ensayos de 1954, en razón de su diferencia al esquema general que utilizará el resto de su obra.

Sin embargo, sólo a finales de los años setenta, Sánchez-Mazas encontrará la manera de corregir la propuesta de Leibniz de modo que con una sencilla modificación de la misma se pueda construir un modelo adecuado para la silogística y la lógica de conceptos. La comunicación de 1977, "Un modèle mathématique de la logique peut-il se fonder sur l'intension?" [pp. 181-202], será la primera presentación pública de su descubrimiento. En los años siguientes, Sánchez-Mazas mostrará sus resultados en los congresos internacionales más importantes sobre lógica y sobre la obra de Leibniz. En esta Segunda Etapa, su propuesta partirá de resolver los problemas del proyecto inicial de Leibniz de asignar a cada término un número, basándose en los principios establecidos por el filósofo alemán en su álgebra de conceptos (que le da la clave de la importancia de asignar un número al término contradictorio o total: *non-Ens*). El resultado que encuentra es un álgebra de Boole aritmética cuyos elementos básicos son los números primos y la contenencia intensional está representada por la divisibilidad (que eran las dos bases de la propuesta leibniziana). Muy pronto se da cuenta también de que hay un análogo aritmético (la composición binaria de los números) que facilita las operaciones y a partir de allí propondrá varias álgebras distintas a la de los números primos.

Un elemento que siempre había quedado pendiente para cumplir el plan inicial de la obra era encontrar la manera de representar los *individuos* en este contexto [p. 245]. Como bien señala De Lorenzo en el prefacio, en este caso las "indicaciones" de Leibniz al respecto son particularmente generales, y por encima de todo, no se encuentra en sus ensayos técnicos ninguna aproximación en este sentido. Sánchez-Mazas las encuentra en sus textos filosóficos. Allí Leibniz se inspira también en las matemáticas para su reflexión lógica y metafísica, y comprende el individuo en analogía con los números irracionales, como límites que sólo son alcanzables a partir de los racionales en un número infinito de pasos. La solución que da Sánchez-Mazas en los artículos de finales de los años ochenta es de una suprema elegancia y belleza en la medida en que convierte lo que parecía una analogía vaga en un isomorfismo en sentido estricto. Y para mayor satisfacción, Sánchez-Mazas consigue hacerlo sin una mayor reforma de su modelo. La estructura es exactamente la misma, lo único que se cambia son los números que corresponden a los términos simples: se sigue tomando como base el modelo binario pero ahora, en lugar de asignársele a cada uno una potencia de 2, se les asigna las potencias inversas de 2.

Durante la década de 1990 Sánchez-Mazas se dedica a recoger todo lo que ha cosechado con los años. Sus escritos finales son verdaderos hipertextos, ya que el hilo central viene siempre acompañado de innumerables notas al pie y una serie creciente de cuadros al final en los que expone sus resultados técnicos. Es justamente el Sánchez-Mazas filósofo el que se revela

en estos últimos artículos dejando ver todas las conexiones que encuentra, los diversos puentes que puede plantear entre los diversos campos del saber. El artículo final, "El poliedro imposible: ciencia, filosofía, tecnología y utopía" (1998), es una especie de coda cuyo tema es distinto al de todos los demás, pero que nos deja ver de manera más clara una filosofía que, inspirada en una devoción por las matemáticas, lleva a cabo una profunda reflexión sobre los temas fundamentales y nos muestra la manera de aplicar su original estilo de investigación a los más diversos campos.

A partir de la década de 1960 Sánchez-Mazas también se dedicó intensamente al estudio de la estructura formal de las normas jurídicas. Su propósito era aplicar sus álgebras numéricas en el campo del Derecho. Asimismo, el conocimiento de la obra de Von Wright sobre lógica deóntica constituyó un punto de apoyo importante para acometer esta empresa. Ésta era una continuación natural de su trabajo anterior inspirado en la obra de Leibniz, ya que éste, habiendo recibido formación en el campo del Derecho, había intentado también extender su *mathesis universalis* en el terreno jurídico. Toda esta faceta de su indagación la hallamos en el Volumen II de la *Obras Escogidas*.

En 1971 Sánchez-Mazas presenta su tesis doctoral "Cálculo de las Normas" en la Universidad de Neuchâtel (Suiza), escrita en francés bajo la dirección de Jean-Blaise Grize, y publicada dos años después en español. Para su confección contó con los consejos de Georges Kalinowski, autor de un importante trabajo sobre "Teoría de las proposiciones normativas" (1952) que influyó grandemente en el primer cálculo normativo de Sánchez-Mazas. Casi al mismo tiempo aparece la importante obra sobre lógica normativa de Carlos Alchourrón y Eugenio Bulygin. El libro de éstos, *Normative Systems*, apareció primero en inglés (1971) y tres años más tarde en castellano. Esta obra marcó fuertemente el trabajo posterior de Sánchez-Mazas, y lo llevó a reformular varias veces sus cálculos normativos.

En *Cálculo de las Normas* (su tesis, publicada en Ariel en 1973 y presente en el Volumen II), Sánchez-Mazas desarrolla tres cálculos en realidad: un Cálculo Normativo Puro, un Cálculo Fáctico Puro, y un Cálculo Deóntico General.

El Cálculo Normativo Puro se aplica a constantes y variables normativas y, por ende, a cualquier norma expresable proposicionalmente. Consta de un solo operador lógico (el de negación) junto a cuatro operadores deónticos específicos: de implicación normativa (o jurídica) y los de incompatibilidad, de compatibilidad y de independencia normativas (o jurídicas).

