THEORIA. Revista de Teoría, Historia y Fundamentos de la Ciencia
ISSN: 0495-4548
theoria@ehu.es
Universidad del País Vasco/Euskal Herriko Unibertsitatea
España

PINTO, Sílvio
Un argumento trascendental para la inducción
THEORIA. Revista de Teoría, Historia y Fundamentos de la Ciencia, vol. 22, núm. 2, 2007, pp. 189-211
Universidad del País Vasco/Euskal Herriko Unibertsitatea
Donostia-San Sebastián, España

Un argumento trascendental para la inducción

(A transcendental argument for induction)

Sílvio PINTO

Manuscrito recibido: 25.09.2006
Versión final: 27.03.2007

BIBLID [0495-4548 (2007) 22: 59; pp. 189-211]

RESUMEN. Aquí lo que me interesa es, primero, distinguir dos problemas de justificación con respecto a la inferencia inductiva: por un lado, el de una justificación persuasiva de este tipo de inferencia y, por otro lado, el de una justificación explicativa de tal inferencia. En segundo lugar, intento mostrar que el argumento de Ramsey-de Finetti a favor de las reglas inductivas de la lógica bayesiana no es capaz de proporcionar una justificación persuasiva de estas reglas. Finalmente, propongo una justificación explicativa para las reglas de condicionalización bayesianas en términos de un argumento trascendental de inspiración kantiana y de estilo davidsoniano.

Descriptores: inducción, causalidad, bayesianismo, condicionalización, justitificación, explicación.

ABSTRACT. *Here, I am interested, firstly, in distinguishing two justification problems concerning inductive inference: on the one hand, the problem of a persuasive justification of induction and, on the other, the problem of an explicative justification of this sort of inference. Secondly, I intend to show that Ramsey-de Finetti's argument in favor of Bayesian inductive rules cannot provide a persuasive justification for these rules. Finally, I propose an explicative justification for Bayesian conditionalization rules in terms of a transcendental argument of Kantian inspiration and Davidsonian style.*

Keywords: *induction, causality, Bayesianism, conditionalization, justification, explanation.*

1. El Problema de Hume

David Hume fue el primero en formular el llamado problema de la justificación de la inducción. El problema para Hume era encontrar un argumento convincente en favor del uso de inferencias de tipo inductivo, cuyo empleo es tan común en los razonamientos cotidianos y, particularmente, en los de la ciencia. La búsqueda humeana de una justificación racional de la inducción está íntimamente ligada a su discusión sobre la posibilidad de justificación racional del principio de causalidad. La primera definición humeana de causa que aparece en la *Investigación sobre el Entendimiento Humano* la concibe como "an object, followed by another, and where all the objects, similar to the first, are followed by objects similar to the second" (Hume 1777, sección VII, parte II). De acuerdo con Hume, la identificación de semejanzas entre objetos o, como se diría más correctamente hoy, sucesos, y su conjunción constante requiere el hábito, un mecanismo psicológico innato que se desencadena a partir de la acumulación de la experiencia del sujeto con su mundo. Pero el hábito también se requiere, según el mismo autor, para la identificación de causas y efectos. En otras palabras, la transición de las premisas:

(P_1) un cierto grupo de sucesos —sucesos del tipo *A*— son semejantes y otro grupo de sucesos —sucesos del tipo *B*— también son semejantes;

(P_2) ha habido una conjunción constante entre ejemplares de sucesos del tipo *A* y ejemplares de sucesos del tipo *B*;

a la conclusión:

(C) los ejemplares del tipo A continuarán a ser seguidos por ejemplares del tipo B;

es inductiva y es el hábito lo que hace posible tal inferencia (Hume 1975, pp. 42-43). Además, de acuerdo con la definición humeana de causa, (C) es equivalente a:

(C*) sucesos de tipo *A* son causas de los sucesos de tipo *B*.

Sin embargo, además de la fundamentación psicológica de la regla de inducción, ¿sería posible una fundamentación filosófica de dicha regla? La famosa tesis negativa de Hume sobre la causalidad y la inducción es que no puede haber una justificación filosófica de ninguna de las dos reglas (Hume 1975, p. 32). Regresaremos a este punto en la sección 3.

2. La Justificación de la Deducción

A pesar de lo que han pensado muchos filósofos —incluso Hume— hay un problema respecto a la deducción que es análogo al problema de la justificación de la inducción. Se podría plantear la cuestión de la justificación de nuestras prácticas deductivas en la ciencia y la matemática así como también en el lenguaje cotidiano. Algunos han intentado justificar, por ejemplo, las aplicaciones de las reglas de la lógica clásica de primer orden mostrando que hay una propiedad que poseen únicamente las reglas de esta lógica y que justifica el uso de estas reglas en los razonamientos llamados deductivos. La validez entendida como preservación de la verdad correspondería a tal propiedad semántica; se demuestra que el sistema de la lógica clásica es el cálculo más extendido que garantiza la validez de las inferencias deductivas en un lenguaje de primer orden con identidad. Estas son las conocidas pruebas de legitimidad y completud del cálculo de predicados de primer orden con la identidad.

El mostrar que las reglas de la lógica clásica son legítimas y en su totalidad completas dejó, quizá, satisfechos a los interesados en la cuestión metalógica de la equivalencia entre las nociones sintáctica y semántica de consecuencia lógica, pero parecen no haber complacido del todo a los filósofos preocupados con el problema de la justificación de la deducción. Muchos de estos últimos empezaron a sospechar de la existencia de circularidad en el argumento justificador. La idea es que cualquier argumento deductivo válido que se use para mostrar que las reglas de esta lógica preservan la verdad tiene que utilizar en alguno de sus pasos una o más de las reglas que se está buscando mostrar que son válidas. No parece haber una manera de escapar a una circularidad de este tipo.

Tomemos como ejemplo la legitimación del *modus ponens* (MP). Para mostrar que el MP es legítimo (o sea, que no puede haber una interpretación bajo la cual sus dos premisas sean verdaderas y la conclusión falsa) tendríamos que apelar en algún momento al *modus ponens* del metalenguaje. Pero si la justificación de una regla lógica tan

básica no puede prescindir del empleo de la misma, entonces o bien somos obligados a aceptar la circularidad del argumento justificador o bien optamos por distinguir entre dos tipos de *modus ponens*: aquél que se aplica al lenguaje objeto (MP_1) y un *modus ponens* metalingüístico (MP_2). Esta segunda alternativa nos abriría la posibilidad de rechazar la circularidad apelando a una regla semejante pero de segundo nivel. Sin embargo, su costo sería elevado: el saltar al nivel del metalenguaje, suscitaría la cuestión de la justificación de (MP_2), lo que requeriría un nuevo salto a un metalenguaje de segundo nivel y a un *modus ponens* de tercer nivel (MP_3) y, a la larga, a una proliferación infinita de metalenguajes y sus respectivos *modus ponens* (MP_n).

La salida menos costosa para los que continúan creyendo en la posibilidad de una justificación racional de la deducción ha sido reconocer la circularidad del argumento justificador y al mismo tiempo proponer —como lo ha hecho, por ejemplo, Michael Dummett (Dummett 1973, pp. 295-297)— una distinción entre argumentos persuasivos y explicativos para mostrar que dicha circularidad no siempre es viciosa (Dummett 1973, p. 296). La idea es que en los argumentos persuasivos la dirección lógica coincide con la dirección epistemológica; o sea, se conocen inicialmente las premisas y el argumento nos lleva a conocer la conclusión. Con los argumentos explicativos pasa el inverso: ya se conoce de antemano la conclusión y se buscan premisas a partir de las cuales se pueda derivar la conclusión. En el caso en cuestión, el que investiga el problema de la justificación de la deducción ya está convencido de la verdad de la conclusión del argumento justificador; a saber, la proposición de que un determinado conjunto de reglas lógicas garantiza la validez de cualquier transición que tenga la forma de alguna de ellas. Lo que busca no es un argumento que pueda persuadirlo de la verdad de esta proposición, sino premisas que puedan explicarla. Tal vez pudiésemos llamar los argumentos de esta especie "inferencias a la mejor explicación".

Ahora bien, la razón por la cual los argumentos persuasivos circulares no son convincentes es que asumimos en las premisas algo que queremos demostrar. En el ejemplo mencionado, utilizamos el *modus ponens* para demostrar que es una buena regla de inferencia en el sentido en que, por ejemplo, es una regla correcta. Claramente alguien que no esté aún convencido de esto no aceptará el argumento. Supongamos ahora que mi interlocutor, a quien deseo convencer de que el *modus ponens* es una buena regla, ya está convencido de esto; quizá debido a su éxito en nuestra práctica deductiva cotidiana o alguna otra razón. Sin embargo, le interesaría conocer una explicación filosófica razonable para tal corrección. En este caso, no importaría que este argumento explicativo utilizara el mismo *modus ponens* una vez que lo que está en juego no es la corrección de la regla sino la cuestión metafísica de que es lo que constituye su corrección.

