# LA INDETERMINACION DE LA RACIONALIDAD Y SU RELACION CON EL PROBLEMA DE NEWCOMB

*Federico Matías Pailos*
*CONICET/UBA/GAF*
*fpailos@hotmail.com*

**Resumen**

¿Qué es actuar racionalmente? Presentaré el esbozo de una teoría acerca de la racionalidad mínima, compatible tanto con los enfoques formales evidencialistas y causalistas. El criterio defendido será el compatible con un número sustantivamente mayoritario de nuestras intuiciones. Argumentaré en su favor. Evalúo a continuación algunos aspectos del problema de Newcomb. Con respecto a la variante del problema que plantea la idea de un Predictor infalible, argumentaré que lo racional es optar solo por la caja opaca. La concepción de la racionalidad defendida permite explicar también por qué es incorrecto afirmar aquí se premia la irracionalidad. Sostendré que los escenarios de Newcomb son de muchos más tipos que los imaginados por Levi. En el hecho de que el planteo original pueda ser completado de formas tan diversas se halla buena parte de la explicación del peso intuitivo que tienen ambas respuestas al problema de Newcomb.

**Palabras Claves:** racionalidad; agencia; intuiciones; explicación.

Cuando se abordan problemas relacionados con la racionalidad, suelen plantearse dos dicotomías que, se estima, vertebrarán la teoría a desarrollar. Ellas son (a) un enfoque normativo contra una aproximación descriptiva al problema; (b) una teoría robusta o sustantiva de la racionalidad, contra una presentación 'mínima' de ella. En lo que sigue me concentraré en algunas ideas normativas y relacionadas con el costado 'mínimo' de la racionalidad.[1]

¿Cuándo una decisión es racional? Quizás, pensando en un enfoque evolucionista (descriptivo), podemos estar tentados a decir: la decisión racional es la que maximiza nuestras ventajas adaptativas. Pero esta respuesta acaso no sea del todo feliz. Imaginemos el siguiente escenario: tenemos que afrontar la decisión de aceptar o rechazar un trabajo. Quien nos presenta la oferta es un individuo que sabe cuál va a ser nuestro futuro en ambos casos, y el futuro no va a ser el mismo. Si aceptamos el trabajo, ganaremos más dinero, tendremos una vida más holgada, plena de satisfacciones. En suma, seremos más felices. Pero no tendremos hijos. Si rechazamos la oferta, por otro lado, ganaremos menos dinero, viviremos de zozobra en zozobra, nuestras satisfacciones serán muchísimo más limitadas. Sin embargo, tendremos una

prole numerosa. Como la reproducción es una de las variables adaptativas, una sustancial, parece sensato afirmar que esta es la conducta adaptativamente más eficaz. Sin embargo, ¿diríamos que rechazar el trabajo constituye la elección racional en este caso? Lo dudo. Por otra parte, una concepción normativa y evolucionista de la racionalidad debe dar satisfacción a algunos interrogantes. No es claro que la unidad de selección, aquello sobre lo que actúa la evolución, sean los individuos. Las opiniones más difundidas más bien tienden a negarlo. Algunos, acaso los más, piensan que lo que se selecciona son los genes. Otros entienden que la unidad de selección son las especies, o que también son las especies. ¿La elección racional es la que maximiza las posibilidades adaptativas de nuestros genes o la de nuestra especie? No veo que haya una respuesta clara al respecto. Más aún: sospecho que cualquiera de ambas vías pierden un punto sustantivo de la idea de 'elección racional', aquella que la relaciona con la maximización de la felicidad, o la satisfacción, o el más eficaz cumplimiento de nuestros deseos. ¿Es la racional aquella elección que maximiza, en cada caso, la satisfacción del conjunto de nuestros deseos? Quizás. Sin embargo, cuando hablamos de elecciones racionales, las pensamos acotadas a situaciones, contextos o escenarios específicos. Al evaluar si una decisión es racional o no, no suele tomarse en cuenta la totalidad de los deseos del individuo, sino un subconjunto relevante de ellos. Esto no quiere decir que no deba evaluarse, en cada caso, la totalidad de ellos. Parece, sin embargo, que podemos contentarnos con esos subconjuntos relevantes de deseos. Hay, al menos en una idea muy extendida de la racionalidad de la elección, un componente contextual. Una elección es racional con respecto a una situación dada, situación que incluye algunos de los deseos del individuo, aunque no (necesariamente) la suma de ellos.

Clásicamente, en el campo de la teoría de la decisión racional, hay dos vertientes principales a considerar, dos tipos de candidatos a dar el equivalente formal de la idea de elección racional: el enfoque bayesiano y el enfoque causal. Veamos primero una versión del bayesiano, la presentada, por ejemplo, por Horacio Arló Costa en su artículo "Racionalidad y teoría de la acción: ¿es la teoría evidencial de la decisión una teoría de la racionalidad mínima?".[2] La idea, sucintamente, es la siguiente: debemos elegir aquella opción que maximice la utilidad subjetiva esperada. La utilidad subjetiva esperada es una utilidad o valor subjetivo que satisface los siguientes tres principios estructurales de una teoría evidencial de la decisión clásica: (1) para cada agente al que la teoría se aplica, es posible identificar un dominio apropiado de entidades sobre las que el agente puede tener preferencias, así como un conjunto adecuado de entidades a las cuáles el agente puede asignar probabilidades (lo llamaremos 'dominio de probabilidad'); (2) los agentes a quienes la teoría se aplica atribuyen probabilidades sobre el dominio de probabilidad que obedecen a los axiomas clásicos de la probabilidad de Kolmogorov, (3) los agentes a los que la

teoría se aplica ordenan el dominio de preferencias de una manera adecuada, siendo el orden en cuestión al menos consistente y completo. Por caso, en el sistema de Richard Jeffrey, uno de los clásicos en teoría evidencial (bayesiana), se parte de un álgebra booleana de proposiciones X, con máximos y mínimos elementos T y F. Un modelo de decisión definido en X es un conjunto M de pares <P, V>, donde cada P es un función de probabilidad sobre X (el dominio de probabilidad) y cada V es una función de utilidad o valor definida sobre X - {F} (el dominio de preferencia). La función de probabilidad satisface los siguientes tres axiomas: (1) la probabilidad de toda proposición A es mayor o igual a 0, (2) la probabilidad de T es 1, (3) la probabilidad de la disyunción de dos proposiciones incompatibles es igual a la suma de sus probabilidades. Para evitar que la satisfacción de los tres principios estructurales convalide como racionales preferencias intuitivamente irracionales, tenemos que introducir un cuarto axioma: $V(A \vee B) = [P(A).V(A) + P(B).V(B)] / (P(A) + P(B))$.[3] Este axioma nos dice cómo calcular la utilidad de una disyunción: la utilidad de una disyunción $V(A \vee B)$ será el promedio ponderado de V(A) y V(B). Esto permite definir a tipos de proposiciones como buenas, malas o indiferentes. Una proposición A será buena si V(A) > V(T), esto es: si su deseabilidad es mayor que la del status quo expresado por V(T). Como se colegirá, A será mala si V(T) > V(A), e indiferente si V(T) = V(A).[4] Este fragmento de la teoría evidencial no dice nada acerca de cómo las evaluaciones subjetivas de utilidad de probabilidad de los agentes racionales se ajustan a la llegada de nueva información. La siguiente extensión determina que un modelo de decisión M en un instante t contiene no sólo el par <P, V> que representa las creencias y deseos del agente en t, sino también, para todo enunciado A, contiene los pares <P'A, V'A>, que determinan cómo <P, V> debiera cambiarse en caso de que A fuese aprendido. Con la introducción los siguientes dos axiomas, quedarán determinadas las relaciones entre P y P'A y las de V con V'A. Todo modelo de decisión M que contenga un par <P, V>, contiene también, para cada proposición A, tal que P(A) es distinto de 0, pares <P'A, V'A> tal que:

(5) $P'A (B) = P(A \wedge B) / P (B/A)$
(6) $V'A (B) = V (A \wedge B)$

La idea de (5) es que la probabilidad P'A (B) está determinada por la probabilidad condicional de B con respecto a A. El axioma (6) determina el valor de todo enunciado B dado por V'A.

