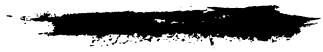
ÉTICA Y DECISIÓN RACIONAL



proyecto editorial

FILOSOFÍA

[hermeneia]

◆

directores

Manuel Maceiras Fafián
Juan Manuel Navarro Cordón
Ramón Rodríguez García

ÉTICA Y DECISIÓN RACIONAL

Gilberto Gutiérrez

EDITORIAL
SINTESIS

Diseño de cubierta
esther morcillo • fernando cabrera

© EDITORIAL SÍNTESIS, S. A.
Vallehermoso 34
28015 Madrid
Tel 91 593 20 98
http://www.sintesis.com

ISBN: 978-84-995838-4-6

Impreso en España - Printed in Spain

Índice

Prólogo

La tradición filosófica occidental, desde sus inicios socráticos, asienta el concepto de hombre sobre tres nociones fundamentales: racionalidad, *agencia* y moralidad. El hombre es concebido como un agente racional, capaz de elegir y obrar bien. Su facultad racional lo capacita para juzgar y discernir lo verdadero de lo falso en el orden del conocimiento pero, sobre todo, para elegir y procurarse lo bueno –la felicidad, el provecho, etc.– en el orden de la acción. Lo decisivo no es tan sólo que el hombre sea capaz de actuar según la razón sin más sino que, como reconoce ya Aristóteles, da por supuesto que ha de actuar según la *recta* razón. En términos literales: por la razón ajustada a un canon, norma o regla. Lo moral –*honestum, justum, rectum, decus,* etc.– es el canon de la razón práctica. Por eso la moralidad es, no sólo pero inevitablemente, *prescriptiva,* es *officium:* lo que *hay que* hacer. Y por ello es, no sólo pero también, *restrictiva:* no se *puede* hacer cuanto se *quiera.*

Por prescribir y restringir ciertas formas de actuar la racionalidad *moral* presenta una relación problemática con la racionalidad *tout court,* de manera muy específica en las relaciones del agente con otros agentes. La moral afirma que no es *justo* obrar con los demás según parezca conveniente al propio interés. Pero aunque parece evidente que las restricciones que la justicia impone a los deseos o a los intereses del agente benefician a los *destinatarios* de los actos justos, no está claro que promuevan siempre el bien propio de quien los ejecuta. Si la ética es una reflexión filosófica sobre la moralidad, su historia sería el intento ininterrumpido de resolver en principio el conflicto entre el interés racional y las exigencias contenidas en las intuiciones, los sentimientos y los juicios morales más reflexivos y ponderados; entre las demandas antagónicas de lo bueno y lo justo, de la virtud y la felicidad, del interés y el deber, de la eficacia y la igualdad, de los derechos individuales y el interés general,

etc. La historia, en suma, del propósito de hallar la respuesta definitiva a la pregunta –tal vez capciosa pero no por ello menos apremiante– *¿por qué actuar moralmente?*

Para ello siempre ha sido inevitable explorar la noción misma de racionalidad práctica, de agente –o, más genéricamente, de *agencia*– racional. La reflexión filosófico moral nunca ha podido desentenderse de los análisis de la acción racional llevados a cabo por otras ramas de la filosofía –por ejemplo, la filosofía de la acción o de la mente– o de las ciencias humanas o *morales*. En tiempos muy recientes se ha producido una fructífera convergencia entre determinados desarrollos de la teoría ética y los de la considerada como teoría *estándar* de la racionalidad práctica, que ha permitido plantear con inusitada claridad lo que implicaría responder a la pregunta antes mencionada. La historia y las razones de esta aproximación han hecho concebir la esperanza de obtener una solución final en los propios términos de la teoría estándar, mostrando que obrar moralmente *interesa*, dando por supuesta la relación analítica entre racionalidad práctica, o *agencia* racional, e interés.

El concepto de *agencia* está en el origen de la discutida distinción epistemológica entre ciencias *naturales* y *morales* –sociales, humanas, etc.– y ésta a su vez presupone la que el sentido común establece entre acontecimientos y acciones. Lo específico de las acciones, a diferencia de los meros acontecimientos naturales, es su carácter significativo, cuya explicación incluso científica parece exigir una metodología interpretativa inevitablemente referida a las intenciones y propósitos de agentes.

Por sus características específicas, la *agencia* humana sólo resulta inteligible si se postula con carácter general, que es el resultado de decisiones inteligentes. De ahí la relevancia metodológica de los modelos de racionalidad para entender, explicar e incluso justificar las acciones. Por razones convergentes tanto las ciencias morales como la filosofía moral han de hacer explícito un determinado modelo de racionalidad práctica o de agente racional. Pero un serio problema lo plantea el hecho de que el concepto de racionalidad presenta a un tiempo características *normativas* y *positivas*. Los modelos positivos –*explicativos*– de racionalidad elaborados por las ciencias morales no son normativamente inocentes: contienen presuposiciones normativas no siempre explícitas; a su vez, los modelos normativos propuestos por las teorías éticas se apoyan en supuestos de hecho (psicológicos, ontológicos, etc.). Es imposible elaborar una teoría ética al margen de toda teoría sobre la naturaleza humana e incluso de la realidad en general.

El concepto de *agencia* implica en diverso grado nociones como temporalidad, libertad, voluntad, intención, preferencia, deliberación, elección, decisión, etc. Hay razones metodológicas para centrar el análisis en el complejo integrado por las de preferencia y elección/decisión. La contemporánea *teoría de la elección racional* –denominación genérica que incluye las teorías de la utilidad, de la decisión, de los juegos, de la negociación y para muchos hasta la de la elección pública– es el resultado de un conjunto de desarrollos históricos y sistemáticos que han puesto a punto un modelo de conducta racional particularmente poderoso y preciso cuyos antecedentes se encuentran en la teoría económica. Demuestra que actuar racionalmente equivale a maximizar la utilidad esperada de las consecuencias de las decisiones. Permite así plantear con una claridad y un rigor extremos los problemas, dilemas y paradojas que suscita la relación entre moralidad y racionalidad en los dos entornos básicos de la decisión racional: el individual o paramétrico y, sobre todo, el colectivo o estratégico.