3. Las justificaciones persuasiva y explicativa de la inducción

Si hubiera una lógica inductiva sería más fácil formular un problema análogo al de la justificación de la deducción. Quisiéramos saber cómo se podría explicar la corrección de las reglas de tal lógica inductiva. Corrección aquí significa simplemente que los argumentos evaluados de acuerdo con la lógica inductiva como inductivamente fuertes —el equivalente más próximo dentro de esta lógica a la validez— llevan la mayor par-

te del tiempo de premisas verdaderas a conclusiones verdaderas. No estamos por lo tanto buscando un argumento persuasivo en favor de la corrección no deductiva de nuestras mejores prácticas inductivas. Lo que nos interesa es cuál sería la mejor explicación de esta corrección.

Tales consideraciones respecto a la mejor explicación de la corrección de la inducción divergen, en mi opinión, de la preocupación de Hume sobre la posibilidad de justificar filosóficamente la inducción. Según lo entiendo, Hume considera sólo la posibilidad de un argumento persuasivo que nos inclinara en favor de la inducción y éste debería ser o bien inductivo o bien deductivo. Su pesimismo acerca de la posible existencia de tal argumento se debe a que, por un lado, si fuera deductivo él probaría demasiado, mientras que, por otro lado, si fuera inductivo nos haría caer en una circularidad. Lo que está detrás de la idea de que un argumento justificador deductivo para la corrección del razonamiento inductivo probaría demasiado es lo siguiente. Para empezar, quien buscara un argumento de este tipo tendría que garantizar que la forma inductiva de razonar que suponemos ha sido correcta en el pasado —en el sentido ya discutido— va a continuar a serlo en el futuro. Pero, ¿cómo un argumento deductivo nos podría arrojar una conclusión sobre el futuro partiendo de premisas sobre el pasado?

Si el único sentido posible de la justificación de la inducción fuera el persuasivo, entonces Hume tendría razón de ser pesimista acerca de cualquier propuesta para justificar esta forma de inferencia. Pues un argumento justificador inductivo obviamente no convencería a nadie que no estuviera ya convencido de la corrección las inferencias inductivas. Por otro lado, un argumento justificador deductivo tendría que garantizar la corrección futura de inferencias inductivas ya reconocidas como correctas en el pasado. Pero esto no es de ninguna manera lo que se está buscando. Lo que se desea argumentar, según Hume, es que la probabilidad de que nuestro uso futuro de las inferencias inductivas sea correcto dado que fue correcto en el pasado es bastante alta; esto es claramente un argumento inductivo.

El dilema humeano entre un argumento justificador necesariamente deductivo o inductivo se disuelve una vez que se ha rechazado su exigencia de que sea un argumento persuasivo. Mi propuesta, siguiendo a Dummett, es que tal argumento debe ser explicativo; él debe proporcionar la mejor explicación para la corrección persistente en el tiempo de nuestras prácticas inductivas. Pero, ¿qué forma lógica tendría una regla inductiva? Hume no nos da ninguna pista al respecto. Tendremos que buscar tal forma en el contexto de una lógica inductiva.

4 La Lógica Inductiva Bayesiana

El llamado enfoque bayesiano o lógica inductiva bayesiana se ha desarrollado a partir de los trabajos del filósofo inglés Frank Ramsey (Ramsey 1926) y del matemático italiano Bruno de Finetti (de Finetti 1937), pero sus orígenes se remontan al descubrimiento del cálculo de probabilidades por Blaise Pascal y Pierre de Fermat en el siglo

XVII. El teorema de Bayes, que utiliza el concepto de probabilidad condicionada[1], es uno de los resultados que demuestran el carácter inductivo de la lógica bayesiana. Una de sus formulaciones es la siguiente:

> (TB) la probabilidad de una determinada hipótesis *h* condicionada a la evidencia *e* es igual a la probabilidad de *e* condicionada a *h* multiplicada por la probabilidad de *h* divididas ambas por la probabilidad de *e*.

Ramsey buscaba introducir una nueva interpretación de la probabilidad como una medida de los grados de creencia de los sujetos racionales en un determinado contenido proposicional. Más específicamente le interesaba encontrar un método para calcular al mismo tiempo los grados de creencia y los grados de deseo de un sujeto racional a partir de la información acerca de sus preferencias por determinados cursos de acción relevantes para la situación de decisión que se pretende investigar. El problema para el cual Ramsey finalmente encontró una solución brillante es en último análisis el problema fundamental de la teoría de la decisión: dada información suficiente sobre la escala de preferencias del agente (*A*) por los diversos cursos de acción posibles para su situación de decisión, encontrar el patrón de deseos y creencias de *A*.[2]

Una de las tesis fundamentales de la lógica bayesiana es que la noción de probabilidad interpretada de manera subjetivista debe ser utilizada para medir los grados de creencia de los agentes humanos. La aplicación del cálculo de probabilidades como escala para la medida de los grados de creencia de los agentes humanos impone diversas restricciones al sistema de dichas creencias de un sujeto que razona más o menos de acuerdo con las reglas de dicho cálculo. Por ejemplo, si un agente atribuye a una proposición la probabilidad subjetiva x y si además es coherente[3], entonces debe atribuirle a su negación la probabilidad subjetiva $1 - x$.

Las reglas del cálculo de probabilidades establecen cómo se debe comportar el patrón de probabilidades subjetivas de un agente supuestamente coherente en un determinado momento[4], pero no nos dice nada sobre cómo deberían cambiar tales probabilidades con el paso del tiempo. La regla bayesiana más sencilla que gobierna los cambios temporales de las probabilidades subjetivas es la llamada regla de condicionalización de Bayes, que afirma lo siguiente:

> (RCB) si la probabilidad subjetiva de una determinada proposición *p* sube a 1 y si la probabilidad condicionada a *p* de cualquier proposición *q* (P(*q*/*p*))

---

[1] Esta probabilidad, que denotamos como P(*q*/*p*), expresa la siguiente idea. Sean todos los mundos posibles en que *p* es verdadera. La fracción de tales mundos en que *q* también es verdadera corresponde a la probabilidad de que *q* dado que *p*.

[2] Esto está en Ramsey (1926).

[3] La coherencia aquí no tiene que ver con la obediencia al principio de no-contradicción de la lógica clásica, sino más bien a la no violación de las reglas de la lógica inductiva bayesiana. Atribuir probabilidades sujetivas a una proposición y su negación cuya suma fuera diferente de 1 ejemplificaría una incoherencia en este segundo sentido.

[4] Este patrón corresponde a la función de probabilidad sobre el dominio de todas las proposiciones asociadas al lenguaje de un sujeto en un tiempo dado.

no cambia como resultado del cambio temporal en la probabilidad de *p*, entonces se debe actualizar la probabilidad de *q* a P(*q*/*p*). Si (i) $P_i(p) = x < 1$ y $P_f(p) = 1$ y (ii) $P_i(q/p) = P_f(q/p)$, entonces $P_f(q) = P_i(q/p)$.[5]

El principio de condicionalización de Bayes nos ofrece el ejemplo más claro de una regla inductiva. Pues, supongamos que se substituye por *p* un determinado contenido proposicional —digamos, que el aumento de temperatura de una cierta olla expresa llena de agua hasta la mitad es acompañado por un aumento de la presión interna del vapor de agua confinado en su interior— cuya probabilidad antes del experimento es menor que 1. Supongamos además que se substituye por *q* la hipótesis de que la presión de un gas ideal a volumen constante es directamente proporcional a su temperatura, cuya probabilidad anterior al experimento en cuestión también es menor que 1. La evidencia proporcionada por este experimento y un sin número de otros similares deberían conducir a un incremento de la probabilidad de la hipótesis *q*. La regla de Bayes describe la corrección de una inducción de este tipo del siguiente modo: sea cual fuera la probabilidad que le asigna un sujeto coherente a *q*, cada vez que se incorpora conocimiento nuevo confirmador de la misma a su sistema de creencias, él debería reajustar la probabilidad subjetiva que le atribuye a *q* de acuerdo con la RCB.

5. El bayesianismo y la justificación persuasiva de sus reglas inductivas

Varios de los que han contribuido al desarrollo de la lógica bayesiana se han dedicado también a una justificación persuasiva de dichas reglas. Debemos a Ramsey y a de Finetti un argumento que se conoce como el teorema del libro de apuestas holandés (*the Dutch-book argument*). En este argumento se propone mostrar que el cálculo de probabilidades —el conjunto de reglas no-inductivas de la lógica bayesiana— es la representación más perfecta de la coherencia del patrón de grados creencias de un agente humano idealizado[6]. El argumento explora la idea de que un apostador cuyo patrón de grados de creencia no obedece a las reglas del cálculo de probabilidades ganaría o perdería —según ocupara la posición de apuntador o la de su oponente— cualesquiera que fueran los resultados de sus apuestas. Bajo la hipótesis de que un agente coherente no va a apostar sabiendo que perderá sea cual sea el resultado, su patrón de creencias debe satisfacer los axiomas y teoremas de dicho cálculo como condición de preservación de la coherencia de sus decisiones con respecto a las apuestas que hace.

---

[5] Estas dos condiciones están claramente explicitadas en Jeffrey 2002, pp. 46, 48. Jeffrey denomina la primera condición necesaria para la condicionalización RCB la condición de certeza y la segunda de condición de rigidez. Las dos juntas constituyen, según él, las condiciones necesarias y suficientes para la aplicación de la RCB. La condición de certeza se llama así porque el tipo de condicionalización sobre una proposición *p* propuesto por la RCB solo se puede aplicar en los casos en que la probabilidad de *p* se cambia a 1. La condición de rigidez debe su nombre a que en el proceso de cambio de la probabilidad subjetiva asociada a *q* cuando la función probabilidad asociada a otra proposición *p* cambia su valor a 1 en el mismo intervalo la probabilidad condicionada P(*q*/*p*) en este mismo intervalo de tiempo no se cambia.