Miremos ahora, sucintamente, una versión, rival a la evidencial, del enfoque causal de la decisión racional. Expondré los rudimentos de la versión de la teoría causal presentada en el artículo clásico de Allan Gibbard y William Harper, "Counterfactuals and two kinds of expected utility".[5] Según los autores, la decisión racional

supone la consideración de proposiciones condicionales. Cuando se evalúa tomar posibles decisiones de peso, es racional que quien considere tomarlas sopese qué pasaría en caso de realizar este o aquél acto. Es racional considerar algunas proposiciones contrafácticas, proposiciones del estilo "Si yo fuera a realizar a, entonces c ocurriría". "$\Box\!\!\rightarrow$" será el símbolo usado para la implicación contrafáctica. La anterior proposición, entonces, se formalizará así: a $\Box\!\!\rightarrow$ c. También se lo encontrará como "Si hago a $\Box\!\!\rightarrow$ c sucederá".

En este tipo de consideraciones, no parece razonable asumir que los individuos saben qué pasará si realiza esta o aquella acción. Sí parece razonable exigirles que asignen probabilidades a esas proposiciones contrafácticas. Con estas probabilidades, más las "desabilidades" o "utilidades", también medidas, que asigne a los hipotéticos resultados de un acto dado a, es posible calcular la utilidad esperada del acto a. Si a tiene como posibles resultados ["outcomes"] $o_1$, …, $o_n$, la utilidad esperada de a es la sumatoria siguiente:

$\Sigma_i$ prob (si hago a $\Box\!\!\rightarrow o_i$ sucederá) D $o_i$

Donde D $o_i$ es la deseabilidad de $o_i$. Allí donde las teorías bayesianas tradicionales, la de Jeffrey, por ejemplo, usan probabilidades condicionales, Gibbard y Harper apelan a probabilidades no condicionadas de contrafácticos. Se asume que los contrafácticos son proposiciones auténticas, y no se exige que tengan antecedente falso: el agente racional estudiado evalúa contrafácticos en los que el antecedente 'a' es una acción que realizará, y otros en los que 'a' es una acción que no realizará. Se entenderá que afirmar A $\Box\!\!\rightarrow$ S no equivale a decir que el que A sea el caso llevará a que S sea el caso. A $\Box\!\!\rightarrow$S será verdadera si esto último ocurre, pero también si S es el caso, sea o no que ocurra A. Sea 'a' un acto que en el tiempo t decido realizar. Un mundo-a es un mundo posible como el actual antes de t, en el que decido en t hacer a y en el que hago a, y en el que rigen las leyes naturales de ahí en más. $W_a$ es un mundo-a que, en t, es el más parecido al actual. Las diferencias en las condiciones iniciales entre $W_a$ y el actual son las mínimas indispensables para hacer a la proposición 'hago a' verdadera, y deben ser todas parte del aparato de decisión racional del agente. (Aquí solo importa la similaridad entre mundos en el momento de la decisión.) Entonces

"Hago a $\Box\!\!\rightarrow$S" es verdadera si y sólo si "Hago a $\Box\!\!\rightarrow$S"es verdadera en $W_a$.

Cada acto tiene diferentes resultados ["*outcomes*"] que el agente (que cada agente, para cada acto) considera posibles. Cada resultado será representado por un letra 'o' más un subíndice. La proposición que rescate (exprese) un acto, significará también todas las consecuencias posibles del acto que interesan ('positiva o negativamente', digamos) al agente. El agente otorga una cierta deseabilidad, cuantificada, a cada

resultado. La deseabilidad del resultado 'o' será señalado por medio de la expresión 'Do'. El agente sabe que un y solo un resultado será el que de hecho corresponda a su acto. La utilidad esperada de A, calculada para probabilidades de contrafácticos [U(A)], será, entonces:

$$U(A) = \Sigma_j \text{ prob}(A \mathrel{\square\!\!\rightarrow} O_j)\, DO_j$$

La utilidad esperada de A, calculada para probabilidades condicionales [V(A)], es decir, el tipo de utilidad esperada defendida por Jeffrey y (con mínimas variaciones, en particular en el tipo de probabilidad empleada) por las teorías bayesianas clásicas, como la presentada más arriba (extraída del artículo de Arló Costa), es, recuerda, la siguiente:

$$V(A) = \Sigma_j \text{ prob}(O_j/A)\, DO_j$$

De acuerdo a Gibbard y Harper, la utilidad esperada de un acto es medida por U(A). V(A), en cambio, mide *el grado de bienvenida que el agente da a la noticia de que está por realizar el acto A*. V(A) mide el valor de A en tanto noticia que recibe el agente. V(A) y U(A) no son equivalentes.[6]

¿Cuál de ambas perspectivas determina lo que es racional hacer? ¿La bayesiana, la causal, quizás ninguna…? Al presentar lo que denomina 'la (así llamada) teoría económica de la racionalidad", Arló Costa expone lo que llama un 'axioma de racionalidad', el principio CDP (por 'comportamiento dotado de un propósito').[7]

(CDP) Cuando un agente X confronta un conjunto de alternativas que están a su disposición, X exhibe un comportamiento racional en el caso de que elija aquellas alternativas que lo llevan a situarse en la mejor posición posible dadas las circunstancias.

Entiendo que esta es una idea correcta. No creo que ni el bayesiano ni el causalista la objeten. Claro: ¿qué es 'la mejor posición posible'? ¿Cuáles son 'las circunstancias' relevantes? El principio es correcto, pero muy vago. Ambos enfoques se presentan como respondiendo estas preguntas de un modo más adecuado que el otro. Insisto con mi pregunta: ¿cuál de ambos, si alguno lo hace, precisa y rescata de modo formal esta idea? Una forma de resolver la cuestión es presentando una serie de casos relevantes, y ver cuál de ambos es compatible con un mayor número de ellos.[8] Un caso relevante tradicional es el problema de Newcomb. Gibbard y Harper presentan un número de situaciones similares.[9] Ahora bien: ni el enfoque causal ni el bayesiano dan cuenta de todas las intuiciones. A veces la intuición coincide con la sugerencia del causalista, otras con la del bayesiano. Ante la pregunta: ¿qué es lo racional hacer

en cada caso?, encuentro que la intuición constituye una buena medida para evaluar esto. La presentación formal de la justificación de la racionalidad de nuestra acción, por tanto, podría apelar, en distintas circunstancias, a diferentes procedimientos. Unas, al bayesiano. Otras, al causalista. Quizás, en otras aún diferentes, a ninguno de ambos. Frente a esto se yergue algunas objeciones: ¿Por qué las razones causalistas valdrían en un caso, pero no en otro? (Lo mismo vale para las razones bayesianas.) Más aún: esta no es ninguna respuesta satisfactoria, porque lo que se espera de una teoría de la racionalidad es que provea un enfoque sistemático. En apoyo de esta idea, cito a Gibbard y Harper:

Quien se hallara libre de falacias, y sin embargo siguiera pensando que lo racional es tomar una sola caja, no hallará en el texto otros argumentos para ser persuadido. Si, además, creyera que en el caso original de Newcomb también se debe optar por una sola caja, que en el caso modificado de Newcomb debe optar por ambas, que Reoboam debe ser severo, que Salomón debe abstenerse de la mujer de su vecino, entonces tiene las intuiciones de un V-maximizador. Si piensa alguna de estas cosas, pero no todas ellas, debe todavía proveer una explicación sistemática de sus opciones. [Gibbard, A. y Harper, W., 1988, p. 372].[10]

(En la cita se recogen distintas soluciones a los problemas mencionados por los autores para hacer persuasiva la opción causalista.) Si lo que se busca, en efecto, es un único sustituto formal que sea compatible con todas nuestras intuiciones, no se lo tendrá. Al menos, ni el bayesiano ni el causal son satisfactorios en este sentido. Quizás el correcto sea aquél que recoja un mayor número de intuiciones de los casos relevantes. Pero, ¿para qué queremos un sustituto formal de CDP? ¿Por qué no conformarnos con nuestras intuiciones? Quizás sea ese el procedimiento sistemático buscado, que explique por qué formalizarlo a veces usando probabilidades condicionales y otras probabilidades contrafácticas. Claro: el procedimiento empleado, entonces, sería sistemático, pero no sistemático en su aspecto formal. Pero, de vuelta, ¿por qué no conformarnos con nuestras intuiciones? Después de todo, la idea de 'maximizar la utilidad esperada', por caso, si bien es más clara que la de 'es racionalmente obligatorio', pretende, además, ser extensionalmente equivalente a esta, y ser su sustituto formal. No es evidente ni lo uno ni lo otro. Parece que allí donde hay alguna intuición, si la intuición es intensa y definida, es todo lo que requerimos. Ninguna cuenta, después de todo, nos persuadirá de lo contrario. De suyo: hay casos en los que no hay una intuición clara. Allí sí sirve apelar a estos sustitutos formales, extensionalmente muy parecidos al predicado 'es racionalmente obligatorio', para despejar la incertidumbre. Pero no veo por qué ello deba revertir sobre los casos en los que nuestra intuición es clara. Quizás, podría argüirse, es necesario para evitar los errores a los que la intuición lleva. Veamos un caso de este estilo.

La así llamada 'paradoja de Allais' es un problema que es visto por algunos como dirigido contra la teoría de la decisión bayesiana. Esta paradoja involucra dos situaciones.[11] En la situación A, se nos insta a optar entre (a) recibir un millón de dólares con plena seguridad, o (b) una lotería. La lotería que se nos ofrece supone que existe 0.1 probabilidades de que ganemos 1 millón de dólares, 0.89 probabilidades de ganar 5 millones de dólares, y 0.01 probabilidades de no ganar nada. En la situación B se nos presenta una opción entre dos loterías. En la primera, (c) hay 0.1 probabilidades de ganar 5 millones y 0.9 de no ganar nada. La segunda (d) presenta 0.11 probabilidades de ganar 1 millón y 0.89 de salir sin nada. Muchos individuos aparentemente reflexivos y racionales tienden a optar por 'a' en A y 'c' en B. Sin embargo, esto contradice la teoría bayesiana. Veamos por qué. Veamos las utilidades de cada acto:

$u(a) = u(1M)$
$u(b) = 0.1\ u(5M) + 0.89\ u(1M) + 0.01\ u(0)$
$u(c) = 0.1\ u(5M) + 0.9\ u(0)$
$u(d) = 0.11\ u(1M) + 0.89\ u(0)$

Pero de aquí se sigue que:

$u(a) - u(b) = 0.11\ u(1M) - [0.1\ u(5M) + 0.01\ u(0)]$
$u(d) - u(c) = 0.11\ u(1M) - [0.1\ u(5M) + 0.01\ u(0)]$

Si se prefiere 'a' a 'b', entonces $u(a) > u(b)$, y por tanto $u(a) - u(b) > 0$. Pero entonces también ocurre que $u(d) > u(c)$, por lo que deberíamos preferir 'd' a 'c', de acuerdo a la teoría y contra la intuición de aquellos individuos presuntamente racionales y reflexivos. Su intuición parece ser la primera reacción de la mayoría.

Savage entiende que este problema no muestra lo errado de la teoría bayesiana, sino las confusiones de la intuición o del sentido común. Lo racional es elegir a y d, o b y c, y ninguna otra combinación. No hay diferencia con otras situaciones en las que sujetos por lo demás inteligentes yerran en calcular las probabilidades de una situación o en llevar adelante argumentos matemáticos complejos. Lo que se necesita es corregir sus juicios, no avalarlos. No es necesario entrar en el detalle de la solución de Savage aquí.

Mi impresión es que puede verse la paradoja como mostrando, una vez más, las limitaciones del dinero (por su utilidad marginal decreciente) para constituir un satisfactorio sucedáneo de la utilidad. Esto, que nadie discute, es el olvido que genera la perplejidad en este caso. En particular, para quien no es rico, la diferencia entre tener 5 millones con respecto a la de tener 1 millón no es, medida en utilidades, de 5 a 1. Para quien no es rico, serlo hace toda la diferencia. Ser rico con 5 millones no es tan diferente a serlo con 1 millón. Al menos, ambas son sustancialmente más distintas de

no ser rico. Esto explica la opción 'a' por sobre la 'b'. Qué elegir entre 'c' y 'd', por otra parte, no parece tan claro. Para que los '5M' signifiquen 5 veces '1M', debería ofrecer más que 5 veces más dinero. Debería, quizás, ofrecer treinta años más de vida y juventud. Y si este fuera el caso, la opción entre 'a' y 'b' no sería tan clara. O sí, al menos para mí si fuera a mí a quien ofrecen la lotería: elijo 'b', en este caso.

Las intuiciones, en este caso, no parecen tan fuertes. Pero quien tenga el impulso, luego de haber elegido 'a', de elegir 'c', puede que razone de la siguiente manera: es altísimamente improbable que termine rico. Tengo casi igual probabilidad a terminar rico con 5 millones a hacerlo con 1 millón. Siendo ese el caso, prefiero tener 5 millones a 1. Por tanto elijo 'c'.

Lo que revela esto, o lo que esto, más modestamente, insinúa, es que la presentación del problema no recoge toda la sutileza de las variantes que se les presentan a quien debe optar. Quien elige puede hacerlo basado en razones, y si lo hace (basándose en las razones correctas), su elección es racional. ¿Qué es una razón? La respuesta tradicional dice que es un compuesto de deseo y creencia. Frederic Schick,[12] sin embargo, piensa que hay algo más. Presenta, como aval de su tesis, una multitud de ejemplos. Uno de ellos, el de los médicos nazis de la S.S. encargados de 'seleccionar' quién iba a ser destinado a las cámaras de gas. En un primer momento, se negaban a hacerlo, pues veían el acto como una falta a sus deberes como médicos de cuidar la vida humana, pues, además, veían a los prisioneros como personas. Eso cambió. Los médicos de la S.S. pasan a ver su acto como el cumplimiento de su deber como nazis, como un acto de servicio a la patria, el asesinato como parte de 'la Solución Final', y a los prisioneros como 'enemigos de la raza'. ¿Qué cambió? Schick sostiene que no cambiaron las creencias ni los deseos. Los médicos siempre quisieron cumplir su deber como médicos y como nazis. Lo que cambió es el modo en cómo veían el acto en cuestión. Dejaron de verlo como una falta a los deberes médicos para pasar a verlo como un cumplimiento de sus deberes como nazis, y así con respecto a los demás tópicos. Lo que cambió fue su interpretación de los actos, y de un modo particular: ahora eran actos que querían llevar a cabo. "Elegir", dice Schick, "es llegar a querer esto *tal como lo interpretamos*".[13]