El Cálculo Fáctico Puro, por su parte, se aplica a constantes y variables fácticas que refieren a *acciones* expresables proposicionalmente. También consta de un solo operador lógico (el de negación) e incorpora asimismo cuatro operadores propiamente fácticos: de implicación fáctica entre acciones, y de incompatibilidad, de compatibilidad y de independencia fáctica entre acciones.

El Cálculo Deóntico General, finalmente, es el núcleo del Cálculo de las Normas de Sánchez-Mazas correspondiente a esta época. Este Cálculo Deóntico General integra los dos cálculos anteriores (el Normativo Puro y el Fáctico Puro) pero agrega además quince operadores deónticos. De estos operadores deónticos los principales son los de obligación, prohibición, permisión y dispensa.

El Cálculo de las Normas se completa con la expresión aritmética de las fórmulas normativas, fácticas y deónticas, para formar un álgebra numérica que expresa la lógica normativa y que constituye así una nueva aplicación de las álgebras numéricas que había desarrollado anteriormente para la lógica intensional y reafirma la orientación leibniziana del trabajo de Sánchez-Mazas, ahora en el campo de las normas y del derecho.

Este primer cálculo normativo de Sánchez-Mazas está basado aún en la lógica deóntica de Von Wright, aunque la amplía y desarrolla en mucho mayor detalle, aumentando así el campo de aplicación. Pero la novedad principal estriba en el desarrollo de un álgebra aritmética isomorfa con el álgebra subyacente al Cálculo lógico Deóntico General.

El desarrollo de esta álgebra es muy importante ya que permite su instrumentación, en términos sencillos y económicos, en programas informáticos para procesamiento de la legislación. Los principales resultados en este campo y que corresponden a esta primera época del trabajo de orientación normativa y jurídica de Sánchez-Mazas, se expresan en una serie de programas. Entre éstos se destacan los denominados "Calculus Ratiocinator" I y II, y "Calculus Consequentiarum" I y II.

Sin embargo, la atención intelectual de Sánchez-Mazas se orienta, años más adelante, hacia el sistema normativo desarrollado por Carlos Alchourrón y Eugenio Bulygin en *Normative Systems* (1971). Esta obra está basada en la noción de "sistema deductivo" de Tarski como conjunto de enunciados que contiene todas sus consecuencias lógicas. Así, un *sistema normativo* para Alchourrón y Bulygin es un conjunto de enunciados que contiene también enunciados normativos e incluye todas sus consecuencias lógicas y normativas (deónticas). Esta noción de sistema normativo se puede ampliar a la de "sistema jurídico" entendido como sistema normativo que contiene normas que prescriben sanciones.

La importancia de la obra de Alchourrón y Bulygin estriba en que es posible dar mejores caracterizaciones y definiciones más claras para las nociones lógicas y jurídicas de completud, coherencia, laguna, independencia, etc., aplicables a diferentes sistemas normativos.

Sánchez-Mazas valoró la obra deóntica de Alchourrón y Bulygin al punto de adoptarla como base para el desarrollo de nuevas álgebras numéricas aplicables al campo del Derecho. La nueva orientación teórica de Sánchez-Mazas, desde los comienzos de la década de 1980, consiste entonces en interpretar los sistemas normativos de Alchourrón y Bulygin como redes deónticas con el propósito de construir, por su parte, redes numéricas isomorfas a aquellas. Basándose en las nociones normativas y jurídicas expresadas lógicamente en los términos de Alchourrón y Bulygin construye un álgebra aritmética compuesta de todos los números naturales en un intervalo finito, construidos como sumas de potencias de 2. Estos números junto a las operaciones aritméticas de supremo binario, ínfimo binario y complemento binario son suficientes para expresar todos los enunciados y relaciones presentes en la red deóntica del sistema normativo. El cálculo así construido constituye una red numérica con estructura de "retículo" y de "álgebra de Boole". Finalmente, este cálculo tiene su implementación práctica en el programa informático llamado "Ars Judicandi" presentado inicialmente en 1985.

Un tercer modelo, al que Sánchez-Mazas prestó mucha atención también, es el *lenguaje jurídico normalizado* de Layman Allen, relacionado con el movimiento surgido en torno a la revista norteamericana M.U.L.L. (*Modern Uses of Logic in Law*). El lenguaje normalizado de Allen también tiene como objetivo la representación unívoca de un conjunto de normas y de sus relaciones a través de reglas de formación estrictas. La novedad aquí consiste en el uso de esquemas, líneas y flechas como forma de expresión de la estructura jurídica subyacente.

En 1984, en la conferencia recogida en el artículo "Álgebra del derecho y procesamiento de la legislación", Sánchez-Mazas declara que los tres modelos (el de los sistemas normativos de Alchourrón y Bulygin, el lenguaje normalizado de Layman Allen, y su álgebra numérica de los sistemas normativos) son: "compatibles, prácticamente equivalentes y complementarios, por poderse utilizar alternativamente según los casos" [p. 59 de *Obras Escogidas*]. Los tres constituyen, entonces, lenguajes jurídicos artificiales desarrollados alternativamente con el propósito de poner de manifiesto la estructura profunda del lenguaje jurídico.

Hemos aludido ya a los programas informáticos de Sánchez-Mazas. Se trata de programas que implementan sus nociones de "red deóntica" y de "red numérica" en pequeños ordenadores de bolsillo como el Hewlett-Packard HP-97, aunque luego fueron aplicados también en ordenadores de mayor tamaño. Desafortunadamente no podemos detenernos en su descripción. Pero la conclusión general es que contrasta la simplicidad relativa de estos programas con respecto a su potencia y robustez.