[6] Un sujeto humano idealizado en el sentido de que sus grados de creencia en un determinado momento obedecen a las reglas del cálculo de probabilidades.

La estrategia usada por Ramsey y de Finetti es la siguiente: ellos tratan de demostrar que el apostador perdería o ganaría siempre para los casos más sencillos en los que sus grados de creencia violaran los axiomas del cálculo de probabilidades. Con esto queda demostrado que lo mismo vale para cualquier violación de sus teoremas[7]. El uso del teorema de Ramsey-de Finetti para justificar las reglas no-inductivas de la lógica bayesiana ha, naturalmente, sugerido la siguiente cuestión: ¿habrá algún argumento semejante al del libro de apuestas holandés que pudiera justificar de manera persuasiva las reglas inductivas de esta lógica y en particular la RCB?

Ramsey y de Finetti no ofrecen ninguna justificación para esta regla inductiva. Pero en el inicio de los 70s, David Lewis y Paul Teller anunciaron haber encontrado una estrategia semejante al libro de apuestas holandés para justificar la RCB. El argumento de Lewis-Teller[8] parte de la suposición de que la probabilidad subjetiva posterior de una hipótesis h es distinta de la probabilidad anterior de h condicionada a e ($P_f(h) \neq P_i(h/e)$). El argumento trata de exhibir una estrategia combinada de apuestas en h dado e, en e, y, finalmente, en h usando las probabilidades $P_i(h/e)$, $P_i(e)$ y $P_f(h)$ en sus respectivos cálculos, tal que el apostador pierde todo cualquiera que sea el valor de verdad de la condición (e) o de la hipótesis (h). Esto muestra, según los autores, que la RCB es la estrategia más razonable para el cambio de nuestros grados de creencia, dada la ocurrencia de una determinada evidencia[9].

Mucho se ha escrito en contra del argumento de Lewis-Teller. Howson y Urbach (Howson & Urbach 1993, cap. 6), por ejemplo, no han aceptado que tal argumento pueda ofrecer algún tipo de justificación a la RCB. Esto porque si la condición (ii) $P_i(q/p) = P_f(q/p)$[10] es indispensable para la aplicabilidad de esta regla de condicionalización, entonces se sigue lógicamente de (ii) y de (i) [$P_i(p) = x < 1$ y $P_f(p) = 1$] que $P_f(q) = P_i(q/p)$.[11] Tal y como la enunciamos arriba la RCB es un enunciado analítico y, por lo tanto, la pregunta por su justificación persuasiva no tendría sentido una vez que en este momento no está en juego alguna posible alternativa al canon de analiticidad que nos ofrece la lógica clásica. En todo caso, para lo que sí tendría sentido pedir una justificación persuasiva serían las condiciones (i) y (ii). Empecemos por pensar en cómo se podría justificar la segunda.

---

[7] La prueba de Ramsey y de Finetti se puede consultar en cualquier texto introductorio sobre bayesianismo. Ver, por ejemplo, Howson & Urbach 1993, cap. 5.

[8] Sobre él, ver Teller 1973, pp. 222-224.

[9] Para una descripción más detallada de la estrategia de Lewis-Teller, además del texto original de Teller, se puede consultar el texto de Howson y Urbach (Howson & Urbach 1993, cap. 6).

[10] La que Jeffrey llama de condición de rigidez, Howson y Urbach denominan condición de invariancia.

[11] Si se supusiera que la condición (ii) no es indispensable, entonces, según Howson y Urbach, obviamente no sería una verdad lógica que:

(RCB') Si (i) entonces $P_f(q) = P_i(q/p)$.

Pero, el argumento de Lewis-Teller tampoco justificaría la nueva formulación de la condicionalización de Bayes (RCB') una vez que para ellos la idea misma de que pudiera haber una incoherencia diacrónica entre el valor $P_i(q)$ y el valor $P_f(q)$ que fuera diferente de $P_i(q/p)$ —o sea, que violara la RCB'— no tiene el mínimo sentido.

Howson y Urbach ofrecen un ejemplo ilustrativo[12] de que no siempre está justificada la condición de invariancia ($P_i(q/p) = P_f(q/p)$) y que además el argumento de Lewis-Teller no ofrece ninguna justificación para ella. Según Howson y Urbach, no hay ninguna incoherencia diacrónica en violarla aunque sí haya una incoherencia (lógica en el sentido clásico) en una estrategia de cambio de creencias que viole la condicionalización de Bayes cuando obtienen las condiciones de certeza e invariancia. Esta última condición estaría justificada solamente en aquellos casos en que la observación no afectara nuestras estimaciones condicionadas a lo que se ha observado sobre la verdad de otras proposiciones. En los casos en que la aplicación de la condición de invariancia no se justifica tendríamos el problema adicional de cual sería la manera correcta de actualizar nuestras probabilidades subjetivas; hasta donde sé, no hay una regla alternativa para estos casos.

Las situaciones en las cuales la condición (i) no se aplica —aquellas en que, como resultado de la observación, no se cambian las probabilidades subjetivas de los enunciados de observación e_1, e_2, …, e_n a 1— son más fáciles de tratar porque ahí sí hay una regla de condicionalización alternativa. Ésta es la llamada regla de condicionalización de Jeffrey que ordena lo siguiente:

(RCJ) si las probabilidades subjetivas de las proposiciones p_n, (n = 1, ..., m) que forman una partición del espacio de probabilidades se cambian de $P_i(p_n)$ a $P_f(p_n)$ y si la probabilidad condicionada a p_n de cualquier proposición q ($P(q/p_n)$) no cambia como resultado del cambio temporal en la probabilidad de p_n, entonces se debe actualizar la probabilidad de q a $P_i(q/p_1) \times P_f(p_1) + ... + P_i(q/p_n) \times P_f(p_m)$. Si (iii) $P_i(p_n) = x \neq y = P_f(p_n)$ y (iv) $P_i(q/p_n) = P_f(q/p_n)$ [para n = 1, ..., m], entonces $P_f(q) = P_i(q/p_1) \times P_f(p_1) + ... + P_i(q/p_n) \times P_f(p_m)$.[13]

También se ha intentado extender el argumento de Lewis-Teller a la RCJ con la finalidad de mostrar que ella representa la única estrategia coherente a seguir si se acepta las condiciones (iii) y (iv); un ejemplo de un intento en esta dirección nos lo ofrece Brian Skyrms[14]. Aquí también se aplican las ya mencionadas observaciones de Howson y Urbach sobre la ausencia de incoherencia diacrónica en la violación de la condi-

---

[12] El ejemplo está en Howson & Urbach 1993, cap. 6, pero es fácil imaginar otros ejemplos semejantes.

[13] Esto está en Jeffrey 1983, cap. 11.

[14] En Skyrms (1987). En este texto magníficamente claro, Skyrms revisa el argumento de Lewis-Teller para la RCB, propone un nuevo libro-de-apuestas-holandés para la RCJ y, finalmente, presenta argumentos inversos del tipo del libro-de-apuestas-holandés en favor de ambas. Un argumento del tipo libro-de-apuestas-holandés muestra que si no respeta la regla de condicionalización en la actualización de las probabilidades subjetivas asociadas a enunciados no-observacionales un agente está sujeto a una estrategia del libro de apuestas holandés. Un argumento inverso del tipo del libro-de-apuestas-holandés muestra más bien que si respeta la regla de condicionalización en la actualización de las probabilidades subjetivas asociadas a estos enunciados el agente es invulnerable a cualquier estrategia del tipo del libro de apuestas holandés. Vale la pena observar que Skyrms, Teller y también Howson y Urbach optan principalmente por discutir la justificación de la formulación (RCB') de la condicionalización de Bayes. De manera análoga para la (RCJ), ellos examinan primordialmente la posibilidad de una justificación persuasiva de:

(RCJ') Si (iii) $P_i(p_n) = x \neq y = P_f(p_n)$, entonces $P_f(q) = P_i(q/p_1) \times P_f(p_1) + ... + P_i(q/p_n) \times P_f(p_m)$

ción (iv) en ciertas situaciones y, por lo tanto, sobre la carencia de justificación persuasiva para el argumento del tipo del libro-de-apuestas-holandés para una regla de condicionalización concebida sin condición de invariancia.

Además, hay seguramente varios otros presupuestos en la teoría bayesiana los cuales también requerirían de justificación persuasiva. Uno de ellos es que no sería racional en el sentido bayesiano[15] apostar a sabiendas de que se va a perder pase lo que pase si los únicos deseos del apostador que son relevantes para la situación de decisión están representados por sus ganancias monetarias. No sería racional elegir este curso de acción en lugar de otros en los que uno no perdería dinero en todos los casos (por ejemplo, aquél en que decidiera no aceptar esta apuesta tan desventajosa) simplemente porque un agente que actuara de esta manera estaría violando el principio básico de la teoría de la decisión bayesiana, el llamado principio de la maximización de la utilidad esperada. Lo podríamos expresar como el siguiente imperativo:

(PMUE) Actúa de manera a maximizar la función de utilidad media asociada al curso de acción que elegiste entre los varios cursos de acción posibles en una dada situación de decisión.