Los entornos *paramétricos* son aquellos en los que el agente adopta su decisión con el propósito de obtener la máxima utilidad de los acontecimientos o estados de cosas futuros, cuyas probabilidades considera como dadas y causalmente independientes de su propia decisión. Es el ámbito de las teorías de la utilidad y de la decisión; de la racionalidad individual y de la prudencia; comprende, por ejemplo, el conjunto de las decisiones que tomaría Robinson en su isla desierta antes de la aparición de Viernes. Resulta problemático afirmar la existencia de obligaciones morales –¿para consigo mismo? ¿con la humanidad?– en este ámbito, dado que en él no existe, estrictamente hablando, conflicto entre obligación e interés. Para Robinson toda obligación sería racionalmente interesada, aunque sin duda se le seguirían planteando conflictos *intrapersonales*, debidos a factores tales como su imperfecta racionalidad, su flaqueza de voluntad o la incertidumbre que afecta a su conocimiento, al mayor o menor peso que conceda a sus intereses futuros en relación con los presentes, etc.

En los entornos colectivos o *estratégicos* interviene más de un agente, cada uno de los cuales adopta su decisión en función de la decisión que espera que adopte el otro, el cual a su vez adoptará la suya en función de sus propias expectativas respecto de la decisión que tomará el primero, y así sucesivamente en lo que podría parecer un *regressus in infinitum*. A diferencia del entorno paramétrico, en este ámbito las decisiones y las probabilidades asociadas son interdependientes. Pero lo más problemático es que en él las utilidades de los agentes no son necesariamente coincidentes: los conflictos interpersonales son inevitables.

Decidir en situaciones de este tipo conforme a los criterios normativos de la racionalidad paramétrica –maximizar directamente la utilidad esperada– conduce a las conocidas paradojas y dilemas de la racionalidad colectiva, tales como el *dilema del prisionero*, la tragedia de los bienes comunales, la tentación del parásito o del *free rider* que se benefician de las contribuciones ajenas, etc.

Gran parte de la filosofía moral contemporánea –muy en particular la representada por la obra de John Rawls, John Harsanyi y David Gauthier– centra su interés en la elaboración de una teoría moral que, sin alterar radicalmente los supuestos de la teoría estándar de la racionalidad representada por la elección racional, dé cuenta satisfactoria de las exigencias éticas de justicia y equidad que inevitablemente se plantean en el ámbito de las relaciones interpersonales. Los resultados, sin embargo, no parecen confirmar sus propósitos. Y cabe pensar que son los propios supuestos los que excluyen toda posible solución.

Entre estos supuestos se cuenta la específica interpretación que la teoría de la elección racional hace de la naturaleza de las razones que son capaces de mover a actuar. La teoría parte de postulados aparentemente irrenunciables como el del "conocimiento compartido de la racionalidad", que estipula agentes racionales "transparentes" o al menos "translúcidos", pero en todo caso movidos únicamente por razones interesadas, prudenciales o *forward-looking*. Pero estos supuestos podrían no sólo ser internamente incoherentes sino incluso hacer racionalmente imposible aplicar los criterios normativos de cooperación en los términos estrictos de la Teoría de Juegos. Las restricciones que la moralidad impone al interés racional así entendido no podrían entonces explicarse ni justificarse apelando a ese mismo interés. La conclusión apunta a la necesidad de revisar los presupuestos psicológico-filosóficos del modelo de la teoría estándar para comprender mejor la función específica que desempeñan en la decisión racional, junto a las preferencias y las utilidades, las razones *back-* o *inward-looking*, en última instancia no-prudenciales, que definen en nombre de principios y valores la actuación constitutiva de la moralidad.

1
Las ciencias morales y la moral

1.1. Acontecimientos y acciones

Una de las distinciones más espontáneas y naturales que hace el hombre en el mundo de la experiencia es la que separa lo que *alguien hace* de lo que, simplemente, *acontece*. Por una parte estarían los *acontecimientos* –fenómenos que se agotan en la exterioridad de su condición de meros objetos– y, por otra, las *acciones* –aquellos otros que presuponen la dimensión interior propia de un sujeto–. Esta dicotomía está profundamente arraigada en la actitud natural del hombre ante el mundo y se impone a la consciencia que tiene de sí mismo. Cuando dirige la mirada a su interior le resulta imposible no separar de forma nítida lo que *él hace* de lo que *le pasa*. La filosofía y las ciencias de la conducta se guardan de considerar la distinción tan tajante como aparenta serlo, pero ello no impide darla por supuesta, al menos como hipótesis metodológica, no sólo en la vida cotidiana, sino en las propias ciencias tradicionalmente llamadas *morales* –el amplio conjunto que incluye, entre otras, la Psicología, la Sociología, la Antropología, la Ciencia Política, la Economía o la Historia– y en diversos sistemas filosóficos o éticos. Las distinciones que hacen Epicteto (1991: 3 [I, § 1]) entre lo que depende y lo que no depende de uno, Sartre entre *être en soi* y *être pour soi* (1984: 32-36; 113-118) o Weber (1964: 20-21) entre acciones significativas y meras reacciones comparten, desde perspectivas y con propósitos muy diversos, el mismo supuesto esencial.

Tomando pie de determinadas características observables de los fenómenos de la experiencia se conjetura con diversos grados

de probabilidad que algunos de ellos responden a la intención o el propósito de alguien, otros no y otros, en fin, resultan dudosos. Para decidir en favor de una de estas interpretaciones se aplican en principio los mismos criterios que en cualquier inferencia inductiva o abductiva, término este último con el que el filósofo norteamericano Charles S. Peirce (1839-1914) designaba el tipo de razonamiento que proporciona una hipótesis conjetural para explicar un conjunto dado de hechos, cuyo valor probatorio estricto depende sin embargo de la prueba experimental, y que guarda notables analogías con el razonamiento del detective que a partir de los indicios descubre al criminal (Eco, U. y Sebeok, T., 1989).

En la cultura occidental contemporánea nadie duda de que un terremoto sea un mero acontecimiento físico literalmente carente de sentido, ni de que una danza sea una actuación significativa; pero, dependiendo del previo conocimiento que se tenga de una persona, se duda entre interpretar cierto peculiar parpadeo como un guiño intencionado o como un tic nervioso. El *mismo* conjunto físico de incisiones en las paredes de piedra de un antiguo valle glaciar puede ser plausiblemente interpretado, bien como el mero efecto del desplazamiento de la morrena, o bien como una inscripción –un texto–; en este caso, aun desconociendo *qué* significa, se conjetura *que* significa algo. Así ocurrió con la piedra de Rosetta o con la escritura micénica conocida como *Linear B* antes de ser descifradas; sigue pasando con *Linear A* y con las inscripciones etruscas e ibéricas que no han podido serlo; y, sin ir tan lejos, forma parte de la experiencia diaria presuponer *que* los titulares de la prensa extranjera en lengua desconocida dicen algo, aun sin saber exactamente *qué*.