De igual manera, quien, enfrentado a la paradoja de Allais, elige 'a' y 'c', no necesariamente es irracional. No ven la opción entre 'a' y 'b' como la alternativa entre la posibilidad de ganar 5 veces más con un riesgo ligeramente mayor si se elige 'b'. La ven, como dije antes, como la alternativa entre ser rico con seguridad, si eligen 'a', y arriesgarse a no serlo, si eligen 'b'. La opción de 'c' sobre 'd' se justifica de modo similar: la ven como la negativa a asumir un riesgo muy menor de una recompensa mucho mayor, pues con amplia probabilidad terminarán pobres de todas maneras.[14]

Creo que la interpretación que hacemos de los posibles actos debe ser un factor a tener en cuenta, y creo que Schick acierta en este punto. Creo, sin embargo, que la visión debería ser refinada. Vemos los actos de muchas maneras, y lo que, muchas veces, hacemos, es sopesar esas visiones. Si vemos un acto como I y como II, y verlo como I nos insta a llevarlo a cabo mientras que verlo como II no, pero lo vemos más claramente como I que como II, entonces terminaremos realizándolo. Pero si, además, lo vemos también como III, y III nos insta a no realizarlo, aunque III sea menos 'vivaz' como visión que I, quizás terminemos no llevando a cabo el acto, pues II y III 'sumados' son más fuertes que I. No me ocuparé de cómo medir y comparar interpretaciones en lo que sigue. Tampoco intentaré profundizar aún más la defensa de esta visión de la racionalidad. Me basta con que haya quedado establecido que es una posibilidad abierta. Trataré de evaluar cómo sirve como guía de conducta en la práctica, y cómo, a partir de ella, puede entenderse el problema de Newcomb.

**Las mil caras de Newcomb**

El problema de Newcomb atañe a quienes están preocupados por entender qué hace de una decisión, una decisión racional. El problema, tal como es planteado por John Collins,[15] es el siguiente: hay dos cajas frente a uno. La primera es opaca; la segunda, transparente. Esta última contiene $1000. La opaca contiene $1000000 o nada. Uno tiene la posibilidad de tomar la caja opaca, o de tomar ambas (y ninguna otra posibilidad). Quien es el encargado de poner el millón o no poner nada es una Predictor de los procesos deliberativos humanos altamente fiable. Si el Predictor puso el millón, lo puso el día anterior a que uno esté frente a las cajas. Si el Predictor predijo que uno agarraría la caja opaca, puso en ella el millón. Si predijo que uno tomaría ambas cajas, no puso nada en la caja opaca. Uno tiene gran confianza en la eficacia del poder predictivo del Predictor. ¿Qué debe uno hacer?

Robert Nozick, quien hizo público el problema en "El problema de Newcomb y dos principios de elección"[16] lo plantea como un conflicto entre dos principios de decisión: el de utilidad esperada y el de dominancia.[17] Nozick cree que allí donde se da influencia de las acciones posibles sobre los estados del mundo (y las probabilidades condicionales de los estados sobre las acciones son todas iguales), y puede darse una acción dominante, hay que realizar la acción dominante. Si no hay acción dominante en estos casos, o si las acciones afectan qué estado del mundo se da, hay que realizar la acción que maximiza la utilidad esperada. El problema de Newcomb es una en el que no hay influencia causal de las acciones sobre los estados, aunque hay divergencia en las probabilidades condicionales y hay una acción dominante. Allí, Nozick recomienda hacer la acción dominante. Traza, para ello, un paralelo con algunos otros casos, que no difieren en lo sustancial del problema de Newcomb, y en los que parece claro que hay que hacer la acción dominante. No expondré, por ahora, ninguno de ellos.

En oposición a Nozick, Isaac Levi, en "Newcomb's Many Problems"[18] afirma que el problema de Newcomb no está descrito con suficiente detalle como para determinar qué es lo que debería hacerse en ese caso. Hay tres tipos de escenarios compatibles con esa situación:

*Caso 1*
**El Predictor predice que uno se llevará solo una caja y:**
Uno elige una caja: 900000
Uno elige ambas cajas: 100000
**El Predictor predice que uno se llevará ambas cajas y:**
Uno elige una caja: 10
Uno elige ambas cajas: 90

*Caso 2*
**El Predictor predice que uno se llevará solo una caja y:**
Uno elige una caja: 495045
Uno elige ambas cajas: 55005
**El Predictor predice que uno se llevará ambas cajas y:**
Uno elige una caja: 55005
Uno elige ambas cajas: 495045

*Caso 3*
**El Predictor predice que uno se llevará solo una caja y:**
Uno elige una caja: 90
Uno elige ambas cajas: 10
**El Predictor predice que uno se llevará ambas cajas y:**
Uno elige una caja: 100000
Uno elige ambas cajas: 900000

Si $H_1$ = hay un millón en la caja opaca, $H_2$ = no hay nada en la caja opaca, $A_1$ = uno toma solo la caja opaca, y $A_2$ = uno toma ambas cajas, entonces en el caso 1 $P(H_1/A_1)$ es alta y $P(H_2/A_2)$ es baja, en el caso 2 son ambas igualmente altas, y en el caso 3 $P(H_1/A_1)$ es baja y $P(H_2/A_2)$, alta. Si, como es habitual, se calcula la utilidad esperada de una acción usando las probabilidades condicionales de los estados del mundo dado los actos, entonces la utilidad esperada de la primera acción es $P(H_1/A_1)1000000 + P(H_2/A_1)0$, y la utilidad esperada de la segunda acción es $P(H_1/A_2)1001000 + P(H_2/A_2)1000$. Si se asume que la utilidad es lineal con respecto al dinero (que es lo que este tipo de ejemplos exige asumir), $A_1$ es la opción que maximiza la utilidad esperada solo en el caso 2. En los casos 1 y 3, la opción que maximiza la utilidad esperada es $A_2$. La conclusión de Levi es la siguiente:

*[F]rom the point of view of someone committed to using the principle of maximizing expected utility, no recommendation can be made concerning what X should do without filling in more details concerning X's predicament than Nozick has done. [Levi, I., 1975, p. 166].*[19]

Levi dedica el resto de ese artículo a mostrar que siempre que la situación esté suficientemente determinada como para que el principio de maximizar la utilidad esperada pueda ser aplicado, debe ser aplicado. Dejo de lado por el momento ese asunto para subrayar una idea de Levi que juzgo muy acertada: el problema de Newcomb está insuficientemente descrito, de modo tal que múltiples escenarios pueden 'instanciarlo', pueden ejemplificarlo.