Su estrategia general, en este campo, fue la de tomar sistemas normativos como base para la construcción de redes deónticas. Luego procede a la construcción de redes numéricas isomorfas a las redes deónticas, y son precisamente estas redes numéricas las que permiten el procesamiento informático de la información de manera económica y eficaz.

Las redes numéricas están compuestas por números característicos que expresan proposiciones normativas o jurídicas en forma de clases de equivalencia. El primer programa, "Calculus Ratiocinator", permite obtener precisamente todos los números característicos de una red dada. Esto es, permite conocer los números característicos de conjunciones de condiciones y acciones, como así también de las combinaciones booleanas de éstas que contienen negación y disyunción.

El programa "Calculus Consequentiarum" corresponde a la segunda etapa deductiva de un cálculo normativo y requiere, por tanto, de los números característicos obtenidos con el programa anterior. Permite la obtención y evaluación de todas las consecuencias normativas y jurídicas de una red deóntica dada, es decir, todas las consecuencias que contienen operadores deónticos (obligación, permisión, prohibición, etc.). Las características particulares de este programa revelan la importancia teórica y computacional de la informática de Sánchez-Mazas, ya que implementa de manera práctica su programa, de impronta leibniziana, de desarrollar una Característica Universal que permita el cálculo exacto del razonamiento normativo y jurídico. Por ello, estos programas no constituyen meras bases de datos judiciales sino que avanzan mucho más lejos en el terreno del razonamiento jurídico.

Finalmente, el programa "Ars Judicandi" lleva mucho más allá estas ideas. Implementa los mismos propósitos que los dos programas anteriores pero permite además evaluar redes y subredes deónticas con respecto a las nociones de coherencia, completud, laguna, etc., y permite también la integración de diferentes sub-redes en una red más general. Como ya se adelantó, este programa está basado en general en el marco teórico de los "sistemas normativos" de Alchourrón y Bulygin, y aporta de manera novedosa la implementación computacional de casi todas las nociones deónticas presentes en la concepción lógica de los autores argentinos. Constituye, por tanto, un programa potente y robusto desde el punto de vista de la informática jurídica, y que requiere además escasos recursos computacionales gracias a la simplicidad relativa de las álgebras numéricas desarrolladas por Sánchez-Mazas.

En conclusión, el segundo volumen de las *Obras Escogidas* muestra muy claramente las líneas generales de investigación del trabajo de Sánchez-Mazas en el campo normativo y jurídico. Pone de relieve la originalidad y la constante inquietud teórica de Sánchez-Mazas en este campo, lo cual le convierte, a nuestro parecer, en uno de los "clásicos" dentro de esta disciplina.

También presenta las aplicaciones prácticas de las lógicas normativas, es decir, la informática jurídica de Sánchez-Mazas, y es este aspecto el que interesa destacar muy especialmente. En general, lo que abunda en el campo informático es el desarrollo de bases de datos, cada vez más grandes y potentes, hasta llegar a los llamados "sistemas expertos", presentes ya en este campo de la informática jurídica desde hace alrededor de 30 años atrás.

Pero la informática jurídica de Sánchez-Mazas intentó superar este nivel y lo logró en gran medida. Ya que sus programas implementan un aspecto más elevado que el correspondiente a las bases de datos, y este aspecto es el del *razonamiento* jurídico propiamente dicho. Sus programas permiten extraer las consecuencias normativas y jurídicas a partir de un conjunto de enunciados y, también, evaluar sistemas normativos enteros. Esto se corresponde con un nivel superior del procesamiento de información jurídica, que Sánchez-Mazas llamaba "metadocumentaria" o "decisional", en contraste con la informática jurídica meramente "documentaria" en el que se sitúan las máquinas que sólo implementan bases de datos.

Este nivel de la informática "decisional" constituye, todavía hoy, una línea de investigación abierta, en el que queda aún trabajo por realizar. En particular, el aspecto semántico de los textos jurídicos, con sus ambigüedades e imprecisiones (muchas veces presentes de manera intencional), constituye un rasgo de difícil implementación al nivel computacional de procesamiento de información. Si las nuevas tendencias en Inteligencia Artificial permiten finalmente la manipulación de los aspectos semánticos de los textos jurídicos, podrán complementarse con los niveles informáticos ya desarrollados. Y si esto se logra, lo estará, seguramente, dentro del mismo espíritu inquieto y progresista de Miguel Sánchez-Mazas.

Alejandro MARTÍN MALDONADO
Dpto. de Matemáticas
Universidad de los Andes (Colombia)
E-mail: alemartin@gmail.com

Gabriel PAINCEYRA
Dpto. de Lógica y Filosofía de la Ciencia
Universidad del País Vasco/Euskal
Herriko Unibertsitatea
E-mail: sfbpapag@ehu.es

GUALA, Francesco (2005): *The Methodology of Experimental Economics*. New York: Cambridge University Press.

Una de las objeciones más escuchadas contra la microeconomía es la de que cuando vamos de compras no nos comportamos como maximizadores racionales de utilidad. En otras palabras, uno de los supuestos centrales en el análisis económico es empíricamente erróneo. Para defender, pese a ello, su valor epistemológico se propusieron distintas réplicas contra esta objeción, apelando a posiciones más o menos instrumentalistas: la teoría no describe de modo adecuado la toma de decisiones del agente individual, pero sus predicciones son acertadas. Hubo quien sostuvo esto de los individuos (v.gr., Friedman respecto a la maximización de la utilidad esperada), pero los economistas se alinearon mayoritariamente con Marshall: incluso si los individuos se confunden, sus "errores" se cancelan al agregarse, y en conjunto tienden a comportarse como establece la teoría de la demanda.