Los datos de la antropología, de la psicología evolutiva y de la historia de la ciencia sugieren que el ejercicio controlado de esta tendencia a distinguir entre acciones y acontecimientos se aprende a lo largo de una lenta y ardua evolución de la especie y del individuo humanos. En las fases tempranas de la filogénesis e incluso de la ontogénesis se tiende más bien a atribuir intencionalidad y sentido no sólo a lo que a todas luces es una conducta humana, sino incluso a aquello que no pasa de ser un mero acontecimiento sometido al juego ciego de las fuerzas naturales. Personificar las fuerzas de la naturaleza o considerarla animada

de intenciones amistosas u hostiles es una actitud mucho más natural que interpretar los fenómenos naturales –y más aún los comportamientos humanos– como puros fenómenos físicos. Prueba de ello son esos atavismos latentes que se cultivan con cierto despego irónico al "creer" en gafes, en rachas de buena o mala suerte o en la validez de la *ley de Murphy* que "explica" por qué las cosas tienden a salir mal (Bloch, A., 1997: 41-48).

Pero lo cierto es que el desarrollo de la razón, entendido como avance histórico del conocimiento filosófico y científico, se ha constituido como un proceso de "des-animación" y "des-humanización" de la naturaleza cuya propia dinámica tendería a la definitiva "naturalización" incluso de la razón y de la *agencia* humanas. La metodología del conductismo clásico obligaba a atenerse a los datos observables y mensurables y excluir toda referencia a lo mental en el estudio de la conducta humana; el objetivo final del materialismo fisicalista era reducir a una única ciencia unificada –la física– todas las ciencias empíricas, incluidas las biológicas, humanas y sociales. Sin entrar en otras consideraciones, lo cierto es que sólo se aprende a ver las acciones humanas en clave fisicalista tras el duro entrenamiento en una artificiosa *Verfremdung* ante la interpretación natural y espontánea de los datos de la experiencia: se cuenta la vieja anécdota de dos conductistas que se encuentran y uno de ellos pregunta al otro "Te veo muy bien. ¿Cómo estoy yo?" (Sen, A., 1986: 78). Por todas estas razones las ciencias morales se enfrentan a un complejo dilema a la hora de definir su *status* epistemológico.

Por una parte, en tanto que *ciencias* que aspiran a explicar la conducta, parecen obligadas a suponerla al menos tan determinada y previsible como cualquier otro fenómeno natural, de modo que el punto de vista interno y subjetivo tiene que subsumirse necesariamente en la perspectiva objetiva y *externa* cuya versión radical representa el conductismo. Los deseos y las creencias de los agentes tienen causas y efectos dentro de un orden natural (cuasi)determinístico gracias al cual aquéllos pueden conseguir lo que quieren mediante sus propias decisiones, aun imaginándolas libres. De hecho, un supuesto ampliamente compartido por las ciencias morales desde la Ilustración –Mandeville, Smith, Marx, Nietzsche, Freud– es que los agentes desconocen las verdaderas

causas y los efectos reales de lo que hacen. Pero este supuesto pugna con el punto de vista *interno* que el agente, en cuanto tal, inevitablemente adopta sobre su propia acción. En la medida en que considera que lo que *él haga* dependerá de la decisión que *él tome* en función de determinadas razones, parece imposible explicar su propia acción como el caso particular de aplicación de una "ley de cobertura" universal y necesaria, al modo de un eclipse o de una reacción química; y en esa precisa medida su acción resulta intrínsecamente impredecible incluso para él mismo. Este peculiar contraste conduce al otro cuerno del dilema.

En tanto que ciencias *morales* –sociales, humanas, de la conducta– su propio objeto sólo les es dado gracias a esa distinción natural entre hacer y acontecer, entre acontecimientos y acciones, la cual presupone una referencia intrínseca al punto de vista interno de un agente. Es dudoso que sea siquiera posible identificar el *explanandum* de las ciencias morales desde una perspectiva puramente externa. Incluso el punto de vista de ese espectador externo y cualificado que es el científico social presupone, al menos como hipótesis o conjetura, el punto de vista interno del agente para poder simplemente "recortar", del continuo de la experiencia, aquellos segmentos que pueden ser descritos como acciones. De hecho es así como el oyente identifica en la corriente sonora emitida por el hablante los conjuntos significativos que son los fonemas, las palabras y las oraciones bien construidas portadoras de sentido, o como interpreta el lector las manchas de tinta en un papel como texto significativo. No extraña, por tanto, que la tarea del científico social al describir y explicar la conducta muestre analogías estructurales con la del traductor de una lengua ajena. Identificar y describir como acción un segmento de comportamiento observable implica imputarle sentido, al menos como conjetura. Presupone que las acciones son *inteligibles* en la proporción exacta en que son efectos de un agente *inteligente*. Como se verá, a esta peculiar condición del objeto de las ciencias morales responde la necesidad de elaborar un tipo o modelo ideal de acción racional –de racionalidad práctica, en definitiva– como único procedimiento viable para descifrar, entender y explicar la conducta humana.

El nudo del dilema radica precisamente en esa excepcional singularidad de la agencia humana. El hombre y, más específica-

mente, la acción humana, es la única entidad natural que puede contemplarse desde esas dos perspectivas contrapuestas y tal vez irreductibles. Una de ellas, *externa,* corresponde al punto de vista del *espectador* que considera las acciones ajenas o incluso propias como algo dado, *ya hecho.* La otra es la ineludible perspectiva *interna* en la que se sitúa el agente que considera sus propias acciones no como algo dado sino aún *por hacer.* Este dualismo natural y en apariencia insoslayable fundamenta la diferencia metodológica y epistemológica entre las ciencias naturales y morales, y el muy diferente respecto bajo el que las ciencias morales y la ética consideran la agencia humana. La acción humana –que, sin dejar de ser un acontecimiento físico, es también producto de una deliberación y una decisión inteligentes– evoca la glándula pineal cartesiana, punto de tangencia de dos ámbitos de realidad y, en consecuencia, de dos universos de discurso.

1.2. Las ciencias morales

La fascinación del dualismo cartesiano fue uno de los factores que impidieron durante largo tiempo que las ciencias morales desarrollaran una metodología específica comparable a la que tan buenos resultados había dado en las ciencias naturales. Esto no sólo explica su retraso histórico, sino incluso la atracción inicial que, como reacción, ejerció sobre ellas el modelo de las ciencias naturales. Hume es un buen ejemplo del modo como la Ilustración inglesa se plantea la necesidad de abrir al conocimiento científico el universo de las acciones humanas, situándose así entre los precursores del enfoque naturalista de la conducta humana. Su propuesta, en el *Tratado de la Naturaleza Humana,* de "introducir el método experimental de razonamiento en las cuestiones morales", responde a su convicción de que todo conocimiento científico es producto de una misma actividad cognitiva determinada por la específica naturaleza del hombre. El conocimiento *de* esa naturaleza, que sólo puede alcanzarse mediante "la experiencia y la observación", es por tanto "el único fundamento sólido" de toda ciencia (Hume, D., 1992: 37 [xx]). No extraña, pues, que la aplicación de la "filosofía experimental" a las cuestiones mora-

les haya debido esperar al desarrollo de su aplicación a las cuestiones naturales, que Hume creía básicamente logrado con la obra de Newton.