Voy a presentar ahora una variante del problema en la que el Predictor es infalible. Ello significa no sólo que cada vez que predijo acertó, sino que cada vez que vaya a predecir acertará. ¿Qué es lo racional hacer en este caso? La respuesta unicajista parece ser la correcta aquí.[20] No importa aquí por qué acierta el Predictor.[21] Es decir: si uno elige tomar una caja, habrá en ella un millón. Si uno elige tomar ambas cajas, no habrá nada en la opaca. Si se razona de la siguiente manera: 'si tomo la caja opaca me llevo un millón. Pero el millón ya está ahí o no está. Si elijo la opaca y no está, no me llevo nada. Si opto por la opaca y está el millón, me pierdo los mil de la transparente. He de optar por tomar ambas cajas', se estará tomando una decisión en contra de los propios intereses, pues el millón no va a estar en la opaca. La idea es la siguiente: si se opta por la opaca, entonces habrá en ella un millón. Si se opta por ambas, en la opaca no habrá nada. Estos condicionales pueden ser tratados como equivalentes a condicionales clásicos.[22] Podemos negar que existan relaciones causales retrospectivas, o suspender el juicio con respecto a ellas, y la opción unicajista en este caso seguirá siendo la apropiada. Hay una correlación entre opciones y plata, y está en nosotros optar por la vía que nos llevará al millón. Es verdad que no podemos influir causalmente en la existencia del millón. Pero insistir en ese punto es desviar el foco de un modo de pensar el problema, el único que nos hará ricos con seguridad: hay una correlación entre opciones y plata en la caja que premia la elección de la caja opaca únicamente. Si creyéramos que la opción racional es la que maximiza la satisfacción de nuestros deseos (y hacernos ricos es uno de nuestros deseos), deberemos tomar solo la caja opaca, la opción preconizada por el principio CDP.[23] ¿Por qué, entonces Gibbard y Harper recomiendan elegir ambas cajas?[24]

Lo hacen porque esta es la acción que maximiza la utilidad U. A muchos les parece que aquí lo racional es optar por una sola caja, porque el sujeto sabe que será millonario si y sólo si toma una caja. Ellos siguen sosteniendo que lo racional es optar por ambas cajas. Supongamos que R = 'me haré millonario'. El sujeto sabe que $R \equiv A_1$.

Pero de esta proposición no se sigue que él se haría millonario si hiciera $A_1$, o que no se haría millonario si optara por $A_2$. Si el sujeto supiera con seguridad que él tomará solo una caja, sabrá con seguridad que se hará millonario. Pero también sabe con seguridad que el millón ya está en la caja opaca, por lo que se haría millonario aún cuando tomara ambas cajas. Por otro lado, si sabe que tomará ambas cajas, sabe que no se hará millonario, pues no hay nada en la caja opaca. Por tanto, aún cuando solo tomara la caja opaca, no se haría millonario.

Si, por otro lado, no sabe qué opción seguir, sabe que o tomará solo una caja y devendrá millonario, o tomara ambas, y no lo será. Pero de aquí no se sigue: (1) que si tomara solo una caja, sería millonario, ni (2) que si tomara ambas cajas no devendría millonario. De (1), el agente sabe que es verdadera si y sólo si hay un millón en la caja opaca; de (2), el agente sabe que es verdadera si y sólo si no hay un millón allí. Por tanto, si (1) se sigue de lo que el agente sabe, el agente puede concluir con certeza que la caja opaca contiene un millón; y si (2) se sigue de lo que él sabe, puede concluir que la caja opaca está vacía. Como el agente no sabe lo que va a hacer, no puede concluir nada con respecto al contenido de la caja: ni (1) ni (2) se siguen de lo que el agente sabe.

La opción racional en el problema de Newcomb depende de lo que el agente sepa. Y lo que sabe es que, no importa qué haya en la caja opaca, se llevará mil dólares más si elige ambas. Lo racional, entonces, es tomar ambas cajas.

Ellos ofrecen tres explicaciones para el hecho de que a muchos les parezca racional tomar solo una caja cuando el Predictor es infalible. (1) Las personas suelen albergar la tendencia a generar indicios del estado del mundo deseado, aunque sepan que el acto que genera el indicio no genera el estado. Tomar solo la caja opaca es ese acto que genera el indicio de que el mundo es tal que hay un millón en la caja opaca. (2) Una objeción que suelen recibir los bicajistas es la siguiente: "Si ustedes, bicajistas, son tan astutos, ¿por qué no son millonarios?". Llevado a la versión del problema de Newcomb en la que un amigo puede ver el contenido de la caja opaca y recomendarnos qué hacer, la conclusión parece absurda: habiendo visto que en la caja hay un millón, el sujeto debería abstenerse de tomar los mil de la caja transparente. Como el argumento prescribe esta conducta tan evidentemente irracional, debe haber algo mal en él. (3) La suscripción del argumento falaz que va de la premisa "o tomo una caja y devengo millonario, o tomo ambas y no lo hago", a la conclusión "si tomara una caja devendría millonario, y si tomara ambas, no".

Nozick y Levi, sin embargo, opinan de forma opuesta a Gibbard y Harper, y recomiendan tomar solo la caja opaca. Ninguno de ambos es unicajista con respecto al

problema original de Newcomb. Levi tiene algo que decir contra la posición general de Gibbard y Harper, algo sin lo cuál ella no se sostiene: los contrafácticos carecen de valor de verdad. Quienes no creen esto, o quienes no tenemos una opinión tan claramente formada como Levi al respecto, pero que creemos que, en efecto, la intuición indica que si el Predictor es infalible, hay que tomar solo una caja, quisiéramos decir alguna otra cosa. ¿Qué?

Creo que Gibbard y Harper tienen razón en este punto: de que sea verdad que uno se hará millonario si y sólo si opta por $A_1$ no se sigue que él se haría millonario si hiciera $A_1$. De todas formas, de que esta inferencia no sea lógicamente válida tampoco se sigue que sea falso que uno se haría millonario si hiciera $A_1$. Creo incluso que esta afirmación es verdadera en este escenario, que esta es la opción intuitiva, y que ninguna traducción de la afirmación a un lenguaje formal, por caso, con semántica de mundos posibles, será suficiente para hacerla falsa. ¿Se cae, entonces, en un impasse en este punto? Si no hay argumentos y solo que queda la intuición, parece que ese es el caso.[25]

Los autores conceden que si el sujeto supiera con seguridad que él tomará solo una caja, sabrá con seguridad que se hará millonario. Pero creen, además, que:
*If the subject knows for sure that he will take just the opaque box, then he knows for sure that the million dollars is in the opaque box, and so he knows for sure that he will be a millionaire. But since he knows for sure that the million dollars is already in the opaque box, he knows for sure that even if he were to take both boxes, he would be a millionaire. [Gibbard, A. y Harper, W., 1988, p. 370].*[26]

Creo que el paso de la primera afirmación a la última tiene algo extraño, de un modo análogo a cómo lo es pasar de que sea verdad que uno se hará millonario si y sólo si opta por $A_1$, a que sea verdad que uno se haría millonario si hiciera $A_1$. La primera afirmación es extensional; la segunda no.[27] Y si bien no toda inferencia de este tipo es inválida, tampoco es evidente. Alguien podría argüir que el 'decisor' no sabe con certeza que si tomara ambas cajas sería millonario, pues el mundo más cercano en el que hiciera esto no es este, sino otro en el que él es muy diferente, y el Predictor actuó de otra manera. Los causalistas no acordarán con esto. El impasse, nuevamente, parece irremediable. ¿Qué debería hacer el individuo enfrentado a este escenario, sin embargo?

Supongamos que este individuo ha leído el presente debate, y, como consecuencia de las dudas que ello le genera, no sabe con seguridad nada de lo que Gibbard y Harper afirman que sabe. Lo que sí sabe es que será millonario si y sólo si elige solo la caja opaca, lo que ninguno de los contendientes dialécticos pone en duda. Sabe entonces

que si elige solo la caja opaca, será millonario, y que si elige ambas cajas, no lo será. El quiere ser millonario. Si debemos basarnos, para actuar racionalmente, solo en lo que sabemos, la salida es evidente: hay que tomar solo la caja opaca. Si hay solo dos caminos a tomar, uno que nos lleva a conseguir lo que queremos y otro que no, lo racional es tomar el primero. El peso intuitivo de esta tesis es muy grande y, como creo haber insinuado,[28] los argumentos de los causalistas no alcanzan a rebatirla, pues el problema de Newcomb con Predictor infalible muestra, precisamente, un escenario de ese tipo.