La pregunta por la justificación persuasiva del principio de maximización de la utilidad esperada se puede extender de manera obvia a todos los principios de la teoría bayesiana del razonamiento: los que anteriormente hemos denominado principios de coherencia del patrón de creencias del sujeto e inclusive los principios de coherencia más clásicos como la no-contradicción, el tercero excluido y otros principios de la lógica clásica que también son parte constitutiva de la teoría bayesiana. Pero si la cuestión de la justificación de los principios básicos del bayesianismo nos lleva directamente a la noción misma de la racionalidad bayesiana caracterizada de manera implícita por el conjunto de tales principios, ¿cómo podríamos justificar persuasivamente la elección de dicha noción frente a sus rivales sin caer en la circularidad viciosa (como ya lo había denunciado Hume) o en un regreso al infinito?

6. Una justificación explicativa para los principios básicos de la lógica bayesiana

Una salida posible sería apelar a algo semejante a la respuesta kantiana al problema humeano de la justificación persuasiva del principio de causalidad. En consonancia con la sugerencia de Dummett respecto a la circularidad no-viciosa de un argumento justificador explicativo, Kant propone que la mejor explicación filosófica de la objetividad del contenido de nuestros juicios sintéticos empíricos descansa en que ésta requiere la intervención de principios organizadores tales como el de causalidad y el de

[15] Este sentido de racionalidad corresponde a la concepción instrumentalista de la racionalidad según la cual ésta se explica en términos de una relación de adecuación entre medios y fines de un agente humano. En este sentido, un agente racional sería aquél que busca los medios más adecuados para lograr satisfacer sus fines. La expresión más precisa de esta concepción de la racionalidad humana está dada por el principio de la maximización de la utilidad esperada que será discutido más adelante. Esta concepción de la racionalidad, que ya está presente en Aristóteles, es la que se encuentra implícita en la mayor parte de la tradición empirista.

la permanencia de la sustancia; esto significa que nosotros no podríamos hablar con verdad o falsedad de objetos o sucesos relacionados de manera causal o como cosas que permanecen a pesar del cambio de sus accidentes si no poseyéramos estas y otras reglas de organización de la multiplicidad de nuestras sensaciones[16].

Según Kant, una aplicación inmediata de los principios organizadores —como la causalidad— a la generación del contenido de nuestros juicios sintéticos empíricos nos ofrece la física newtoniana, en donde el principio de causalidad funciona como una regla para buscar causas del movimiento de un cuerpo material. Kant denomina esta aplicación del principio de causalidad en la mecánica el principio de inercia (Kant 1970, pp. 104-105).

En tanto regla constitutiva del contenido objetivo de nuestro conocimiento científico, la causalidad es considerada acertadamente por Kant como un principio de la razón teórica en la medida en que hace posible la existencia del contenido causal de los juicios sintéticos de experiencia. La discusión de su justificación está en la sección de la *Crítica* llamada "Analítica Trascendental" que a su vez es parte de la "Lógica Trascendental". La justificación kantiana de la causalidad es netamente circular aunque no viciosa, pues tiene la forma de una inferencia a la mejor explicación del hecho, según Kant universalmente aceptado, de que nuestros juicios empíricos son objetivos. La objetividad de estos juicios se explica a su vez en términos de la posibilidad de verdad y falsedad del contenido de ellos y estos contenidos se constituyen a partir de la operación de los principios organizadores de la multiplicidad sensible como, por ejemplo, el principio de causalidad.

La justificación explicativa de la aplicación de tales principios para constituir el contenido de los juicios sintéticos *a posteriori* tanto de la ciencia de la naturaleza como del conocimiento cotidiano toma la forma, según Kant, de una argumentación trascendental, esto es: se conciben estos principios como condiciones de posibilidad de la experiencia, o como se dijo arriba, condiciones de posibilidad del contenido objetivo de todos los juicios empíricos. La circularidad de la justificación explicativa kantiana de la causalidad estriba en que la experiencia requiere de causalidad para su constitución mientras que la causalidad requiere de la experiencia para su correcta aplicación. La virtuosidad de la circularidad de esta explicación consiste en que la justificación kantiana es más bien explicativa. Quien acepta tal explicación ya está convencido de la corrección del empleo de la causalidad al mundo de la experiencia; lo que busca es simplemente una respuesta filosófica al porqué de dicha corrección.

Vamos a ilustrar el argumento justificador kantiano para la causalidad por medio de un ejemplo que aparece en los *Prolegómenos a toda Metafísica Futura*. Tomemos el juicio de percepción —en la terminología kantiana, un juicio cuyos elementos de contenido se relacionan de manera meramente subjetiva (Kant 1783, § 18)— con el siguiente contenido *si el sol brilla sobre la piedra ella se calienta*. Para que tal contenido como un todo se pueda concebir como objetivo y, por lo tanto, pueda ser verdadero o falso, es necesaria la intervención de la regla de la causalidad que lo transforma en el siguiente

[16] Sobre la causalidad, por ejemplo, ver Kant 1787, B 234. Esta es la manera *standard* de referirse a la paginación de la segunda edición de la *Crítica de la Razón Pura* (Kant 1787).

contenido: *la iluminación del sol es la causa del calentamiento de la piedra* (Kant 1783, § 20, nota **).

No me interesa aquí entrar demasiado en la terminología kantiana, por ejemplo, sobre la conexión necesaria entre conceptos o la validez universal de juicios. Me basta que se me conceda la interpretación según la cual un juicio de percepción, si bien puede incluir contenidos conceptuales, el juicio mismo no tiene todavía un contenido proposicional objetivo —esto es, que sea verdadero o falso. El punto de Kant es que se requiere de principios organizadores como la causalidad y la permanencia de la sustancia para constituir tal contenido proposicional. Sin duda, hay que reconocer el enorme esfuerzo kantiano para mostrar, apelando a la forma lógica de cualquier juicio, que estos principios operan confiriendo contenido a todos los juicios con una cierta forma lógica. Por ejemplo, el principio de causalidad daría un contenido causal a juicios de la forma hipotética y el principio de permanencia de la sustancia otorgaría el contenido de la relación sustancia-accidente a juicios de la forma categórica —esto es, de la forma sujeto-predicado.

Sin embargo, muy probablemente, hoy nadie aceptaría que la totalidad de los principios del entendimiento propuestos por Kant como los responsables por el contenido objetivo de nuestros juicios empíricos lo sean realmente aunque algunos de ellos como la causalidad se puedan reconocer todavía hoy día como principios constitutivos de nuestra cognición de, e interacción con, el mundo. Más específicamente, se rechazaría la tesis de que la causalidad pudiera fungir como principio semántico para todo un conjunto de proposiciones empíricas de una determinada forma. La idea de un principio semántico que estaría en juego aquí es la que se conoce desde Alfred Tarski; en este sentido, los principios semánticos determinarían las condiciones de verdad y falsedad de todas las proposiciones de este conjunto[17]. Por otro lado, también se rechazaría la idea de que un principio semántico deba operar sobre un contenido ya dado (una relación meramente subjetiva entre los contenidos ya establecidos) para generar un contenido proposicional objetivo. Esto sería concebir la teoría del contenido de nuestras representaciones —conceptos y juicios— como descansando sobre algún tipo de contenido más primitivo: una de las maneras, como diría Wilfrid Sellars, de caer en el mito de lo dado (Sellars 1956).

A pesar de todas estas críticas y muchas otras que se ha dirigido a la justificación explicativa kantiana de la causalidad, opino que se la podría aprovechar de la siguiente manera. Podríamos afirmar que la regla de causalidad y formulaciones adecuadas de las reglas inductivas son principios constitutivos de la cognición y acción humanas en tanto que son indispensables para la construcción de una teoría de la mentalidad y del contenido lingüístico de un hablante en una situación radical en que tenemos a nuestra disposición únicamente evidencia en términos de las conductas lingüísticas y no-

[17] En una semántica tarskiana (ver Tarski 1933, secciones 2, 3 y 4), un teorema semántico expresaría la condición de verdad de una oración del lenguaje para la cual se está explicitando su semántica; en esta misma semántica, un axioma semántico daría, por ejemplo, las condiciones de satisfacción de los predicados primitivos del lenguaje. La causalidad no podría fungir como principio semántico ni el sentido de un teorema de la teoría semántica ni tampoco en el sentido de un axioma de la misma.

lingüísticas del hablante —en resumen, la situación de interpretación radical. Esto lo han sugerido Willard van Quine[18] y más claramente algunos de sus discípulos —por ejemplo, Donald Davidson—; ellos han afirmado que sin tales principios no podríamos atribuir contenido a las palabras de nuestros semejantes y motivación a sus conductas intencionales[19]. En otras palabras, no podríamos interpretarlos como hablando con sentido y actuando intencionalmente a la luz de sus deseos y creencias. Lo dice Davidson en los siguientes pasajes:

> The question whether a creature "subscribes" to the principle of continence, or to the logic of the sentential calculus, or to the principle of total evidence for inductive reasoning, is not an empirical question. For it is only by interpreting a creature as largely in accord with these principles that we can intelligibly attribute propositional attitudes to it, (…) (Davidson 1985, p. 352)

> (…) I should never have tried to pin you down to an admission that you ought to subscribe to the principles of decision theory. For I think everyone does subscribe to those principles, whether he knows it or not. This does not imply, of course, that no one ever reasons, believes, chooses, or acts contrary to those principles, but only that if someone does go against those principles, he goes against his own principles. (Davidson 1985, p. 351)

No me voy a detener a discutir en este momento muchos de los principios que Davidson toma como constitutivos de una teoría unificada de la interpretación y de la acción de un hablante-agente como, por ejemplo, los mencionados en la cita: los principios de continencia y de evidencia total[20]. Lo importante a recalcar ahora es que, de acuerdo con Davidson, las reglas inductivas de la teoría bayesiana están claramente incluidas entre estos principios. A continuación, daremos a conocer el lugar indispensable de las reglas inductivas bayesianas en una teoría general del lenguaje y de la mentalidad humanas.