Con todo, no era fácil para las ciencias morales seguir el paso que marcaban las naturales. Ni Hume ni ninguno de los grandes filósofos y científicos morales desconocieron que explicar la conducta humana añadía dificultades específicas a las que planteaba la explicación de los fenómenos de la naturaleza. Para algunos –por ejemplo, Mill– esta dificultad añadida no justificaba un *backward state* que podría remediarse aplicándoles los métodos de la ciencia física debidamente extendidos y generalizados. En el libro VI de su *Sistema de la lógica*, dedicado a la lógica de las ciencias morales, afirma Mill que "todo lo que una obra como ésta puede hacer por [dicha lógica] ha sido, o ha debido ser hecho en los cinco libros anteriores [...] puesto que los métodos de investigación aplicables a la ciencia moral y social deben haber sido ya descriptos (sic) en la medida en que he podido lograr numerar y caracterizar los de la ciencia en general" (Mill, J. S., 1917: 837). El objeto de estudio científico de la naturaleza humana son las leyes de formación del *carácter* (íd.: 871). Si se llama Psicología a la ciencia de las leyes elementales del espíritu *(mind)* bien podría denominarse *Etología* "a la ciencia ulterior que determine el género del carácter que producirá, según estas leyes generales, un conjunto cualquiera de circunstancias físicas y morales" (íd.: 876). Sería una ciencia exacta de la naturaleza humana porque las verdades que la constituyen no son generalizaciones aproximativas como las leyes empíricas que de ellas dependen, sino leyes verdaderas; pero como ocurre con todos los fenómenos complejos sus proposiciones no son exactas sino simplemente hipotéticas y afirman tendencias, no hechos. La Psicología se basa en la experiencia y la observación, mientras que la Etología es deductiva: los principios de ésta son los axiomata media de aquélla (íd.: 877).

Mill considera que "quizás la etimología haría este término [etología] aplicable a la ciencia entera de nuestra naturaleza mental y moral" (íd.: 876). El propio Aristóteles, cuando distinguía dos clases de virtud, la *dianoética* o intelectual y la *ética* o del carácter, atribuía el origen de esta última a la práctica de la costumbre *(ex éthous)*, y creía que hasta su nombre mismo *(eethiké)*

procedía de "una pequeña modificación" –el cambio de la *e psilón* inicial de *éthos* (costumbre) por la *eta* de *eethos* (carácter) (1959: 19 [1103a14-18]). Cicerón (1950, 1) acuña el término *moralis* para traducir *eethiké* porque *mos* (costumbre) traduce el griego *eethos*. En última instancia lo cierto parece ser que una misma raíz indoeuropea: **seswodh-*, **swedh-* se encuentra en el origen, tanto del latín *suesco* –del que proceden el *consuetudo, costumbre* y *coutume*– como de los términos griegos *éthos* (costumbre) y *éethos* (carácter).

Para otros, como Dilthey (1986: 40-41) o Weber, exigía por el contrario adoptar una metodología peculiar adecuada al carácter esencialmente significativo e intencional de la acción humana. Este último, por ejemplo, propone escapar al dilema mediante una *verstehende Soziologie* capaz de entender e interpretar la acción social como condición previa para explicarla causalmente (Weber, M., 1964: 5): esta sociología comprensiva no renunciaría al canon nomológico-deductivo de la explicación propia de las ciencias estrictas cuyo modelo es la física, pero tampoco prescindiría del criterio heurístico –el sentido, la intención, las razones– que permite literalmente identificar aquello mismo que ha se ser explicado: la acción.

Bajo distintas formulaciones la cuestión de fondo reaparece en distintos *loci* del debate filosófico de las últimas décadas. De una u otra forma está implicada en el análisis de las relaciones entre la mente y el cuerpo y, por extensión, en la definición del concepto de acción racional y de la función que desempeñan en ella las razones y las causas. La solución que se le dé condiciona parcialmente incluso la opción entre distintos modelos de fundamentación –por ejemplo, consecuencialista o deontologista– de la acción moral. En el propio campo de la epistemología de las ciencias humanas, un siglo largo de debates en torno a la distinción entre *Naturwissenschaften* y *Geisteswissenschaften*, entre explicar –*erklären*– y comprender –*verstehen*– no han bastado para fijar de forma definitiva el *status* metodológico de las ciencias morales frente a las naturales (Wright, G. v., 1979; Manninen, J., 1980). La verdadera naturaleza del problema se pone de manifiesto cuando se examinan los tipos o modelos ideales que las ciencias morales elaboran para conceptualizar la acción humana.

1.3. Modelos de racionalidad

El recurso a modelos ideales no es, por supuesto, privativo de las ciencias de esta clase. Toda teoría científica puede en principio servirse de modelos de diverso tipo que cumplen variadas e importantes funciones gracias a su isomorfismo, a las analogías formales o materiales que guardan con aquello de lo que son modelos. En los sistemas puramente formales –lógicos o matemáticos– compuestos de conjuntos de axiomas y de sus consecuencias deductivas –teoremas– los modelos representan conjuntos de proposiciones que los satisfacen y que constituyen una de sus posibles interpretaciones. En tales sistemas no existen definiciones de los términos aparte de los propios axiomas en los que aparecen: los modelos formales que los contienen son estrictamente autocontenidos y carentes *per se* de toda proyección empírica.

Otros modelos, en cambio, se caracterizan por implicar algún tipo de referencia a la realidad en forma de idealizaciones, simplificaciones o incluso falsificaciones de esa misma realidad por conveniencias explicativas. En la mecánica de fluidos se establece la ley de los gases perfectos o ideales gracias a la ficción del vacío absoluto; en los modelos estadísticos se "suavizan" o regularizan las curvas representativas de los valores reales de dispersión, etc. En un sentido bastante más laxo, las ciencias sociales han adoptado determinadas perspectivas idealizadas que simplifican a efectos explicativos la complejidad humana y la reducen a una única dimensión, ofreciendo un variado repertorio de *tipos* de hombre –*sapiens, sociologicus, ludens, faber* o *economicus*– o incluso de sociedades, baste recordar las tipologías sociales acuñadas en conceptos como los de solidaridad *orgánica* y *mecánica* (Durkheim), *Gemeinschaft* y *Gesellschaft* (Tönnies), *status* y *contract* (Maine), *folk society* (Redfield), etc. Así, por ejemplo, para cierta sociología el *homo sociologicus* es básicamente un seguidor de reglas, un actor o *role player* que ajusta sus decisiones a las expectativas normativas del papel social que le ha sido asignado, mientras que la economía neoclásica ve al *homo economicus* como un individuo movido sólo por su propio interés racional, que decide únicamente en función de la utilidad máxima que puede esperar de sus acciones.