Claro: el problema de Newcomb no incluía un Predictor infalible, sino sólo uno suficientemente fiable. Allí el Predictor es solo muy bueno. Más aún: ha fallado en algunas ocasiones. ¿Qué hacer en este escenario? No hay correlación infalible: puedo elegir solo la caja opaca y no llevarme nada, puede elegir ambas y llevarme un millón mil dólares. Más aún: ya ha habido gente que no se llevó nada, ya ha habido gente que se llevó el millón más los mil. El Predictor ya predijo, y dejó o no dejó el millón. Nada de lo que uno pueda hacer, ni siquiera forjar la intención sincera de que voy a tomar sólo una caja, podrá alterar este hecho.

Supongamos que sé que solo voy a ser sometido una vez a un problema del estilo del de Newcomb, y que sé que lo que decida en esa ocasión no afectará sustancialmente mi futuro en ningún sentido relevante (más allá de, eventualmente, perder mil dólares). La opción razonable, en este nuevo escenario, antes de hacer cuentas, parece ser tomar las dos cajas. Tomar las dos cajas parece ser la opción dominante: no importa cuál sea el escenario, haya un millón o no lo haya, tomando dos cajas nos quedamos siempre con más plata que si hubiéramos optado por sólo la caja opaca. En este escenario, la opción bicajista sí parece la razonable.[29] Nótese, sin embargo, que este no es el problema de Newcomb original, sino que hay nueva información: es el único problema del estilo al que voy a ser sometido en mi vida, y nada de lo que decida alterará sustancialmente mi futuro. Señalo esto porque el que ello no sea dicho, me parece, es lo que origina buena parte de las disputas. Digo: no que *esto* no sea dicho al plantear el problema, pues no forma parte del problema. Mi impresión, reitero, es que el problema de Newcomb está insuficientemente descrito, y que puede ser 'rellenado' con distinto tipo de nueva información, que de lugar a distintos escenarios. El que acabo de plantear es uno, y solo uno de ellos.

Ahora supongamos que de hecho se nos plantea el problema de Newcomb original, pero que ese caso va a ser el primero de una serie de diez problemas de Newcomb que se nos planteará a lo largo de una cierta cantidad de tiempo. Si no tenemos ninguna información al respecto, y por ello quiero decir que no hay ningún indicio de que ese pueda ser el caso (es decir, que sea el caso que seremos sometidos a diez problemas

de Newcomb), *prima facie* parece que la opción más razonable sigue siendo agarrar las dos cajas. ¿Cuán probable es en ese escenario que se nos planteen diez problemas de Newcomb, es decir, diez veces distintas -y sucesivas, claro- la situación conocida como 'el problema de Newcomb'? Parecen remotas. Al menos no son nada claras, y no hay nada que indique que ese escenario será más probable que otros varios futuros posibles en los que no se nos plantea ningún problema de Newcomb más. Lo seguro es que estamos frente a un problema de Newcomb, que el Predictor no es infalible, que ya ha fallado, y que ya puso o no puso el millón en la caja opaca. No hay ninguna razón de peso, ninguna razón relevante, ninguna posibilidad futura que sea sustancialmente más plausible que las demás, que nos haga reconsiderar nuestra elección bicajista. Conviene, entonces, utilizar, por caso, el principio de maximizar la utilidad del peor escenario posible, y asegurarnos al menos mil dólares.

Ahora imaginemos que sabemos que esa elección será la primera de diez que debamos realizar, la primera de diez veces que nos enfrentemos con casos de problemas de Newcomb. Supongamos que el Predictor tendrá en cada nuevo caso el dato de nuestras elecciones anteriores, y que esos datos influirán en sus decisiones. Ahora sí parece conveniente elegir una caja. Parece conveniente hacerlo siempre, al menos hasta el problema de Newcomb número 10. En cada uno de los nueve casos previos, sabemos que nuestra elección afectará la decisión siguiente del Predictor. Conviene, entonces, elegir en ellos una caja. Quizás la primera vez no nos llevemos nada, pero parece que sí nos llevaremos más plata al finalizar las nueve rondas primeras, pues el Predictor es muy bueno, y, en particular, es muy bueno en nuestro caso,[30] por tanto acertará al menos en un alto porcentaje de nuestras elecciones. Y, recordemos, lo que prediga está relacionado con lo que hayamos elegido en el pasado (asumiendo que esa relación puede glosarse así: si antes eligió p, ahora es más probable que elija p a que elija no-p -donde esta regla es claramente derrotable por una variopinta gama de consideraciones). Ahora bien: llegado al caso 10, ¿qué nos conviene hacer? Supongamos que el único dato que el Predictor tiene sobre nosotros es nuestro comportamiento previo. Más aún: supongamos que su decisión se basa únicamente en ese dato y en consideraciones generales acerca de la conducta humana.[31] El Predictor puede razonar de esta manera: 'antes eligió siempre una caja. Es difícil que ahora cambie'. O quizás pueda razonar así: 'antes eligió siempre una caja. Lo hizo así porque creía que así iba a lograr que yo pusiera más veces un millón. Pero esta es la última elección que va a afrontar, y ya no tiene que crear en mí ninguna impresión particular. Ahora va a elegir las dos cajas, porque esa es la opción dominante'. Cómo razone dependerá de qué tipo de 'consideraciones generales acerca de la conducta humana' tenga presentes. En cualquier de ambas situaciones, la opción dominante es tomar las dos cajas. No es claro, por otra parte, que la opción que maximice la utilidad esperada sea tomar solo una caja, tal como señaló Levi (el escenario general

está indeterminado en su sentido, también). Por tanto, conviene, en este último caso, tomar ambas cajas.

Consideremos una última variante del problema. Estamos frente a un problema de Newcomb, y se nos dice: 'esta puede ser el primero de diez problemas de Newcomb a los que se enfrentarán, siempre con el mismo Predictor, y en cada caso el Predictor tendrá el dato de sus elecciones previas. Pero quizás sea el único problema de Newcomb al que se enfrenten'. ¿Qué hacer en este caso? Suspendamos por un momento uno de los presupuestos de este tipo de problemas: que el dinero es linear con la utilidad. Qué hacer, entonces, dependerá mucho del valor asignado de los mil dólares, claro. Cuanto mayor sea éste, más racional será asegurárselos eligiendo las dos cajas. Pero supongamos que la utilidad de esos mil dólares no es excesivamente alta. Ahora el escenario de diez problemas de Newcomb no es tan improbable como antes. De hecho es una posibilidad 'destacada' del contexto de las demás, y 'destacada' no por nosotros, sino por quien nos plantea el problema. Que algo sea una posibilidad 'destacada', supone que su probabilidad es significativa, que no es mucho menos probable que la opción más probable, y que el individuo que decide la tiene presente, o podría y debería tenerla presente.[32] Si, entonces, el escenario de tener frente a nosotros diez problemas de Newcomb constituye una probabilidad muy destacada, la opción unicajista se hace, frente al primer problema de la serie, muy razonable. ¿Y si los problemas de Newcomb no son diez, sino dos?