No obstante, el escepticismo respecto al análisis económico del comportamiento individual era, en general, intuitivo, tanto entre objetores como entre sus propios defensores. Hasta después de la Segunda Guerra Mundial el control experimental de la conducta económica no comenzó a desarrollarse sistemáticamente y su conversión en una subdisciplina académicamente respetable se obtuvo sólo con la concesión del Nobel de la especialidad a Kahneman y Smith en 2002. Es imposible subestimar, por tanto, la importancia metodológica de la economía experimental, pues viene a ordenar definitivamente algunas de nuestras intuiciones centrales sobre el valor empírico de la teoría económica. O eso creíamos, pues a menudo los partidarios y detractores de esta coinciden en señalar la miseria metodológica de la economía experimental: es imposible reproducir en un experimento las condiciones en las que los agentes toman realmente sus decisiones y, por tanto, sus resultados no sirven para convalidar nuestras intuiciones a favor o en contra de nuestros modelos teóricos.

En *The Methodology of Experimental Economics*, Francesco Guala examina y defiende el estatuto metodológico la economía experimental de un modo que quiere resultar asequible a economistas y filósofos. Para aquellos, buena parte de sus 11 capítulos constituyen una introducción «situada» a muchos debates actuales sobre metodología científica. Estos se verán sorprendidos sobre su rendimiento al aplicarse a un caso tan singular como es el de la experimentación en ciencias sociales. Así, junto a un buen número de cuestiones clásicas (evidencia, explicación nomológico-deductiva, causalidad, ...), Guala introduce también algunas tesis propias del neoexperimentalismo, como la distinción entre datos y fenómenos, la pluralidad de la ciencia o las mediaciones entre teoría y experiencia. Aunque no se propone un desarrollo completo de cada una de ellas, la claridad de la presentación y el interés de los ejemplos con que se ilustran lo convierte en un texto muy adecuado para su uso en cursos de metodología económica.

La estrategia argumental de Guala en su vindicación de la economía experimental es dúplice. Por una parte, intenta establecer cuál es el alcance del conocimiento que nos proporcionan los experimentos. Su tesis aquí es que nuestro control causal de la conducta económica es efectivo (a partir de los datos aparecen regularidades fenoménicas estables), pero restringido por unas condiciones experimentales concretas. De ahí que debamos constreñir nuestras inferencias sobre los resultados experimentales conforme a tales condiciones, explicitando el conocimiento de fondo subyacente (*background knowledge*) mediante sucesivas inducciones eliminativas donde se establezca objetivamente su valor empírico. Guala establece su tesis contra un buen número de alternativas filosóficas. Desde luego, el falsacionismo (durante años predominante en la metodología económica) y, en general, contra las posiciones exageradamente deductivistas (inevitablemente abocadas a los dilemas de Duhem-Quine), pero también contra la interpretación bayesiana del conocimiento de fondo (y se diría que también del propio diseño experimental). De este modo, Guala se opone a quienes apelan-

do a posiciones de principio cuestionan el valor de los experimentos económicos por su carácter excesivamente particular: no cabe alternativa mejor, y el metodólogo debe dar cuenta de ello.

La segunda parte del argumento de Guala (y también de su libro) se concentra en el dilema de la validez externa de los experimentos económicos: ¿pueden sus conclusiones extrapolarse a los auténticos mercados? Nuestro autor despliega aquí su tesis sobre los experimentos como *mediadores* entre la teoría y el mundo. Nuestras inferencias serían antes *analógicas* desde experimento al mundo que directas (de la teoría a su aplicación): sólo en la medida en que el experimento reproduzca satisfactoriamente aquellas circunstancias reales por las que nos interesamos (a veces tan sólo por motivos prácticos, como en las subastas de telecomunicaciones) podremos considerar justificadas nuestras analogías, y no sólo por su congruencia con nuestro modelo teórico. De nuevo, se trata de un razonamiento de lo particular (nuestras circunstancias experimentales) a lo particular (el caso analizado). De ahí su condición de *mediadores*: no son sin más el objeto de análisis al que se aplica la teoría, sino que lo representan de un modo no exclusivamente teórico, que tiene un interés en sí mismo. Frente al particularismo radical (*radical localism*) defendido por Bruno Latour, Guala afirma así un particularismo «intermedio»: intentar imponer normas metodológicas universales es tan nocivo como negar absolutamente su existencia.

Nadie podrá negar el interés y la solvencia de semejante argumento y, en esa medida, la economía experimental quedará metodológicamente vindicada contra sus críticos. No obstante, cabe preguntarse también qué perspectivas nos abre esta posición sobre el conjunto de la metodología económica. Por ejemplo, queda abierta la cuestión de qué inferencias podemos establecer de los experimentos ya no al mundo, sino a la teoría. Uno de los casos más ampliamente discutidos por Guala en la primera parte de su libro es el de la *preference reversal*, claramente contradictorio con uno de los ingredientes centrales en la teoría de la elección racional: la relación de preferencia es asimétrica. Se trata de un resultado experimental bien establecido, frente al cual proliferan las respuestas que, en el mejor de los casos, o bien minimizan la importancia del axioma en cuestión, o bien ofrecen modelos de decisión alternativos. Ambas opciones son válidas para Guala. El dilema que cabe aquí plantearse (y sobre el que apenas encontramos mención en este trabajo) es *qué prueban los experimentos respecto a la teoría*. Por el momento, los fracasos experimentales no parecen dar suficientes motivos para la adopción de enfoques alternativos, sin que sepamos muy bien cómo afecta esto al estatuto científico de la teoría económica. Dado el gusto por la generalidad matemática de sus partidarios, se diría que la particularidad de los resultados experimentales, tan bien defendida por Guala, constituye un buen motivo para no prestarles demasiada atención. El economista teórico podría considerar su actividad como un puro ejercicio de matemática aplicada del que podría aprovecharse independientemente el experimentador para articular modelos contrastables, sin que su fracaso empírico les restase justificación. Todo depende, desde luego, cómo se conciba la unidad de la economía como ciencia, y en una perspectiva neoexperimentalista como la de Guala no parece que tengamos demasiados motivos para exigir tal unidad —que era, justamente, la que creaba dificultades empíricas en la concepción positivista clásica de los modelos económicos. Por tanto, la teoría económica gozaría de una envidiable salid a este respecto y las objeciones de sus críticos revelarían únicamente incomprensión respecto a cómo funciona su disunidad. ¿Es esto aceptable? Júzguelo el lector, a modo de ejemplo del interés que los debates que previsiblemente nos traerá el desarrollo de esta posición.