Por último, la explicación científica se ha servido con frecuencia de determinadas entidades naturales o artificiales poseedoras de ciertas características que reproducen otras características presentes en aquellas cosas de las que son modelos: es el caso de las réplicas, los modelos a escala, los diagramas de flujo, las simulaciones, etc.; es en este sentido lato como los relojes sirvieron de modelo del universo a los mecanicistas del siglo XVIII, o los seres vivos a los sociólogos organicistas del XIX o a los funcionalistas del XX.

Todas las acepciones antes reseñadas del concepto de "modelo" coinciden además en poseer un declarado carácter descriptivo, positivo o teórico. En tanto que constructo artificial todo modelo posee una dimensión sintáctica que se agota en su coherencia formal e interna, como típicamente ocurre en los modelos formales autocontenidos del primer tipo. Pero los modelos que presentan analogías materiales con aquello de lo que son modelos están diseñados como instrumentos del conocimiento objetivo, y para cumplir esa función científica –descriptiva, explicativa, predictiva e incluso hermenéutica– no pueden ser autocontenidos ni estar vacíos de contenido. Ni sus conceptos primitivos ni las premisas que los contienen pueden ser simplemente estipulados. Si el valor epistémico o heurístico de un modelo depende de su adecuación o ajuste –su isomorfismo– a la realidad, ha de incorporar necesariamente una hipótesis que permita su interpretación sustantiva y material. En este caso la dirección del ajuste va del modelo a la realidad, y puede hablarse de modelos más o menos realistas o plausibles, dotados de mayor o menor potencial explicativo, etc.

La elección de un determinado modelo condiciona además las estrategias de investigación: si se acepta el *homo economicus* como tipo ideal de la racionalidad práctica, para explicar una determinada conducta en primer lugar habría que partir del supuesto de que es egoísta; si hay evidencia empírica de que no lo es, se la supone racional en un sentido menos fuerte; y sólo si tampoco lo es habría que admitir que, aunque irracional es al menos intencional (Elster, J., 1987: 22). Pero esta misma gradación de supuestos requiere una interpretación muy diferente del concepto de modelo, en la que la dirección del ajuste se invierte,

y en vez de ser el modelo un calco estilizado de la realidad, es la realidad la que debe ajustarse al modelo. Los modelos *normativos* establecen cierto canon o criterio de lo que es un *x* al que han de adecuarse los hechos para ser considerados, o contar como, un *x*. El problema con el modelo de acción o de agente racional es que parece imposible soslayar el carácter normativo del concepto mismo de racionalidad. Y no sólo en el sentido relativamente trivial de establecer un criterio para identificar ciertas acciones como racionales desde el punto de vista del espectador, sino en el más problemático de ofrecer criterios de actuación que el agente debería adoptar para actuar racionalmente.

En efecto, existe un sentido mínimo de "racional" –llámese *racional*$_1$– que hace verdadero por definición que toda acción es racional por el hecho mismo de ser acción y no una mera reacción o respuesta refleja, esto es, algo que alguien *hace* y que no simplemente *le pasa*. Éste es el sentido implícito en la afirmación "por algo lo hará" a la que se recurre cuando la conducta de alguien resulta ininteligible, desconcertante o estúpida, pero sigue considerándosela una *actuación* y no una mera reacción o un acto reflejo. Y no se trata tan sólo de una desesperada profesión de fe en el principio de razón suficiente: se postula la racionalidad del agente porque es imposible disociar lo que se interpreta como una acción del hecho de que el agente ha debido (de) tener razones de algún tipo para actuar. Se conjetura que de hecho las ha tenido porque es ineludible tenerlas. En este sentido mínimo, "racional" entra en la definición misma de acción; aunque ello no equivale a considerarla meramente tautológica y que no informe de nada no sabido de antemano.

Pero de hecho ese sentido mínimo de racionalidad no es el único ni el principal que ordinariamente se aplica a las acciones. Hay al menos otro sentido de "racional" –por ejemplo, *racional*$_2$– según el cual no toda acción que es *racional*$_1$ es ipso facto *racional*$_2$. Es el empleado más arriba al calificar una acción de inteligente o acertada, de estúpida o imprudente o incluso de inmoral. Este uso presupone un canon, norma o criterio, no necesariamente moral, al que la acción *debería* –el verbo es importante– haberse ajustado. *Racional*$_1$ es un requisito *débil* que, por poder ser satisfecho por cualquier conducta intencional, se incluye en esque-

mas explicativos de amplio alcance, pero que poseen poco potencial predictivo o explicativo, precisamente por excluir pocos comportamientos observables. En este sentido amplio, a poco que se haga un mínimo esfuerzo, puede "entenderse" –humanamente, e incluso demasiado humanamente– tanto la conducta del empresario inteligente como la del empresario torpe. Pero para predecir qué tipos de conducta harán sobrevivir una empresa o explicar qué tipo de decisiones la hicieron fracasar hay que complementar el supuesto genérico de racionalidad con ulteriores postulados que lo hagan operativo. La racionalidad del *homo economicus* explica el funcionamiento del mercado; pero la estructura de interacción del mercado determina qué decisiones pueden considerarse racionales. Por eso *racional*$_2$ es un requisito *fuerte* empleado "en teorías destinadas a poseer un fuerte potencial explicativo y predictivo, [...] para formular una teoría que elabore un tipo ideal al que las condiciones reales pueden aproximarse, pero nunca representar plenamente" (Benn, S., 1976: 1). Toda conducta *racional*$_2$ es, por definición, *racional*$_1$, pero la inversa no es verdadera.