Para empezar, habría que indagar sobre cuál sería el papel de una teoría radical de la interpretación y de la acción. Lo que se busca es dar una respuesta plausible para el problema de la constitución y de la epistemología de la mentalidad intencional (creencias, deseos, comprensión del lenguaje, etc.). El problema está íntimamente ligado al que ya le preocupaba a Kant (a saber: el de cómo se constituyen los contenidos objetivos de nuestros juicios sintéticos a posteriori a partir del múltiple sensible que nos son dados en la experiencia), una vez que los contenidos o significados de las oraciones que escuchamos de nuestros interlocutores así como sus actitudes proposicionales deben ser inferidos a partir de sus proferencias, de su conducta observable y de nuestro conocimiento previo sobre ellos.

Desde la perspectiva interpretativista más conocida a partir de los escritos de Quine y Davidson, la estrategia más adecuada para solucionar este problema debería partir

---

[18] Quine lo dice por primera vez en los primeros capítulos de su monumental Quine (1960).

[19] Davidson, por ejemplo, lo dice en varios textos. Uno de ellos es Davidson (1985).

[20] El principio de continencia sería equivalente al principio de la maximización de la utilidad esperada. Alguien que violara este principio estaría actuando en contra de su mejor juicio—este sería un caso claro de debilidad de la voluntad (*akrasia*). El principio de evidencia total ordenaría al agente a aceptar la hipótesis que esté más corroborada por la totalidad de la evidencia a su disposición. Se podría interpretar este principio como la formulación de sentido común de los principios de condicionalización bayesianos.

de la siguiente cuestión: ¿qué tendríamos que conocer para poder entender el lenguaje de un hablante y al mismo tiempo conocer sus deseos y creencias? Debemos privilegiar la situación radical en que no tenemos conocimiento previo ni de los significados de sus palabras ni de su mentalidad para evitar la circularidad presente en cualquier intento de caracterizar el contenido lingüístico en términos de la mentalidad o viceversa. Como parte esencial de la mentalidad humana, la competencia lingüística se requiere de manera indispensable para identificar las actitudes proposicionales por lo menos parcialmente a través de sus contenidos; por otro lado, el contexto de sus actitudes proposiconales también es esencial para la individuación de los contenidos de las oraciones proferidas por un hablante. En el caso humano, lenguaje y mentalidad están inextricablemente ligados.

Además del carácter radical del enfoque davidsoniano, está presente la idea de que la explicación de nuestro conocimiento del lenguaje y de la mentalidad debe empezar por el acceso de tercera persona a los contenidos lingüísticos y a las actitudes proposicionales y no, como era usual en la tradición anterior, por el conocimiento de primera persona sobre ellos. Una vez que se acepta este punto de partida, entonces lo más natural sería pensar que el conocimiento de tercera persona sobre la mentalidad intencional debe tomar la forma de una teoría empírica del significado y de la acción. Pero, ¿qué forma tendría que tener tal teoría? Inspirado en el trabajo de los etnólogos que se ocuparon de entender comunidades indígenas hasta entonces aisladas de otras culturas humanas, Quine había sugerido que la parte del significado de dicha teoría debería ser caracterizada como un manual de traducción que especificara la sinonimia entre cada oración del lenguaje alienígena y su contraparte en el lenguaje del etnólogo (o traductor radical).

Más explícito sobre la forma de la teoría unificada del significado y de la mentalidad[21], Davidson ha mantenido que la parte de la red de contenidos lingüísticos sería regimentada por una teoría tarskiana de las condiciones de verdad de todas las oraciones de por lo menos un fragmento bastante significado del lenguaje desconocido[22]. La construcción de tal teoría de la verdad para el lenguaje desconocido tomará en cuenta la evidencia dada por un sin-número de oraciones aceptadas como verdaderas o rechazadas como falsas por un hablante, según un intérprete adecuadamente posicionado. A continuación, veremos más detalladamente como funciona la parte de la corroboración empírica de la teoría unificada.

La parte de las actitudes proposicionales representada por el patrón de creencias y deseos de los extranjeros deberá ser sistematizada por una teoría de la decisión bayesiana en los moldes de la formulación propuesta por Richard Jeffrey (1983) modifica-

---

[21] Sobre la forma de la teoría unificada, ver, por ejemplo, Davidson (1980b), Davidson (1984) y Davidson (1990).

[22] La idea es que se pueda fijar el significado del fragmento constituido por las oraciones declarativas del lenguaje extranjero por medio de una teoría de sus condiciones de verdad. Esto sería suficiente para determinar el significado de sus palabras constituyentes, lo que contribuiría para determinar el significado de las oraciones no-declarativas una vez que fuéramos capaces de identificar, vía la teoría unificada, los actos de habla que ejemplifican.

da de tal manera a aplicarse a las oraciones del hablante y no a sus respectivos contenidos proposicionales. El cambio se debe a que no podemos en el contexto de una teoría radical de la constitución del lenguaje y de la mentalidad presuponer como ya conocidos los contenidos proposicionales del agente-hablante. Como ya sabemos, la sistematización teórica de los deseos y creencias en conexión con la acción de una persona guiada por sus preferencias hacía diversos cursos de acción en detrimento de otros debe arrojar su red de deseos y creencias expresada en términos de intensidades de deseo (deseabilidades) y grados de creencia (probabilidades subjetivas)[23]. Los contenidos de tales deseos y creencias será sistematizado por la parte de la teoría unificada que trata de los significados.

Al igual que la teoría del significado, la teoría de la mentalidad también será una teoría empírica a ser contrastada con evidencia dada por las preferencias relativas[24] observables de un agente. Para los propósitos de la teoría unificada, será necesario buscar un tipo de evidencia que sea adecuado a las dos partes de la misma (a saber: la parte del significado y la parte de las actitudes) y que no presuponga de antemano nada de lo que se desea explicar. Hablaremos más sobre ello a continuación.

También sería importante decir algo sobre los principios *a priori* de la teoría unificada. En tanto empírica, la teoría unificada contiene una parte *a posteriori* o revisable que se ejemplifican, en el caso de la teoría del significado, por las T-oraciones que dan las condiciones de verdad, desde un metalenguaje suficientemente poderoso, de las oraciones de un determinado lenguaje objeto *L*[25] como también por los axiomas que dan las condiciones de satisfacción de los predicados y la referencia de los nombres propios de *L* respectivamente. En el caso de la teoría de la decisión, es empírica la escala de preferencias del agente así como también el patrón de intensidades de creencia y de deseo a ser inferido de tal escala. Lo que no puede ser *a posteriori* son los principios constitutivos de ambas teorías. Me refiero a los principios que las identifican *como* teorías tarskiana de las condiciones de verdad de las oraciones de *L* y bayesiana de la decisión racional respectivamente.

Pero, ¿qué principios serían estos? Ya hemos mencionado por lo menos dos de los principios constitutivos de la teoría de la decisión: el principio de continencia (maximización de la utilidad esperada) y el principio de la evidencia total (alguna de las reglas de condicionalización bayesianas). Más adelante, hablaremos del papel constitutivo de las reglas de condicionalización en la teoría bayesiana. Otro principio constitutivo de esta teoría es el llamado principio de transitividad aplicado a las preferencias de un agente racional (*S*), la regla que afirma que si *S* prefiere la situación descrita por la proposición *p* a aquella expresada por *q* y si además *S* prefiere *q* a *r*, entonces también

---

[23] Por supuesto, hay muchas otras actitudes proposicionales a ser descubiertas. La expectativa aquí es que todas las otras actitudes distintas de los deseos y creencias puedan ser vistas como combinaciones diversas de un elemento cognitivo (una creencia) y de un elemento conativo (un deseo).

[24] Relativas en el sentido de que se trata de una comparación para un agente entre por lo menos dos situaciones posibles y su inclinación por una de ellas en detrimento de la otra.

[25] Si el metalenguaje fuera, por ejemplo, el inglés una T-oración para el español sería:

(T) The sentence "la nieve es blanca" is true in Spanish if and only if snow is white.

preferirá *p* a *r*.[26] Digamos que este es un principio constitutivo de la noción bayesiana de preferencia; esto significa que si lo sometemos al tribunal de la experiencia y eventualmente lo rechazamos por razones empíricas nos quedaríamos sin una teoría de la acción humana en el sentido de que la pudiera conectar causal y racionalmente con sus motivaciones psicológicas (los deseos y las creencias del agente).