No tengo una respuesta muy definida. Lo que quise ilustrar con esta serie de variantes del problema son varias cosas. Primero, que el problema, tal como suele planteárselo, es suficientemente impreciso como para que distintos detalles con los que se lo complete arrojen distintos resultados; más, incluso, de los que Levi se imaginaba, o al menos más de distinto orden. Segundo, que no siempre la opción bicajista es la recomendable. En particular, que hay muchos escenarios en los que la alternativa unicajista es la recomendable. Tercero, que en varios de estos escenarios no hay una opción intuitiva, y que para averiguar qué opción es la racional hay que hacer cuentas. ¿Qué cuentas? Digo: ¿qué sumatoria, la evidencial o la causalista? Eso, también, puede que dependa de los detalles del ejemplo. Como señalé en el primer capítulo, no veo necesario contar con un solo procedimiento de decisión para todo escenario. La intuición podría ser una mejor guía para la acción que muchos procedimientos, o conjuntos de procedimientos de elección.[33]

## Notas

[1] Una teoría normativa y mínima de la racionalidad no tiene por qué ser incompatible con una perspectiva evolucionista de la misma. De serlo, quizás, la teoría normativa en cuestión tendría varias cuestiones más a las que atenerse, y necesitaría más razones para avalar su viabilidad. Con esto, también, intento que se me exima de dar ese tipo de razones.

[2] Arló Cosa, H., "Racionalidad y teoría de la acción (1996). "¿Es la teoría evidencial de la decisión una teoría de la racionalidad mínima?", en *La Racionalidad: Su Poder y Sus Límites*, de Oscar Nudler, Paidós, Buenos Aires, Argentina, pp. 295-329.

[3] Arló Costa muestra cómo, basándonos solo en los tres principios estructurales, y prefiriendo que mañana no llueva a que mañana llueva, se llega a la antiintuitiva conclusión que es más racional preferir que mañana llueva o no llueva, cuando lo intuitivo es preferir que mañana no llueva (y no la disyunción citada). El ejemplo aparece en las páginas 314/5.

[4] Arló Costa llama al principio de racionalidad presentado al inicio del apartado, 'segundo principio de racionalidad'.

[5] Gibbard, A. y Harper, W. (1988). "Counterfactuals and two kinds of expected utility", en *Decision, Probability, and Utility*. Peter Gardenfors y Nils-Eric Sahlin, eds. C.U.P.

[6] Que no lo son es ampliamente aceptado. Hay pruebas de ello en ese artículo de Gibbard y Harper, así como en el artículo de Lewis, D. "Causal Decision Theory" (1988). *Decision, probability, and utility*. Peter Gardenfors y Nils-Eric Sahlin eds. C.U.P.

[7] El artículo de Arló Costa pretende probar que la teoría de la decisión bayesiana clásica no es psicológicamente neutral, y que por tanto no es un candidato apropiado (o tan bueno como sería de esperar) para una teoría mínima correcta de la racionalidad. Pero ese propósito no es el del presente trabajo.

[8] Este es un criterio metafilosófico o metodológico para dirimir disputas que goza de cierto prestigio. No es el único, y quizás no sea el mejor. Sí merece, de ello no hay mayor duda, tenerse en cuenta. Un defensor de prestigio de este criterio es Jason Stanley. Para un esboza de una defensa de él, puede verse la introducción que Stanley escribió para su libro *Knowledge and Practical Interests* (Stanley, J. (2005). Knowledge and Practical Interests. Oxford University Press).

[9] Son cuatro, con algunas variaciones, y están desperdigados a lo largo del artículo.

[10] Ibid., página 372. La traducción es mía. Se podría argumentar que el bayesiano no recomienda todo lo que los autores creen que recomienda, y esto es algo que se ha hecho. Como este no es el asunto principal de mi interés aquí, no ahondaré tampoco en él.

[11] Sigo, aquí, la exposición de la paradoja debida a Michael Resnik, planteada en su libro *Choices: an Introduction to Decision Theory* (Resnik, M. (1997). *Choices: an Introduction to Decision Theory*. University of Minnesota Press).

[12] Schick, F. (2000). *Hacer elecciones. Una Reconstrucción de la Teoría de la Decisión*, Gedisa, Barcelona, página 20.

[13] Ibid., página 27.

[14] Quizás la decisión racional no esté determinada solo por deseos y creencias e interpretaciones, sino, por caso, también por miedos y aprehensiones. La consideración de esta posibilidad, sin embargo, nos llevaría demasiado lejos del interés de este trabajo.

[15] Collins, J. (2005). "Newcomb's Problem", *Encyclopedia of Social and Behavioral Sciences*, Smelser, N. y Baltes, P., eds.

[16] En Nozick, R. (2004), *Puzzles Socráticos*, Cátedra, Colección Teorema. Originalmente, Nozick, R. (1969). "Newcomb's Problem and Two Principles of Choice". *Essays in Honor of Carl G. Hempel*. Rescher, N., ed. Dordrecht, Reídle, pp. 114-146.

[17] Por supuesto que hay situaciones en las que las recomendaciones de ambos principios son equivalentes. Ello es así, por ejemplo, cuando los estados del mundo relevantes son probabilísticamente independientes de las acciones, cuando para todo estado relevante H y toda acción A, P(H/A)=P(H).

[18] Levi, I. (1975), "Newcomb's Many Problems", Theory and Decision 6, pp. 161-175. Desarrolla también esta posición en Levi, I. (1982) "A note on Newcombmania", *The Journal of Philosophy,* 79(6), pp. 337-42. En menor medida (y de modo lateral), también puede verse algo de todo esto en otro artículo suyo, Levi, I. (1983). "The wrong box", *The Journal of Philosophy*, Vol. 80, No. 9, 534-542. Sep.

[19] "Newcomb's Many Problems", página 166.

[20] En esto acuerdan tanto un bicajista (alguien que cree que lo correcto en el problema original es tomar ambas cajas) como Robert Nozick (nuevamente, en "Newcomb's Problem and Two Principles of Choice"), como Isaac Levi (quien, sin ser unicajista, tampoco es bicajista. Levi expone esta postura, por caso, también en "Newcomb's Many Problems"). Gibbard y Harper, sin embargo, creen que, aun en este escenario, la solución correcta es optar por ambas cajas. (El lugar donde aparece la justificación de esta respuesta es el ya citado "Counterfactuals and Two Kinds of Expected Utility".) La posición unicajista, claro está, es la que recomienda tomar solo la caja opaca.

[21] No importa si acierta porque sabe, porque está causalmente relacionado con los fenómenos relevantes del modo apropiado, o por mero azar. Lo que hay, y supongamos que sabemos esto, es una correlación infalible (más que meramente fiable) entre elecciones de los electores y predicción del Predictor.

[22] Condicionales que expresan relaciones de condición necesaria y condición suficiente. No expresan relaciones causales de ningún tipo.

[23] Cuán plausible es escenario, cuán similar o diferente a nuestro mundo, no son temas que me parezcan relevantes aquí. Puede, en última instancia, suponerlo un escenario de ciencia ficción. Puede encontrarse una defensa de este tipo de planteos (de problemas de este tipo) para la filosofía en dos artículos de Peter Slezak: Slezak, P. (2006). "Demons, Deceivers and Liars: Newcomb's *Malin Genie*", *Theory and Decision*, Springer y Slezak, P. (2006). "Realizing Newcomb's Problem" (inédito).

[24] Como señalé en la nota 7, la posición de estos autores, Gibbard, A. y Harper, W.L., aparece en el artículo "Counterfactuals and Two Kinds of Expected Utility".