La distinción entre acontecimientos y acciones y la consiguiente dualidad de las perspectivas del espectador y el agente se reflejan en una esencial diferencia entre los modelos explicativo-descriptivos y los modelos normativos. El concepto de norma simplemente está fuera de lugar en el ámbito de la naturaleza. Los acontecimientos naturales son lo que son y acontecen como acontecen, de forma tal vez necesaria y, como tales, ni son ni dejan de ser conformes a norma alguna. Para una epistemología realista la propia naturaleza en su facticidad misma es, en todo caso, la norma a la que deben ajustarse los modelos teóricos que aspiran a ser verdaderos u objetivos. En este sentido toda teoría *científica* pretende ser la respuesta articulada y sistemática a la pregunta por lo que hay, hubo o habrá. Pero los agentes no pueden ver sus propias acciones, aquellas sobre las que deliberan, como naturaleza –como algo ya dado– o como meros acontecimientos, sino que necesariamente las ven como lo que resultará de su personal decisión, y ésta implica de un modo u otro la respuesta a otro tipo de pregunta: "¿qué hacer?". Y esa respuesta es forzosamente normativa –ciertamente de formas muy diversas: estrictamente deóntica, técnica, prudencial, etc.–, o sea, *práctica,* en cuanto no se limi-

ta a *describir* lo que hay sino que *prescribe* qué hacer. El problema con los modelos de racionalidad de que se sirven las ciencias morales es que parecen ser al mismo tiempo descriptivos y prescriptivos.

El reconocimiento de los diversos usos de "racional" antes indicados ha obligado a distinguir, por ejemplo, entre un concepto "estrecho" y uno "amplio" de racionalidad. En el primero la racionalidad es una propiedad formal de las acciones y se reduce a la consistencia interna "del sistema de creencias, [...] del sistema de deseos; y entre las creencias y los deseos, por una parte, y la acción para la que son razones, por otra", sin entrar a examinar unos ni otras (Elster, J., 1987: 10). En el segundo, exige tanto a los deseos como a las creencias que satisfagan determinados requisitos materiales y sustantivos. De hecho es sólo la racionalidad entendida en el primer sentido, como exigencia formal de que las preferencias del agente "satisfagan ciertos axiomas de consistencia y de continuidad", la que permite a la teoría bayesiana de la decisión demostrar que actuar racionalmente equivale a, o consiste en, maximizar la utilidad esperada, sin que la utilidad sea a su vez más que la medida de la preferencia entre alternativas (Harsanyi, J., 1976b: 318-322).

En todo caso parece evidente que la tarea de elaborar un modelo de racionalidad implica un amplio y complejo conjunto de cuestiones que requiere el concurso de muy distintas disciplinas, cuyos análisis convergen sobre un mismo punto focal: la acción humana, que por su misma naturaleza se sitúa en la interfaz de las ciencias morales y naturales, de los modelos normativos y positivos, de las perspectivas del espectador y del agente. El supuesto implícito es que *un mismo modelo* de racionalidad práctica debería servir tanto a los fines de las ciencias sociales –entender, explicar y predecir la conducta real de los agentes– como a los de la teoría ética –justificar racionalmente la conducta moral–. Explicitar ese supuesto y precisar el concepto de racionalidad práctica es una tarea que sólo puede llevarse a cabo en el marco de una constante interacción de las ciencias morales y la filosofía.

Es justamente en ese marco en el que se plantean actualmente las relaciones entre la ética y la decisión racional. Pocas veces como en el presente se ha dado una convergencia tan deliberada y sis-

temática entre los resultados de las ciencias explicativas de la conducta y la filosofía moral. Importantes filósofos contemporáneos sostienen incluso que la propia ética es parte de la *teoría de la elección racional,* que ha desarrollado un modelo elegante y preciso de la racionalidad práctica aceptado como estándar en múltiples enfoques interdisciplinares (Elster, J., 1986: 9-65; 1989: 189-197; Colomer, J., 1990; Gauthier, D., 1994: 18; Green, D., 1994: *passim*; Cabrillo, F., 1996; Friedman, J., 1996a: *passim*; Marsh, D., 1998: 85-101).

1.4. Moralidad y racionalidad

Como ocurre con otras ramas de la filosofía, no existen ni una definición canónica de la ética, ni un acuerdo unánime sobre su naturaleza, su objeto o su metodología. La mayor parte de la tradición filosófica la incluye, en cuanto disciplina filosófica, dentro del género de la filosofía práctica, pues trata de la acción moral, que es un cierto tipo de acción racional. Se discute si existe, y en su caso, en qué consiste, alguna peculiaridad típica que caracterice específicamente las acciones –pero también el lenguaje, o los sentimientos, o los principios– morales frente a los no-morales. Es decir, alguna propiedad cuya presencia sea condición necesaria, o suficiente, o ambas cosas, para considerar moral, por ejemplo, un criterio de decisión. Pero en ningún caso se prejuzga la continuidad básica del ámbito de la acción racional. Podría considerarse que una excepción es Kant, cuya exigencia de autonomía para la razón pura práctica convierte la Ética en un ámbito *sui generis* escindido o transcendido del resto de la racionalidad práctica en la que intervienen motivos empíricos. Pero ni siquiera Kant renuncia a mostrar la unidad de la razón práctica. Su concepto de *bonum consummatum* garantiza la continuidad de la acción racional a lo largo de la escala de bienes –*nil appetimus nisi sub ratione boni*– y la supremacía en esa escala del bien moral.

Por ello puede compartirse la convicción de Moore para quien una gran parte de los filósofos morales creen adecuado definirla como "aquella disciplina que trata de la cuestión de lo que es bueno o malo en la conducta humana" (Moore, G., 1983: 2). Ya

advertía Aristóteles que, aunque esta cuestión es en principio teórica, pues se trata de *saber* en qué consiste obrar bien o mal, en última instancia es práctica, pues la pregunta a la que intenta responder es la del agente que se plantea *qué hacer* y *por qué hacerlo:* el propósito de la Ética no es la teoría, como lo es el de los demás tratados, pues en ella no se investiga para *saber* qué es la virtud, sino para *ser* buenos (Aristóteles, 1959: 20 [1103b26]). En tiempos relativamente recientes el debate en torno a este asunto en la filosofía inglesa y norteamericana se ha planteado como respuesta a la pregunta *why be moral?*, es decir *¿por qué actuar moralmente?* O, más radicalmente, *¿por qué ser (o convertirse en) un agente moral?*

Así planteada, la pregunta parece inocente, pero en modo alguno lo es. Muchos filósofos morales desde los inicios de la reflexión ética han creído que sólo puede responderse a ella como a cualquier otra pregunta práctica: demostrando que obrar moralmente –obrar bien– hace bien, es decir, *beneficia* de un modo u otro al propio agente y en consecuencia le interesa; tal vez no sólo ni principalmente a él –admitiendo intereses no egoístas– pero siempre y también a él. Éste es el clásico supuesto socrático según el cual toda virtud favorece el propio interés, que es desarrollado en distintos diálogos platónicos.