David TEIRA SERRANO
Dpto. de Lógica, Historia y Filosofía de la Ciencia
Universidad Nacional de Educación a Distancia
E-mail: dteira@fsof.uned.es

CHANG, Hasok (2004): *Inventing Temperature. Measurement and Scientific Progress*. New York: Oxford University Press.

La invención de la temperatura es un libro que necesariamente interesará a quienes se preguntan hoy cómo articular de nuevo tres disciplinas: Historia de la ciencia, Filosofía de la ciencia y Epistemología. Desde luego, son muchos los que creen que es mejor cultivarlas por separado (por dar un ejemplo reciente, Jesús Zamora en su *Cuestión de Protocolo* [Tecnos, 2005]), sospechando acaso que suelen perder en la "mezcla". A estos, el ensayo de Hasok Chang se les ofrecerá como reto, pues lo que nos propone es, precisamente, revisar algunos conceptos fundamentales en Filosofía de la ciencia (*e.g.*, observabilidad, progreso, operacionalismo) a partir, por un lado, del análisis de algunos casos señalados en la Historia de la termometría y, por otro, de una reconsideración de algunos de sus supuestos epistemológicos más característicos (la disyuntiva entre fundacionismo y coherentismo).

Chang parte, desde luego, de una posición filosófica donde la conjunción entre estas tres disciplinas no resulta extraña (y así se refleja en su propio itinerario académico, desde Stanford a Londres, según puede seguirse en los agradecimientos). Chang argumenta en las inmediaciones del neoexperimentalismo de Hacking y Cartwright, desde donde aborda la constitución de la temperatura como *dato* en *leyes fenomenológicas* o *regularidades causales de bajo nivel*. Esto explica, en buena parte, la estructura del libro: así, Chang nos presenta un buen número de tentativas experimentales para fijar un punto fijo en el termómetro a partir ebullición de diversas sustancias (cap. 1); para establecer la compatibilidad entre distintas escalas de medida (cap. 2); y para ampliarlas a temperaturas extremas, como pretendieron los primeros pirómetras (cap. 3). El cap. 4 sigue el camino inverso y examina mediante qué tipo de experimentos se pretendió conferir significado empírico al concepto de *temperatura absoluta*, en particular a partir de su elaboración por Kelvin. En estos cuatro capítulos se procede según una división entre una primera parte *narrativa* y una segunda *analítica*. En la primera se presenta de un modo erudito cada caso tal como lo expusieron inicialmente sus autores, en un ejercicio de *Historia interna* iluminado por la discusión filosófica que se desarrolla en la segunda parte. Es decir, quien sólo se interese por la Historia de la termometría echará seguramente en falta buen número de consideraciones (en particular, contextos, deudas, etc.), pero esto no implica que el análisis que se nos propone no resulte original: Chang rescata los trabajos de experimentalistas como De Luc, Regnault o Wedgwood precisamente porque su juicio metodológico le permite vindicarlos.

La tesis filosófica que construye sobre ellos es básicamente la siguiente: el caso de la termometría nos muestra cómo es posible alcanzar un consenso científico sobre la base de un refinamiento gradual de nuestras convenciones métricas, aun si no se disponga de un punto de partida incuestionable (puntos fijos), ni de un concepto teórico bien elaborado, y con independencia de que coexistan durante amplios periodos escalas difícilmente convertibles de precisión muy diversa. Chang reivindica aquí una posición epistemológica *coherentista* para interpretar como *progresivo* este proceso: es el apoyo mutuo que se van prestando las distintas mediciones —juzgado no por la verdad, sino por otras virtudes epistémicas (tales como la exactitud, fecundidad etc.)— lo que justifica su aceptación. Su argumento también puede leerse a la inversa: sería imposible dar cuenta racionalmente del progreso de la termometría si se exigiera de sus protagonistas una justificación al modo fundacionista, basada en datos empíricos autoevidentes y, por ello, incontestables para todos los implicados. La observabilidad es, para Chang, un *logro*, antes que un dato. Así, nos propone un modelo para analizar este progreso sobre la base de una rectificación operacionalista del convencionalismo (pp. 92-96), donde se muestra cómo puede avanzar iterativamente el proceso de medición observando ciertos principios metodológicos que garanticen su coherencia (p. 152).