[25] Se podría, quizás, sostener que uno se haría millonario si hiciera $A_1$ porque en el mundo más cercano a este dónde se haga $A_1$ uno se hará millonario. Si ese mundo es este, uno se hará millonario porque habrá un millón en la caja. Si ese mundo no es este (porque no hay un millón en la caja), será uno en el que el Predictor haya puesto un millón en la caja. ¿Por qué? Porque si uno en este mundo toma ambas cajas, el mundo más cercano en el que toma solo la opaca es uno en el que uno es muy diferente a quien es aquí. En ese mundo, uno tendrá, quizás, la predisposición a tomar solo una caja, y el Predictor, con sus métodos de avanzada, quizás, habrá detectado esta disposición, y por tanto habrá puesto el millón en la caja.

No creo que sea muy fructífero explorar esta vía. David Lewis, por caso, podría replicar que el mundo más cercano en el que uno elige solo la caja opaca es uno en el que el Predictor ya puso un millón, y en el que, por caso, el cambio en nuestras disposiciones ocurrieron después de este evento. Lewis podría sostener que este es un mundo en el que el número y el tipo de cambios con respecto al actual son menores. Yo no lo veo así pero, insisto, me parecen asuntos elusivos, y no voy a adentrarme en ellos.

[26] "Counterfactuals and two kinds of expected utility", página 370.

[27] Pues incluye contrafácticos.

[28] Porque no creo haber probado o demostrado nada. Solo pretendí dejar la impresión de que los debates acerca de afirmaciones condicionales no extensionales, en particular si incluyen semánticas de mundos posibles, pueden prolongarse, si no indefinidamente, al menos sí mucho más de lo que a primera vista podría parecer.

[29] Por supuesto que alguien como Levi no estará de acuerdo con esto. Pero quizás aún él pueda acordar en que el peso de la intuición recae en este caso, aunque no de modo claro, en la opción bicajista.

[30] Por supuesto que a medida que se engrosa el número de casos de referencia con los que medir la calidad del Predictor, disminuye la necesidad de que el Predictor acierte en nuestros casos para que continúe siendo un muy buen Predictor, o un muy buen Predictor con al menos el mismo grado de eficacia que luego de fracasar en la predicción de alguna de nuestras acciones. Pero pensemos, entonces, para este caso, en un Predictor que es muy eficaz *también* con referencia a estos diez casos.

[31] Donde 'se basa' quiere decir 'está justificado por' y no meramente 'está causado relevantemente por' (otras cosas que no son justificaciones). Quiero dejar fuera de la determinación de su decisión cualquier otro mecanismo causal que no sean las razones que pueda manipular. Quiero excluir a un Predictor cuya fiabilidad se explica no únicamente por las razones que sopesa, sino por algún otro mecanismo que no involucre la evaluación de razones, y del que quizás el Predictor no sea conciente.

[32] Para una ampliación de qué signifique que un escenario constituya una posibilidad 'destacada', pueden verse los artículos de contextualistas como Keith De Rose y Stewart Cohen. Ellos apelan a esta noción para desestimar los planteos escépticos, a la vez que explican su persuasividad. Pueden verse, como ejemplo de ello, el artículo de Cohen llamado "Contextualism, Skepticism and the Structure of Rehaznos" (Cohen, S. (1999). "Contextualism, Skepticism and the Structure of Reasons", *Philosophical Perspectives*, 13; Epistemology, J. Tomberlin, ed. Oxford: Blackwell, pp. 57-89.), y el de De Rose, "Contextualism and Knowledge Attributions" (De Rose, K. (1999). "Contextualism and Knowledge Attributions". *Philosophy and Phenomenological Research*, 52/4, pp. 913-29).

[33] El tratar de mantener tantas intuiciones como podamos es la posición metodológica (o metafilosófica) de muchos. Por citar solo a uno de ellos, es la que defiende Jason Stanley en la introducción de su libro *Knowledge and Practical Interests* (Stanley, J. (2005). *Knowledge and Practical Interests*. Oxford University Press).

## Bibliografía

Arló Cosa, H. (1996). "Racionalidad y teoría de la acción. "¿Es la teoría evidencial de la decisión una teoría de la racionalidad mínima?", en La Racionalidad:

Su Poder y Sus Límites, de Oscar Nudler, ed. Editorial Paidós, Buenos Aires, Argentina, pp. 295-329.

Bara, B., Barsalou, L. y Bucciarelli, M., eds. (2005). Proceedings of the 27th Annual Conference of the Cognitive Science Society, Lawrence Erlbaum, Mahwah, NY.

Broncano, F. (1996). "Las dimensiones de la racionalidad". La racionalidad: su poder y sus límites, de Oscar Nudler, ed. Editorial Paidós, Buenos Aires, Argentina, pp. 29-65.

Burgess, S. (2004). "The Newcomb Problem: An Unqualified Solution". Synthese, 138.

Cohen, S. (1999). "Contextualism, Skepticism and the Structure of Reasons", Philosophical Perspectivas, 13; Epistemology, J. Tomberlin, ed. Oxford: Blackwell, pp. 57-89.

Collins, J. (2005). "'Newcomb's Problem". Encyclopedia of Social and Behavioral Sciences, Smelser, N y Baltes, P, eds.

De Rose, K. (1999). "Contextualism and Knowledge Attributions". Philosophy and Phenomenological Research, 52/4, pp. 913-29.

Elster, J. (1989). Juicios salomónicos, Gedisa, Barcelona.

Gardenfors, P. y Sahlin N., eds. (1988). Decision, probability, and utility, C.U.P.

Gibbard, A. y Harper, W. (1988). "Counterfactuals and two kinds of expected utility". Decision, Probability, and Utility. Peter Gardenfors y Nils-Eric Sahlin, eds. C.U.P.

Kripke, S. (1988). "Outline of a theory of truth", The Journal of Philosophy 72, pp. 53-81.

Levi, I. (1975), "Newcomb's Many Problems", Theory and Decision 6, pp. 161-175.

Levi, I. (1982) "A note on Newcombmania", The Journal of Philosophy 79(6), pp. 337-42.

Levi, I. (1983). "The wrong box", The Journal of Philosophy, Vol. 80, No. 9, 534-542. Sep.

Lewis, D. "Causal Decision Theory" (1988). Peter Gardenfors y Nils-Eric Sahlin eds. C.U.P.

Lewis, D. (1981). "Why Ain'cha Rich?". Nous 15: 377-80.

Matisen, S. y Wilson, G. (2003). "Newcomb's hidden regress". Theory and Decision 54, pp. 151-162.

Nozick, R. (1969). "Newcomb's Problem and Two Principles of Choice". Essays in Honor of Carl G. Hempel. Rescher, N, ed. Dordrecht, Reídle, pp. 114-146.

Nozick, R. (2004), Puzzles Socráticos, Cátedra, Colección Teorema.

Priest, G. (2002). "Rational dilemas", Analysis 62, pp. 11-16.

Resnik, M. (1997). Choices: an Introduction to Decision Theory. University of Minnesota Press.

Schick, F. (2000). Hacer elecciones. Una Reconstrucción de la Teoría de la Decisión, Gedisa, Barcelona.

Slezak, P. (2005). "Newcomb's problem as cognitive ilusion". Proceedings of the 27th Annual Conference of the Cognitive Science Society, Bara, B., Barsalou, L. y Bucciarelli, M., eds. Lawrence Erlbaum, Mahwah, NY, páginas 263-274.

Slezak, P. (2006). "Demons, Deceivers and Liars: Newcomb's *Malin Genie*". Theory and Decision, Springer.

Slezak, P. (2006). "'Realizing Newcomb's Problem" (inédito).

Sorensen, R. (1987). "Anti-expertise, instability, and racional choice". Australasian Journal of Philosophy 65, pp. 301-15).

Stanley, J. (2005). Knowledge and Practical Interests. Oxford University Press.

*Artículo recibido el 30/11/07*
*Aceptado para su publicación el 12/02/08*