En el que lleva su nombre, el joven *Menón* asiente al argumento de Sócrates: todos reconocen que "somos buenos gracias a la virtud" y, por lo tanto también útiles, porque todo lo bueno *–tagathá–* es útil *–ofélima–*, de lo que se concluye que también la virtud es útil (Platón, 1970: 40-41 [87d-e]). En *Gorgias* Sócrates arguye a Polo que obrar justamente es mejor *–ámeinon–* y peor hacerlo injustamente. El hombre honrado es feliz y el malvado, desgraciado (Platón, 1983a: 58-59 [470c-d]). Cometer injusticia es peor que padecerla (Platón, 1983a: 57[469b]; 63 [473a-b]). En la opinión común, que hasta Aristóteles compartirá, la justicia es la virtud más perfecta porque quien la posee "puede usar de la virtud con otros *–pros héteron–* y no sólo en sí mismo": entre todas las virtudes, es la única "que parece consistir en el bien ajeno, porque se refiere a los otros" (1959: 71-72 [1129b35-1130a5]). Pero ésa es justamente la característica de la justicia –y por extensión de la moralidad– que lleva a Trasímaco a considerarla bene-

ficiosa para los demás pero perjudicial para uno mismo. De ahí que en la *República* Sócrates intente demostrar que la justicia beneficia realmente al individuo, a pesar de que según todas las evidencias sus beneficiarios directos sean *los demás* incluso en perjuicio de quien la practica. Si la justicia tuviese como único objeto el propio bien, no es evidente que el justo haya de llevar a cabo acciones justas en beneficio ajeno a costa del propio. A menos que pueda defender la racionalidad de una virtud que beneficia a los demás no logrará refutar las objeciones de Trasímaco y Glaucón (Irwin, T., 1977: 200-217).

Frente a Trasímaco, Sócrates no admite que la injusticia sea más ventajosa que la justicia "ni aun cuando se dé a aquélla rienda suelta y no se la impida hacer cuanto quiera" (Platón, 1969: 36 [345a]). Todas las cosas a las que se atribuye una operación –*érgon*– poseen una virtud gracias a la cual desempeñan mejor esa operación propia: ver, los ojos; oír, los oídos. Y lo mismo, para el alma, "dirigir, gobernar y deliberar": el alma mala lo hace mal y la buena, bien, gracias a su virtud propia –*oikeía areté*– que es la justicia. Por eso el justo es dichoso y el injusto desgraciado. Y como *no conviene* ser desgraciado, sino dichoso, jamás es la injusticia más *provechosa* que la justicia (Platón, 1969: 52-55 [353b-354b]). Por esa razón es peor cometer la injusticia que padecerla.

Adimanto considera que no es con meras definiciones como puede probarse que la justicia es mejor que la injusticia y exige a Sócrates que lo demuestre no sólo con palabras –*mónon toi lógoi*– sino haciendo ver lo que por sí mismas *producen* –*ti poioúsa*– en quien las practica (Platón, 1969: 73 [367e]). Poco antes Glaucón ha recogido el argumento de quienes afirman que cometer injusticia es por naturaleza –*physei*– un bien, y sufrirla un mal; pero como "es mayor el mal que recibe el que la padece que el bien que recibe quien la comete" –en el discutible supuesto, del que se hablará más adelante, de que puedan compararse las utilidades personales– por razones de utilidad común se establecieron convenios –*nómoi*– para no cometer ni padecer injusticias. Pero la misma consideración de utilidad que lleva a *suscribir* públicamente los acuerdos es la mejor razón para no *cumplirlos* en privado, si hay garantías de no ser descubierto. El que se atiene a la justicia no lo hace por su gusto –*hekón*– sino por fuerza –*anan-*

kazómenos–. Prueba de ello es que quien pudiera defraudar a escondidas –por ejemplo, haciéndose invisible mediante un anillo como el de Giges – "cometería todo tipo de tropelías en provecho propio", persuadido de que "resulta mucho más ventajosa personalmente la injusticia que la justicia". De no hacerlo así sería considerado el ser más desgraciado –*athliótatos*– e irracional –*anoetótatos*– (Platón, 1969: 58-61 [358e-360c]; Gutiérrez, G., 1979: 43-47).

No es esencial para el argumento el específico tipo de deseos –por ejemplo, los llamados *bajos*– a los que daría rienda suelta quien se supiera invisible; no hace falta suponer que, libres de toda restricción, todos los hombres estarían determinados por naturaleza a "dirigirse al mercado para tomar de allí sin miedo alguno lo que quisieran, entrar en las casas ajenas y fornicar con quienes se les antojase, matar o liberar a su arbitrio y obrar, en fin, como un dios entre mortales" (Platón, 1969: 61 [360b-c]). El problema está en parte en la naturaleza humana, pero se plantea específicamente por la diferente naturaleza de los bienes individuales o privados y los comunes o públicos.

Estos últimos son producto del acuerdo cooperativo suscrito y cumplido por múltiples individuos que, por ejemplo, pagan impuestos o el billete de autobús, donan sangre, respetan las normas de tráfico, participan en una huelga o votan en las elecciones. El problema es que entre sus beneficiarios potenciales se cuentan todos los miembros de la colectividad, por lo que no es posible excluir a los "gorrones" y "parásitos" que se benefician de ellos sin contribuir a producirlos o mantenerlos: el anonimato o el disimulo cumplen la misma función que el anillo de Giges para el asesino en serie, para el gran evasor de impuestos y para el que se cuela en el autobús sin pagar, todos son igualmente *free-riders* (Gutiérrez, G., 1990b: 13-14). Desde la perspectiva de los antagonistas de Sócrates en la *República* todos los que cumplen con las obligaciones contraídas en el acuerdo asumiendo el correspondiente coste serían tildados de auténticos "primos" –*suckers,* en el acertado término del lenguaje de germanía rescatado por la teoría de los bienes públicos– (De Jasay, A., 1989: 1-8, 134-137). Es precisamente un lugar común en el análisis de los bienes públicos hacer notar que "a casi todos los bienes públicos cuya provi-

sión requiere un gasto de recursos, tiempo o prudencia moral, se les puede aplicar una matriz de estrategias análoga al Dilema del Prisionero" (Mueller, D., 1984: 25) sobre el que habrá que volver más adelante.