Desde este punto de vista, uno de los aspectos más interesantes del libro es que explicita muchas tesis clásicas en filosofía de la ciencia que en el neoexperimentalismo suelen encontrarse de modo más bien oblicuo. Esto es particularmente cierto en lo que respecta a los dos últimos capítulos. El quinto articula todo lo expuesto en las partes analíticas de los anteriores y el sexto plantea cuál sea la posición de la filosofía de la ciencia "en el conjunto del saber". Chang defiende aquí es que se trata de la continuación de la ciencia "por otros medios": puesto que cualquier ciencia normal es dogmática respecto a su propio paradigma, a la filosofía le corresponde mostrar hoy en qué medida fue crítica su aceptación inicial respecto a las alternativas existentes. Lo peculiar de esta posición es que, para Chang, el conocimiento que se nos proporciona así sería también *científico* (pp. 240-47): en la medida en que la filosofía opera sobre argumentos científicos previos, su recuperación produciría conocimiento igualmente científico, bien por coincidir críticamente con lo que hoy aceptamos como dogma, bien por sugerir una alternativa eventualmente desarrollable. De hecho, del lado epistemológico, es probable muchos lectores encuentren poco desarrollada conceptualmente la posición de Chang. Su respuesta más probable es que todo desarrollo deberá venir de la mano de análisis de argumentos científicos, si es que ha de ser relevante para su posición.

Es mérito indiscutible de este libro poner sobre la mesa cuestiones tan políticamente incorrectas (para la convivencia departamental, por ejemplo) como puedan ser todas las anteriores. En qué medida resulten aceptables sus conclusiones lo determinará el debate ulterior. Para contribuir a éste, vaya al menos una objeción: vista la posición de la filosofía de la ciencia como *ciencia complementaria*, ¿qué nos queda de la vieja distinción entre disciplinas positivas y normativas?

La base del consenso público sobre la ciencia, en la vieja perspectiva empirista, radicaba en que el conocimiento que nos procuraba se justificaba, en última instancia, sobre nuestras sensaciones (instancia positiva), como nuestro conocimiento ordinario, y esta era una justificación que cabía compartir universalmente (instancia normativa). Chang no renuncia a esta base empírica (p. 86), pero entiende que la justificación descansa más bien sobre la coherencia entre argumentos científicos, que al filósofo le corresponde ahora rescatar por vía informal (y ya no axiomática). La cogencia de estos argumentos no puede ser superior a la que el propio científico podría obtener de ellos y, en ese sentido, nos proporcionarán una justificación equivalente para su aceptación pública. Es decir, lo que pierde la instancia positiva lo gana la normativa: la filosofía se convierte en ciencia complementaria, porque el tipo de argumentación específicamente científico se asimila al propio análisis filosófico.

El nudo de esta posición coherentista es que un sociólogo podría apropiarse perfectamente del análisis de Chang cambiando las coordenadas: la coherencia argumental, como muestra la tradición de Bloor, es perfectamente interpretable en términos de coincidencia de intereses particulares. El público acepta los resultados científicos porque los intereses subyacentes son una proyección de los suyos propios. Ahora bien, quienes no pertenezcan a la comunidad (social o científica) no tendrán por qué compartirlos. ¿Esta universalidad es deseable para el neoexperimentalista? Implícitamente se diría que el análisis de la práctica experimental que Chang nos propone pretende mostrarlo, pero el tipo de argumentos que nos proporciona (en particular, la teoría del significado que se asume) no parecen darnos demasiadas razones para esperarlo.

David TEIRA SERRANO
Dpto. de Lógica, Historia y Filosofía de la Ciencia
Universidad Nacional de Educación a Distancia
E-mail: dteira@fsof.uned.es

YEARLEY, Steven (2005): *Making Sense of Science. Understanding the Social Study of Science*. London, Thousand Oaks, New Delhi: Sage Publications.

El objetivo del libro *Making Sense of Science. Understanding the Social Study of Science* es doble: tematizar la materia oscura de la sociedad, es decir, el rol que juegan en la sociedad y en nuestras vidas sociales la evidencia científica, la experticia técnica, las leyes científicas así como los riesgos y los sistemas tecnológicos como agencias; y, reconocer la sociología de la ciencia como un componente principal de la teoría sociológica. Hasta el momento, según Steven Yearley, cuando los teóricos de la sociedad han pretendido incorporar cuestiones del ámbito de la ciencia y la tecnología han tomado como referencia principalmente los estudios sobre el posmodernismo o sobre la comunicación del riesgo y sus derivados, y no por el contrario los estudios sociales de la ciencia. Yearley propone la integración en un cuerpo único de análisis los estudios de ciencia y la teoría social: sólo el reconocimiento del valor analítico de los estudios de ciencia nos permitirá avanzar en la comprensión sociológica del significado de la ciencia y la experteza técnica.

A lo largo del libro toman relevancia dos ideas: la materia oscura o la masa ausente, y la objetividad de los análisis científicos. La primera intenta mostrar la importancia de las prácticas científicas y su interpretación para que el análisis sociológico dé cuenta de las dinámicas sociales. La segunda, por su parte, revisa la idea de que las teorías y los descubrimientos científicos resultan autónomos y objetivos, de modo que el reconocimiento de las cuestiones de ciencia y tecnología como materias oscuras de las ciencias sociales no suponga la claudicación ante las ciencias naturales. Yearley propone que las creencias sobre el estado del mundo natural no están determinadas por la información que podamos extraer del mismo mundo natural sino por las decisiones que adoptan los grupos sociales, y especialmente los científicos, y es por ello por lo que la comprensión de las materias oscuras de la sociedades tiene una dimensión sociológica.