Es indiscutible que ciertos principios lógicos como la no-contradicción deberían necesariamente entrar en esta lista. Si no fuera así, ¿cómo podríamos tener éxito en la atribución de valores de verdad a las oraciones de nuestros interlocutores? A este respecto, dice Quine:

> Or consider the familiar remark that even the most audacious system-builder is bound by the law of contradiction. How is he really bound? If he were to accept contradiction, he would so readjust his logical laws as to insure distinctions of some sort; for the classical laws yield all sentences as consequences of any contradiction. But then we would proceed to reconstrue his heroic novel logic as a non-contradictory logic, perhaps even as familiar logic, in perverse notation. (Quine 1960, p. 59)

En tanto gobiernan las relaciones inferenciales deductivas entre los contenidos proposicionales de las creencias o de las oraciones de un agente-hablante, las reglas de lógica deductiva contribuirían para fijar, por lo menos parcialmente, tales contenidos. La razón es que una oración o una creencia tiene el contenido que tiene en parte por su posición en una red de otras proposiciones o otras creencias ligadas entre sí por medio de relaciones inferenciales deductivas e inductivas. Esto significa que también principios lógicos inductivos como las reglas de condicionalización bayesianas contribuen para fijar el contenido proposicional de las oraciones de nuestro lenguaje y de nuestras actitudes. Veamos esto con más detalle.

Un papel crucial de las reglas bayesianas de condicionalización[27] es que permiten al intérprete (*I*) actualizar las probabilidades subjetivas por él asociadas a todas las oraciones no-observacionales del hablante (*S*) como consecuencia de los cambios de las probabilidades subjetivas asociadas a las oraciones de observación de *S*[28] que *I* identifica en la conducta lingüística de *S*. La interpretación de las oraciones observacionales de *S* procede por la observación de los cambios de las probabilidades subjetivas asociadas a estas proposiciones como resultado de las situaciones en el mundo que *I* supone que están causando estos cambios. Pero, la empresa interpretativa empieza por la identificación de las constantes lógicas en el lenguaje de S. Veamos resumidamente cómo describe Davidson esta empresa[29] y qué papel constitutivo le atribuye, por un lado, a las re-

---

[26] La noción de transitividad que manejan algunos psicólogos cognitivos como Amos Tversky es ligeramente distinta de la que expresa el axioma mencionado arriba. Tversky, por ejemplo, habla de transitividad estocástica débil (Tversky 1969, p. 31). Consideramos que la diferencia es irrelevante para los propósitos de la presente discusión.

[27] RCB o su refinamiento: la ya mencionada regla de condicionalización de Jeffrey (RCJ).

[28] Relacionadas lógicamente estas últimas con las primeras como evidencia e hipótesis.

[29] Descripciones quizá más detalladas de la empresa interpretativa se encuentran también en Davidson (1980b) y Davidson (1984).

glas de condicionalización bayesianas y, por otro lado, al principio de causalidad en este proceso:

> Jeffrey's version of decision theory, applied to sentences, tells us that a rational agent cannot prefer both a sentence and its negation to a tautology, or a tautology to both a sentence and its negation. This fact makes it possible for an interpreter to identify, with no knowledge of the meanings of the agent's sentences, all of the pure sentential connectives, such as negation, conjunction and the biconditional.[30] This minimal knowledge suffices to determine the subjective probabilities of all of the agent's sentences […] and then, in turn, to fix the relative values of the truth of those sentences […]. The subjective probabilities can then be used to interpret the sentences. For what Quine calls observation sentences, the changes in probabilities provide the obvious clues to first-order interpretation when geared to events and objects easily perceived simultaneously by interpreter and the person being interpreted. Conditional probabilities and entailments between sentences, by registering what the speaker takes to be evidence for his or her beliefs, provides the interpreter with what is needed to interpret more theoretical terms and sentences. (Davidson 1995, pp. 9-10)

Sabiendo, por ejemplo, que, de acuerdo con la teoría bayesiana, la escala de preferencias de *S* respecto a la verdad de las oraciones de su idioma debería ser tal que su preferencia por la verdad de una tautología siempre se ubique entre su preferencia por la verdad de una oración a la de su negación respectivamente, *I* ya puede identificar la negación y, en general, los conectivos proposicionales del lenguaje de *S*,[31] sin conocer todavía los significados de las oraciones en juego. Una vez identificada la estructura lógica proposicional del lenguaje (*L*) de *S* y la escala de sus preferencias por la verdad de las oraciones de *L*, se puede entonces echar a andar la maquinaria de la teoría bayesiana para calcular las probabilidades subjetivas que *S* asocia a cada una de estas oraciones y sus respectivas deseabilidades.[32]

La interpretación de oraciones deberá empezar al nivel observacional y de preferencia por las oraciones observacionales más simples de *L*.[33] Pues, por medio de una especie de triangulación involucrando situaciones en el mundo, las reacciones lingüísticas del hablante a tales situaciones y la observación por el intérprete de estas reacciones y de las situaciones mismas, logrará este último interpretar estas primeras oraciones de su interlocutor. Todo lo que *I* necesita es relacionar los cambios en el patrón de probabilidades subjetivas asociadas a estas oraciones con los cambios en las configuraciones relevantes de objetos (o sucesos) en el mundo que *I* observa simultáneamente a

---

[30] Una descripción más detallada del proceso de identificación de las constantes lógicas utilizando el aparato de la teoría de Jeffrey nos ofrece Davidson en el apéndice de su artículo del 90 sobre la verdad (Davidson 1990).

[31] Una vez descubierta la negación, *I* puede encontrar la disyunción observando qué oración se comporta como una tautología respecto a la preferencia por su verdad cuando se la obtiene a partir de la oración y de su negación. Los otros conectivos proposicionales se obtienen fácilmente a partir de sus equivalencias lógicas con oraciones cuyos únicos conectivos lógicos son la negación y la disyunción.

[32] No es que las probabilidades subjetivas y las deseabilidades se apliquen a las oraciones del idioma de *S*; ciertamente se aplican a sus contenidos proposicionales. Es que en esta etapa de la aplicación de la teoría de la decisión el intérprete no conoce estos contenidos, por lo que utiliza las oraciones de *S* como sus representantes.

[33] Este procedimiento es el que aparece descrito de manera bastante resumida en la última cita.

las reacciones lingüísticas de su interlocutor y a la evolución temporal de su patrón de preferencias respecto a estas oraciones[34]. Esta triangulación le permite al intérprete mapear sus propias oraciones observacionales más simples sobre las de su interlocutor.

La interpretación de las oraciones no-observacionales y de los términos teóricos de *L* requiere conocimiento de las relaciones lógicas deductivas y inductivas entre estas oraciones y las observacionales. Una relación lógica entre ellas extremamente relevante en este contexto es la de probabilidad condicionada; pensemos, por ejemplo, en la probabilidad de una hipótesis (*h*) condicionada a la evidencia observacional (*e*) a su favor. Conociendo dicha probabilidad para diversas oraciones observacionales e_i ($i = 1, \ldots, n$), la relación lógica entre *h* y las e_i y además la interpretación de las e_i, estará el intérprete en condiciones de inferir el contenido de *h* y de sus términos teóricos.

Las reglas de condicionalización entran en la escena para explicar la evolución temporal de las probabilidades subjetivas asociadas a las oraciones más teóricas del hablante pues los cambios temporales en estas probabilidades gobernados por las condicionalizaciones bayesianas contribuyen para revelar al intérprete las relaciones lógicas entre estas oraciones y otras menos teóricas del lenguaje *L*. ¿Cuál sería entonces el papel del principio de causalidad en tal explicación? Hemos hablado de un tipo de triangulación a través del cual el intérprete llega a mapear oraciones de su propio idioma sobre las de su interlocutor[35]. Ahora bien, la posibilidad de dicha triangulación descansa sobre el discernimiento de relaciones causales entre las situaciones en el mundo y las reacciones lingüísticas del hablante (*S*), entre estas mismas situaciones y ciertas creencias del intérprete (*I*) sobre estas situaciones y, finalmente, entre las reacciones lingüísticas de *S* tal y como las observa *I* y sus creencias sobre la causa mundana de estas reacciones.

Si Davidson tiene razón con relación a la atribución de contenido a las oraciones observacionales más simples de *S* por triangulación, entonces son relaciones causales las que conforman los lados del triángulo intepretativo, cuyos vértices serían el mundo, el sujeto *S* y un intérprete suyo. Vemos que para Davidson, así como para Kant, la causalidad es un principio indispensable para la existencia de contenido proposicional. La diferencia entre los dos filósofos radica en que, mientras que para Kant la causalidad es una de las reglas que capacitan a los sujetos humanos para sintetizar tales contenidos a partir de los datos sensibles, para Davidson la causalidad es una regla indispensable para la atribución interpretativa de significado y mentalidad a otras personas.

Sin embargo, se podría legítimamente preguntar: ¿en que sentido estaría Davidson ofreciendo una explicación de la corrección de las reglas inductivas bayesianas? Una pregunta análoga se aplicaría al principio de causalidad respecto a las supuestas explicaciones kantiana y davidsoniana de su corrección. Ciertamente que una explicación

---

[34] Mencionamos que para la teoría unificada se hace necesaria la búsqueda de un tipo de evidencia que no presuponga conocimiento previo ni de la estructura de significados ni de la red de creencias y deseos del sujeto-agente. Davidson considera que la preferencia por la verdad de una oración en comparación con la verdad de otra oración satisface esta condición.

[35] Sobre la triangulación davidsoniana, ver Davidson (1992).

posible de la corrección de nuestro empleo de estas reglas sería la de que hay un mundo con el que interactuamos y que está gobernado por leyes causales; en cierto sentido, el uso exitoso de los principios de causalidad y de inducción (entendido este último a la manera bayesiana) reflejaría una especie de armonía entre nuestras mentes y el mundo. Pero, tal vez una explicación más simples fuera que hay un mundo que se comporta de manera regular aunque la causalidad y la inducción son reglas que nosotros imponemos sobre los estímulos lingüísticos y no-lingüísticos que provienen de este mundo para generar contenido proposicional y mentalidad. La mayor simplicidad de esta explicación estaría en que no requiere la presuposición de la existencia de relaciones causales o inductivas en el mundo independientes de nuestra intervención cognitiva y práctica sobre él. Por otro lado, no puede explicar la corrección de tal aplicación de estos principios apelando a sus contra-partes en el mundo; corrección, para Davidson y tal vez también para Kant[36], significa apenas que ésta es *la* aplicación de tales reglas capaz de generar contenido objetivo[37]. De acuerdo con Davidson, dicha aplicación es la única capaz de producir no solo contenido como también mentalidad objetiva.

Ahora me gustaría discutir, aunque no exhaustivamente, algunas críticas a la explicación davidsoniana del lenguaje y de la acción humanas. Empiezo por la que afirma que esta explicación se apoya sobre una teoría falsa de la acción humana. Me refiero a la siguiente afirmación sobre la teoría bayesiana: se ha insistido de manera vehemente en que los seres humanos violan de manera sistemática los principios constitutivos de la teoría de la decisión cuando actúan a la luz de sus deseos y creencias y que, por lo tanto, la teoría bayesiana, en su aplicación a la acción humana en conexión con sus motivos internos, es empíricamente falsa. Si esto fuera correcto, seríamos forzados a abandonar la propuesta davidsoniana del lenguaje y de la mentalidad humanos. El espacio de este ensayo no comporta una discusión detallada y abarcadora de la verdad o falsedad empírica ni del carácter *a priori* o *a posteriori* de la teoría bayesiana. Sin embargo, la cuestión de la posibilidad de someter los principios constitutivos de la teoría bayesiana al tribunal de la experiencia es demasiado relevante para que la dejemos de lado en la presente discusión.

Tomemos como ilustración el ya mencionado principio de la transitividad de la preferencia. Ahora bien, varios experimentos psicológicos cuidadosos han sido propuestos para mostrar que los agentes sistemáticamente violan este principio[38]. Confrontado entonces con experimentos tales como los llevados a cabo por Amos Tversky en que, dispuestas en una misma escala, las preferencias de los sujetos en ciertas situaciones de decisión son ineludi-

---

[36] Hago la calificación respecto a Kant porque muchos de sus intérpretes probablemente no estarían de acuerdo en aproximarlo tanto a Davidson.

[37] Recordemos la afirmación kantiana según la cual la aplicación de la causalidad a las cosas en sí (*noumena*) resultan en juicios que no son ni verdaderos ni falsos. Por ejemplo, el juicio "Dios es la causa de la existencia del mundo".

[38] Uno de los más conocidos está reportado en Tversky (1969).

blemente intransitivas[39], ¿estaría Davidson obligado a conceder que la teoría bayesiana ha sido refutada o que algunos de sus principios constitutivos son falsos y por lo tanto deben ser eliminados? Antes de arriesgar una respuesta, hagamos algunas aclaraciones pertinentes.

Hasta el momento, hemos supuesto que la teoría bayesiana podría ser considerada como una teoría empírica con principios más centrales, los cuales hemos llamado constitutivos de la misma, y otros más periféricos, que son aquellas afirmaciones pertenecientes a una cierta aplicación de la teoría cuyos contenidos dependen de la evidencia dada por la conducta de los sujetos experimentales. Estamos suponiendo obviamente que la teoría bayesiana es aplicable para describir la mayoría de las acciones humanas a la luz de deseos y creencias; esto es, además de los axiomas que definen implícitamente las nociones de preferencia, utilidad y creencia también se está proporcionando una interpretación de la teoría en la medida en que se ofrecen escalas para medir estas cantidades[40]. También es esencial observar que en su aplicación a la explicación y predicción de la acción humana, la teoría bayesiana no es meramente empírica; también la estamos considerando como una teoría normativa[41]. Esto significa que la estamos utilizando no solamente para describir la conducta de los seres humanos en situaciones concretas sino también para evaluarla a luz de los cánones de racionalidad impuestos por la teoría bayesiana.

Regresemos entonces al caso de la supuesta falsedad del principio de transitividad de las preferencias. El propio Tversky, por lo menos en un primer momento, ha tomado los experimentos en cuestión como refutatorios de la aplicación de la teoría bayesiana para sistematizar la acción humana en relación con preferencias, deseos y creencias; consistente con tal actitud ha buscado teorías alternativas. Sin embargo, si nuestro propósito es dar una explicación sistemática del lenguaje y de la acción humana tal salida no puede satisfacernos del todo.

En primer lugar, no queda claro que los experimentos hayan realmente mostrado la falsedad empírica del axioma de transitividad, pues se podría reinterpretar los resultados experimentales de manera a eliminar la aparente intransitividad. Esto porque dichos experimentos dan por sentado que la interpretación que están dando experimen-

---

[39] Es importante resaltar que los experimentos propuestos por Tversky en este artículo (Tversky 1969, p. 32, 33-34, 37-38) siempre involucran elecciones dependientes de valoraciones en por lo menos dos dimensiones distintas. Así, por ejemplo, en una situación en que se debe escoger entre varios solicitantes para un puesto de trabajo las elecciones tomaban en cuenta, por un lado, su inteligencia medida por algún test de tipo IQ y, por otro lado, su experiencia laboral. Los casos de intransitividad pueden aparecer cuando dentro del universo de solicitantes la escala de utilidades asociada a la inteligencia para cada candidato es inversamente proporcional a la escala de utilidades asociada a su respectiva experiencia laboral. El problema aquí es que las dimensiones parecen interactuar generando preferencias intransitivas.

[40] El concepto observacional aquí es el de preferencia; podemos observar las preferencias de un agente en situaciones de decisión concretas. Los conceptos de deseo y creencia son teóricos. Sin embargo, también para ellos se está proporcionando escalas de medida: para el caso de las creencias, la escala de probabilidades; para los deseos, una escala más o menos arbitraria de utilidades.

[41] Así la están tomando Davidson y Suppes (Davidson & Suppes 1957, cap. 1).

tador y sujeto a las opciones presentadas verbalmente por el primer es la misma[42]. No obstante, a la luz del problema más general que estamos intentando resolver, tal presuposición es exageradamente optimista. También se podría cuestionar la aparente falla de la transitividad en los experimentos en cuestión si lanzamos alguna sospecha sobre la adecuación de la caracterización de la escala de utilidades atribuida por los experimentadores a los sujetos. Por ejemplo, muchos de los experimentos para mostrar la intransitividad involucran apuestas; pero, ¿cómo saber si las ganancias o pérdidas económicas miden confiablemente las utilidades de los sujetos? Muchos otros valores como, por ejemplo, la aversión al riesgo, interfieren normalmente con el valor del dinero inclusive en situaciones de apuestas.

En segundo lugar, también se podría mencionar una serie de experimentos que muestran, al contrario, que con el paso del tiempo los sujetos se tornan cada vez más consistentes respecto a sus preferencias[43]. A pesar de lo sorprendente que puedan parecer estos experimentos cuando se toma en cuenta que los experimentadores tuvieron el cuidado de que esta mejora de la consistencia de los sujetos no se debiera al condicionamiento ni tampoco al aprendizaje, Davidson no los ha utilizado para concluir que la transitividad está empíricamente confirmada. Esto porque excluyendo el aprendizaje y el condicionamiento, no hay ninguna buena explicación de tal 'mejora'. Recordemos que la teoría de la decisión es inicialmente una teoría estática sobre el patrón de preferencias, creencias y valores de un sujeto; no nos dice como él debe evolucionar absolutamente en el tiempo[44].

En tercer lugar, lo que se vaya a concluir sobre la transitividad de la preferencia también debe extenderse naturalmente a los otros principios constitutivos de la teoría bayesiana y, en particular, a las reglas de condicionalización. Con relación a estas reglas, la crítica más frecuente ha sido que no siempre lo más racional es actualizar las probabilidades subjetivas de acuerdo con las reglas de condicionalización[45]. Pero, ya hemos discutido en la sección anterior una manera de incorporar tal crítica, a saber: imponiendo restricciones sobre la aplicación de las reglas de condicionalización; en los casos en que no se satisface la condición de invariancia, la condicionalización podrá no ser la estrategia más racional a seguir.

---

[42] Ver, por ejemplo, Davidson 1974, pp. 235-8 y Davidson 1976, p. 270.

[43] El propio Davidson menciona una serie de experimentos de este tipo en Davidson 1974, pp. 235-6.