En definitiva, la preocupación de la sociología no pueden ser únicamente las personas y las instituciones, pero tampoco resulta suficiente analizar los supuestos elementos accidentales de la ciencia (los beneficios económicos de los nuevos inventos y la aceptabilidad política que resulta de las nuevas ideas) según propone la postura demarcacionista a través de un criterio externista. De la misma manera, la relevancia sociológica de los estudios de ciencia gana posiciones con la crítica a la figura contemplativa, guiada por la curiosidad, el desinterés y el reconocimiento de sus colegas, que ha tenido el científico. Para ello, sin ánimo de exhaustividad, conviene enumerar algunas ideas que resaltan tanto en el libro *Making Sense of Science. Understanding the Social Study of Science* como en los estudios sociales de la ciencia, a saber: la crítica a la racionalidad científica; las discusiones en torno a los métodos científicos y las virtudes cognitivas que parecían facultar a la ciencia reflejar lo natural y adquirir por tanto un conocimiento isomórfico e independiente de toda influencia social; los cambios estructurales en los sistemas de ciencia; un nuevo contexto de las políticas científicas; y la emergencia de principios pragmáticos (riesgo e incertidumbre, actitudes públicas críticas, sociedad civil) como elementos constitutivos de las prácticas sociotécnicas.

Con el objetivo de reflexionar estas cuestiones, el libro se divide en dos bloques: en el primero se evalúan la excepcionalidad de la ciencia y las diferentes escuelas que componen el campo de los estudios de ciencia; en el segundo bloque el autor trata de explicar las dificultades que atraviesa la ciencia en su relación con las tomas de decisión en política, tribunales y sociedad. Para ello, en lo que al primer bloque se refiere, en las dos primeras partes y a lo largo de siete capítulos del libro Yearley se centra en analizar el estado de la cuestión del interés sociológico sobre los estudios de ciencia. Antes de introducirse en el estudio crítico de las diferentes

tradiciones en los estudios de ciencia, Yearley aborda por separado las principales preocupaciones intelectuales en los estudios de ciencia. Para ello, en el primer capítulo revisa las cuatro ideas (según la perspectiva empirista, metodológica, normativa y valorativa) que se han utilizado para destacar el carácter especial de la ciencia y posteriormente en el segundo capítulo presenta el programa fuerte y el programa empírico del relativismo. En estos dos primeros capítulos me permito identificar dos conclusiones: Yearley argumenta no sólo las debilidades que tienen las teorías de la excepcionalidad, sino que también reconoce las implicaciones que los presupuestos dominantes de la ciencia tienen con sus transposiciones normativas en la arena política; y por otra parte, expone una serie de consideraciones que condicionan las posibilidades de los estudios de ciencia, una vez que los fundamentos sociológicos reproducen a través de polos opuestos las dicotomías que caracterizan la perspectiva realista.

Con estas apreciaciones, en los cuatro capítulos de la segunda parte el autor recoge las diferentes escuelas de los estudios de ciencia. En primer lugar, la "Escuela de Edimburgo" (cap. 3) que, a través de la incorporación de los intereses sociales en el desarrollo del conocimiento científico, supone la primera corriente teórica en la sociología del conocimiento científico; la teoría de actor-red y sus contribuciones con los términos enrolamiento y traducción de intereses, la generalizada simetría en acción, y la incorporación del mundo natural (cap. 4); los estudios de género que, junto a los intentos teóricos de superar el feminismo empirista, tiene el reto de reformular su comprensión de cómo operan los valores científicos (cap. 5); la etnometodología y el análisis del discurso científico (cap. 6). En cada uno de estos cuatro capítulos Yearley distingue una parte teórica y otra más empírica, y en todas ellas se hace una constante alusión a los tres principios que dio a conocer el programa empírico del relativismo (la relevancia de la flexibilidad interpretativa de los resultados científicos; el análisis del proceso social que cierra los debates sobre los resultados; la investigación de la conexión entre los procesos y fuerzas sociales más allá de las comunidades científicas).

Es así como pretende mostrar los déficits de cada corriente: esta tarea, en mi opinión, a pesar de su extensión, no tiene especial relevancia para el libro, una vez que las críticas que se realizan son conocidas incluso en anteriores trabajos del mismo autor. Ahora bien, sí debemos reconocer que las críticas resumidas en el quinto y último capítulo de la segunda parte resultan importantes para los siguientes capítulos (cap. 7). El autor establece, especialmente respecto a la sociología del conocimiento científico, tres insuficiencias en sus prácticas: (i) si bien es de interés analítico la identificación de lo que resulta común a todas las ciencias (según la escuela: la traducción, los valores constitutivos, etc.), no menos importante es el rol social del conocimiento científico y sus cambios en los últimos años (nuevas formas de producción del conocimiento, crisis en la regulación de la ciencia, etc.); (ii) la no-reflexividad sobre sus propias prácticas, la limitación que les han supuesto su aparato teórico y la renuncia a asumir otras tendencias sociológicas, en gran medida debido a la exclusiva atención que les han merecido las cuestiones epistemológicas; (iii) la limitada preocupación por la sociología de la organización interna de la ciencia.

Estas tres cuestiones importan para la cuestión que preocupa a Yearley: el análisis de la masa ausente en sociología. Mas Yearley evita una conclusión negativa de los estudios de ciencia, sin por ello cerrarse a una escuela concreta, y es partidario de resaltar la diversidad de aportaciones que derivan de las diferentes escuelas: (i) el "finitismo": la gente o las comunidades son las que deciden lo que es el mundo; (ii) la dependencia entre la gente y las comunidades para determinar lo que se sabe, así como la actitud crítica para con las instituciones de control y decisión; (iii) la importancia de la confianza para establecer y mantener el conocimiento; (iv) la constatación de que la producción del conocimiento requiere el juicio, y que no resulta sufi-

ciente un método estándar. En opinión de Yearley, estas modestas conclusiones sobre los estudios de ciencia permitirán un análisis diferente de la ciencia en la sociedad y la reconsideración de la materia oscura de la sociología.