[44] La excepción serían las reglas de condicionalización que imponen ciertas restricciones sobre la evolución temporal de las probabilidades subjetivas asociadas a ciertas creencias del sujeto y del experimentador (por ejemplo, sus creencias más teóricas) condicionada a la dinámica de las probabilidades subjetivas asociadas a otras de sus creencias lógicamente conectadas con las primeras (por ejemplo, algunas de sus creencias observacionales). Pero, ya sabemos que estas reglas apenas establecen una relación entre la evolución temporal de unas y la evolución temporal de las otras y además que todavía en los casos en que se aplican lo hacen solo bajo ciertas restricciones (ver sección 5).

[45] Vimos en la sección 5 la crítica de Howson & Urbach a la supuesta incoherencia diacrónica de la estrategia de condicionalización cuando no se satisface la condición de invariancia. Lo mismo se puede concluir del artículo de Frank Arntzenius (Arntzenius 2003). Este último contiene, además, varios ejemplos distintos del que presentan Howson & Urbach de la falla de la condición de invariancia.

¿Qué se puede entonces concluir sobre el carácter empírico de los principios centrales de la teoría de la decisión? Me parece que las consideraciones precedentes son suficientes para confirmar lo que ya habíamos afirmado páginas arriba, a saber: que quizás la mejor estrategia sería tomar estos principios como constituyendo la parte, digamos, *a priori* de la teoría de la decisión bayesiana.

En todo caso, se podría intentar rechazar la teoría bayesiana como lo han intentado Tversky, en algún momento, y muchos otros psicólogos cognitivos. Según Davidson, ésta sería la alternativa más costosa una vez que en su opinión no hay una teoría igualmente poderosa y simples que permita identificar los patrones de creencias y de deseos de un agente en relación causal y racional con su acción[46]. Pero, ¿qué justificación ofrece para una opinión tan pretensiosa? En mi opinión, su justificación es circular aunque no viciosa. Como vimos, su proponente la presenta como la mejor explicación filosófica de la posibilidad de atribución de contenido a nuestras palabras como también de la posibilidad de una teorización sobre nuestra acción intencional. Nosotros ya estamos convencidos de manera independiente[47] de que hay contenido intencional y mentalidad. Sin embargo, en tanto filósofos buscamos una explicación que nos satisfaga respecto a la cuestión de cómo el contenido intencional y la mentalidad son posibles. Apelamos a los principios constitutivos de la mentalidad para mostrar cómo se podría generar una mentalidad de este tipo por medio de los principios en cuestión; por otro lado, apelamos al hecho mismo de la existencia de la mentalidad humana para señalar la correcta aplicación de sus principios constitutivos.

Como argumento explicativo, la justificación davidsoniana de los principios básicos de la teoría de interpretación radical comparte con otros argumentos parecidos la característica de que su dirección lógica —de los principios constitutivos de la teoría unificada del discurso y de la acción como premisas al contenido intencional y a la mentalidad como conclusión— no coincide con su correspondiente dirección epistemológica que parte del conocimiento de que hay mentalidad y contenido lingüístico para acceder al conocimiento de que los principios de constitución de tal mentalidad y contenido deben ser los que permiten construir una teoría del significado y de la mente de un agente-hablante.

7. Conclusión

No conozco ningún texto suyo en que Davidson afirme estar ofreciendo una justificación en el sentido explicativo para el principio de causalidad y las reglas de condicionalización bayesianas. Tampoco sé de algún pasaje donde Michael Dummett reconozca

---

[46] Insisto en la afirmación davidsoniana de que los principios de la teoría de la decisión son constitutivos de los conceptos mentales. Esto de ninguna manera implica que la teoría psicológica como un todo sea *a priori*; más bien, según lo entiendo, Davidson propone que tales principios definen implícitamente los conceptos mentales en el contexto de la teoría bayesiana, la cual solamente en conjunción con la psicología popular se torna una teoría empírica. Sobre esto, ver Davidson 1976, pp. 272-274 y también Davidson 1970, pp. 220-221.

[47] Si no estuviéramos convencidos de ello no trataríamos los otros como si tuvieran mentalidad y tampoco esperaríamos que los otros nos trataran como si tuviéramos mentalidad.

la posibilidad de la justificación explicativa de la lógica deductiva y inductiva en términos de una teoría general del significado y de la mentalidad. Seguramente, Hume no reconoce la posibilidad de una justificación que no sea persuasiva para la causalidad y la inducción. Así que si Hume tiene razón en su preocupación filosófica por una justificación de estos principios fundamentales y si Dummett está en lo correcto en que tal justificación puede tomar la forma de una explicación filosófica de la corrección de su aplicación y además si Davidson cumple su promesa de darnos una teoría correcta de la interpretación del discurso y de la acción humanas, entonces por lo menos dicha teoría tiene la ventaja de proporcionarnos un argumento trascendental a favor del equivalente contemporáneo de los principios de causalidad y de inducción.[48]

REFERENCIAS

Arntzenius, F. (2003). "Some Problems for Conditionalization and Reflection", *The Journal of Philosophy* 100 (7), 356-370.

Davidson, D. (1970). "Mental Events", Davidson (1980a), 207-227.

——— (1974). "Psychology as Philosophy", Davidson (1980a), 229-239.

——— (1976). "Hempel on explaining Action", Davidson (1980a), 261-275.

——— (1980a). *Essays on Actions and Events*. Oxford: Oxford University Press.

——— (1980b). "A Unified Theory of Thought, Meaning, and Action", *Grazer Philosophische Studien* 11, 1-12.

——— (1984). "Expressing Evaluations", Lindley Lecture monograph, University of Kansas.

——— (1985). "Incoherence and Irrationality", *Dialectica*, 39 (4), 345-354.

——— (1990). "The Structure and Content of Truth", *The Journal of Philosophy*, 87 (6), 279-328.

——— (1992). "The Second Person", *Midwest Studies in Philosophy* 17, 255-67.

——— (1995). "Could there be a Science of Rationality", *International Journal of Philosophical Studies* 3 (1), 1-16.

——— & P. Suppes (1957). *Decision Making*. Stanford: Stanford University Press.

de Finetti, Bruno (1937). "La prevision: ces lois logiques, ces sources subjectives", *Annales de l'Institut Henri Poincaré* 7, 1-68.

Dummett, M. (1973). "The Justification of Deduction", Dummett (1978), 290-318.

——— (1978). *Truth and Other Enigmas*. London: Duckworth.

Howson, C., & P. Urbach (1993). *Scientific Reasoning*. 2nd ed. Chicago: Open Court.

Hume, D. (1777). *Essays and Treatises on several Subjects*, vol. 2. London: T. Cadell.

——— (1975). *Enquiries concerning Human Understanding and concerning the Principles of Morals*. 3rd ed. Oxford: Clarendon Press.

Kant, E. (1783). *Prolegomena zu einer jeden künftigen Metaphysik*. Riga: Hartnoch.

——— (1787). *Kritik der reinen Vernunft*. Riga: Hartnoch.

Kant, E. (1970). *Metaphysical Foundations of Natural Science*. Traducción inglesa de Kant (1786). New York: The Bobbs-Merrill Company.

Jeffrey, R. (1983). *The Logic of Decision*. Chicago: The University of Chicago Press.

——— (2002). *After Logical Empiricism*. Lisboa: Colibri.

Quine, W. (1960). *Word and Object*. Cambridge: the MIT Press.

Ramsey, F. (1926). "Truth and Probability", in D. Mellor (1990), *F.P. Ramsey Philosophical Papers*. Cambridge: Cambridge University Press.

Sellars, W. (1956). "Empiricism and the Philosophy of Mind", *Minnesota Studies in the Philosophy of Science* 1, 253-329.

Skyrms, B. (1987). "Dynamic Coherence and Probability Kinematics", *Philosophy of Science* 54, 1-20.

Tarski, A. (1933). *Pojecie prawdy w jezykach nauk dedukcyjnych*. Varsovia: Warsaw Academy of Sciences.

[48] Agradezco a las excelentes observaciones de los dos dictaminadores de *Theoria* a una versión anterior de este artículo.

Tarski, A. (1956). "The Concept of Truth in Formalised Languages". Traducción inglesa de Tarski (1933), Tarski (1956), pp. 152-278.
——— (1956). *Logic, Semantics, Metamathematics*. Oxford: Oxford University Press.
Teller, P. (1973). "Conditionalization and Observation", *Synthese* 26, 218-258.
Tversky, A. (1969). "Intransitivity of preferences", *Psychological Reviewk,* 76, 31-48.

Sílvio Pinto es professor titular del Departamento de Filosofía de la Universidad Autónoma Metropolitana, Unidad Iztapalapa, desde el 2002. En esta Universidad, ha dirigido la revista *Signos Filosóficos* de 2003 a 2005 y actualmente coordina la Línea de Historia y Filosofía de la Ciencia del Posgrado en Humanidades. Sus intereses de investigación giran en torno a la filosofía del lenguaje, de la mente y de las matemáticas. Entre sus más recientes publicaciones están la coordinación del libro *Bertrand Russell y el análisis filosófico a partir de "On denoting"* (UAM/Juan Pablo's 2007) y los artículos "Naturalism and the Metasemantic Account of Concepts" (*Abstracta* 2006), "Los conceptos abiertos y la paradoja del análisis" (*Theoria* 2005).

Dirección: Departamento de Filosofía. UAM-Iztapalapa. Av. San Rafael Atlixco, 186, Col. Vicentina, 09340 Iztapalapa, México D.F. E-mail: pint@xanum.uam.mx.