Precisamente, en el segundo bloque del libro, compuesto por los cinco capítulos de la tercera parte, Yearley intenta mostrar que los principales problemas identificados en la teoría sociológica respecto a cuestiones de ciencia y tecnología se comprenden de manera más cabal si nos dotamos del utillaje conceptual que viene proponiendo el campo de los estudios de ciencia. En los siguientes cinco capítulos se abordarán cinco cuestiones de especial relevancia: las controvertidas relaciones del público con la autoridad científica (cap. 8), el riesgo (cap. 9), la ciencia en las decisiones jurídicas (cap. 10), las políticas científicas y las dinámicas políticas del conocimiento (cap.11), y los problemas de las representaciones tradicionales de la ciencia y su credibilidad (cap. 12). El uso de las ideas que emergen en los estudios de ciencia servirán para acometer de una manera más "conflictiva" y "realista" cada una de las cinco controversias.

El reconocimiento que merece el libro no resulta tanto de las observaciones y críticas que se hacen a lo largo del trabajo, sino por los objetivos marcados y por el modo en que el autor trata de articular la problemática en un marco de consideraciones más amplias. Como también han indicado estudiosos del conocimiento científico como Jasanoff y Wynne, es probable que los estudios de ciencia tengan ante sí diferentes desafíos respecto a los que se han confrontado en los finales de los setenta, ante todo aquéllos que hacen referencia a la necesidad de pensar nuevas formas de relación entre ciencia, tecnología y sociedad, entre otras razones por la centralidad de la ciencia y la tecnología en nuestras sociedad y por su poder de reproducir organizaciones sociales, justas e injustas. Es la conclusión, en mi opinión, que falta en el libro de Yearley, ante todo por desconsiderar las aportaciones que se vienen haciendo en otras áreas de las ciencias sociales, y limitar su objetivo a complejizar la teoría sociológica a través de los estudios de ciencia. Un libro, por lo demás, necesario para la tarea académica y social que suponen las nuevas revoluciones sociotécnicas.

Andoni EIZAGIRRE
Unidad de Estudios de la Ciencia y la Tecnología
CSIC-UPV/EHU
E-mail: skxeieia@ehu.es

THEORIA

REVISTA DE TEORIA, HISTORIA Y FUNDAMENTOS DE LA CIENCIA

SUMARIO ANALITICO / *SUMMARY*

Vol. 21/3, Nº 57, pp. 241-360, Septiembre/*September* 2006
ISSN 0495-4548

ARTICULOS / *ARTICLES*

Es conocido que Kripke argumentó que la ilusión de contingencia en el caso de la conciencia no puede explicarse del modo en que se explica en el resto de casos familiares de enunciados necesarios a posteriori. En un artículo reciente, Pérez Otero (2002) argumenta que hay una explicación alternativa, en términos de mera aposterioridad. Argumento en contra de la corrección exegética y de la verdad de esta tesis.

Keywords: aposterioridad, consciencia, ilusión de contingencia, verdades necesarias aposteriori, bidi-mensionalismo, Saul Kripke.

Se examinan dos argumentos en favor de que la superveniencia fuerte y la global (debidos respectivamente a Kim y Sider) implican la reducción. Ambos se basan en sendas nociones de propiedad maximal que resultan problemáticas. Bajo una lógica infinitaria irrestricta con clases arbitrarias de estructuras infinitas la conclusión de Sider resulta falsa, mientras que el argumento de Kim no se puede formalizar. Concluimos que los dos argumentos no son válidos *en general.*

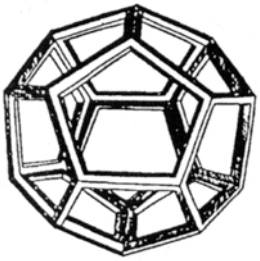
Descriptores: superveniencia, propiedades maximales, teoría de modelos, reducción, fisicismo no reductivo.

Evaluative realism is characterized, as rejecting the flexibility of values. The intuitive case against it is presented. Two further considerations against it are provided: one concerning the internalist connection between values and motivation, and the other concerning the intuitive causal inefficacy of evaluative properties.

Descriptores: evaluative realism, flexibility, metaethics, internalism, causal efficacy.

THEORIA

REVISTA DE TEORIA, HISTORIA Y FUNDAMENTOS DE LA CIENCIA



Ὁ Θεὸς ἀριθμετίζει

Vol. 21

FUNDADA EN 1952 - SEGUNDA EPOCA
FUNDADOR: Miguel SÁNCHEZ-MAZAS (†)

Revista asociada a la *Sociedad de Lógica, Metodología y Filosofía de la Ciencia en España* y a la *Sociedad Española de Filosofía Analítica*

Coeditan:

Centro de Análisis, Lógica e Informática Jurídica (CALIJ)

Servicio Editorial Universidad del País Vasco / Euskal Herriko Unibertsitatea

eman ta zabal zazu
universidad del país vasco
euskal herriko unibertsitatea

SAN SEBASTIAN / ISSN 0495 - 4548

No 55

SECCIÓN MONOGRÁFICA: *Scientific Representation*

Guest Editors: José DÍEZ and Roman FRIGG

ARTICULOS / *ARTICLES*

RECENSIONES / *BOOK REVIEWS*

No 56

No 57

ARTICULOS / *ARTICLES*

ESTADO DE LA CUESTIÓN / *STATE OF THE ART*

RECENSIONES / *BOOK REVIEWS*

INDICE DE AUTORES/ *AUTHOR INDEX*

THEORIA

REVISTA DE TEORIA, HISTORIA Y FUNDAMENTOS DE LA CIENCIA

Vol. 21

INDICE DE LIBROS RESEÑADOS/ *INDEX OF REVIEWED BOOKS*

T H E O R I A

REVISTA DE TEORIA, HISTORIA Y FUNDAMENTOS DE LA CIENCIA

Vol. 21