Pero Sócrates tiene asimismo otros antagonistas que niegan que el concepto de bien sea unívoco y sostienen que no hay que dar por supuesto que puedan invocarse beneficios o intereses *no morales* para obrar moralmente; incluso la propia pregunta revelaría una *ignoratio elenchi* sobre la naturaleza específica de la moral. De esta opinión era el filósofo inglés Harold A. Prichard cuyo célebre artículo de 1912 en la revista *Mind* cuestionaba ya desde su mismo título si la propia filosofía moral no se basaría en un error (Prichard, H., 1912) y que reprocha a Platón que en ningún momento pusiera en duda el supuesto que compartían Sócrates y los sofistas de que, para que una acción sea verdaderamente justa, ha de ser ventajosa, porque sólo sobre esta base pueden concluir que lo que de ordinario creemos justo en realidad no lo es; y que es imposible que una acción sea realmente justa, es decir, un deber, a menos que sea ventajosa para el agente (Prichard, H., 1970: 693).

Es evidente que en todo caso el planteamiento y las soluciones del problema de las relaciones entre moralidad, racionalidad e interés dependen de la posibilidad de justificar la existencia y en su caso precisar la naturaleza de la distinción entre razones, intereses y motivos morales frente a los no morales. La cuestión afecta a la definición misma de la racionalidad práctica y a la condición racional de la moralidad.

Aunque pecaría de reduccionista todo intento de identificar alguno de los múltiples problemas a los que se enfrenta la filosofía moral como *el* problema fundamental de la ética, hay que rendirse a la evidencia de que tanto la estructura teórica de la disciplina como su historia desde los sofistas a nuestros días ponen de manifiesto que al menos esta cuestión –la de la relación entre moralidad e interés– justificaría por sí sola la existencia de la reflexión filosófica que constituye la ética. Todas las demás la presuponen o conducen a plantearla. Si se procede a reconstruir la génesis y la naturaleza de las relaciones entre moralidad e interés se pondrá de manifiesto que, bajo diversas formulaciones la cuestión de fondo

es intuitivamente simple. En todo caso parece haber razones de peso para sostener que *este* problema lo suscita necesariamente la condición social del hombre. Como se verá a continuación, difícilmente se les puede plantear a Adán en el paraíso o a Robinson en su isla mientras permanecen a solas un conflicto entre deber moral e interés, pero no *pueden* escapar a él una vez en compañía.

Una de las primeras evidencias a las que tiene que enfrentarse el ser humano en su desarrollo personal es la de no poder hacer cuanto desea o cuando lo desea –en términos freudianos: el principio de realidad se opone al principio del placer–. En algunos casos la imposibilidad es genéricamente física e impuesta desde el exterior: no puede volar, no le dejan seguir jugando. A medida que el individuo madura, la restricción, sin dejar de ser en cierto modo física y externa, se interioriza y se sitúa en una perspectiva temporal que antes le era ajena: no se trata ya de la mera *imposibilidad* física de satisfacer un deseo actual, sino de la *inconveniencia* de satisfacerlo al precio de una satisfacción futura mayor. El glotón avisado que, a la vista de la carta, resiste a su deseo presente de hartarse de un primer plato apetitoso para mejor gustar un segundo más apetitoso, efectúa ya un rudimentario cálculo que, aunque materialmente hedónico, es formalmente racional. Este argumento cuenta con una ilustre prosapia. Cicerón lo pone en boca del epicúreo Torcuato, quien afirma que nadie rehuye el placer por serlo sino porque, si no se goza con discernimiento, se siguen grandes dolores; ni busca el dolor por ser dolor sino porque, al precio de algún dolor, se consigue a veces algún gran placer. Por eso son censurables y despreciables quienes, seducidos por el placer del momento, no prevén los dolores y trabajos que les esperan (Cicerón, 1987: 70 [I, § 32-33]).

Parece consustancial a la condición misma de agente racional, y en particular a su dimensión temporal, la necesidad de distanciarse de los deseos y aun de los intereses presentes, restringiéndolos racionalmente con el propósito de satisfacer intereses o deseos propios, cualitativamente superiores o simplemente ulteriores. La historia de la filosofía abunda en distinciones que reconocen que los deseos, los motivos, los intereses o las preferencias no están situados en el mismo plano sino que manifiestan algún tipo de jerarquía. Platón distingue entre diferentes partes del alma:

apetitiva *(epithymetikón)*, emotiva (*thymoeidés)* y *racional (logistikón).* La escolástica lo hace entre *appetitus sensitivus* y *rationalis.* Hume admite la oposición entre *passion* y *reason,* aun manteniendo reservas sobre su verdadero alcance. Kant distingue entre arbitrio *(Willkür)* que requiere de un incentivo para actuar y voluntad *(Wille)* que no lo necesita. Al distinguir entre *desire* y *will* o entre cualidades –y no simples cantidades– de placer Mill puede incluso afirmar que un Sócrates insatisfecho es preferible a un tonto satisfecho. Entre los filósofos contemporáneos de la economía es habitual distinguir entre *gustos* y *valores* (Arrow), entre preferencias *de hecho* y *preferencias hipotéticas o morales* (Harsanyi), o entre preferencias *reveladas* y preferencias *reales* (Sen).

En una perspectiva meramente temporal, ser mínimamente racional equivale a ser literalmente *prudente –prudens* es el *providens,* capaz de prever, de anticiparse al futuro y por tanto de proveer para sí–. En la fábula clásica, enfrentadas a la elección elemental entre consumir ahora o consumir ahora, la horaciana cigarra *carpit diem,* mientras la hormiga previsora trueca su disfrute veraniego presente por un invierno abastecido. El prudente *no puede* –no por supuesto en sentido físico– hacer todo lo que ahora querría, porque prevé que más adelante querrá no haberlo querido. El *no poder* trasciende su mera condición primitiva de imposición externa y se convierte en una restricción autoimpuesta. Aparece ya aquí una noción rudimentaria y elemental de obligación, ciertamente interesada.

En este tipo de situaciones resulta difícil entender en qué podría consistir un conflicto en sentido estricto entre obligación e interés, al menos en el ámbito de la racionalidad individual. En lo que atañe a sí mismo, en efecto, el individuo no se impondría ninguna restricción –ninguna obligación– que no redundase en su propio interés racionalmente determinado. Solitario en su isla, un Robinson prudente reconocería tanto la existencia de leyes naturales ineluctables cuanto las limitaciones de su personal racionalidad. Por ser imperfectamente racional, su limitado conocimiento de la naturaleza afectaría de incertidumbre sus previsiones de futuro. Igualmente limitados serían su conocimiento y sobre todo su dominio de sí mismo, lo que se manifestaría en el conflicto interior entre su "razón" y sus "pasiones"; en la debilidad de su voluntad voluble e irresoluta, afectada por mecanismos

inconscientes que le hacen proclive al autoengaño; en su incapacidad para saber, no ya lo que realmente le conviene, sino ni siquiera lo que realmente quiere.

Teniendo todo eso en cuenta, si realmente deseara promover su propio beneficio e interés, según su propia concepción de lo bueno, Robinson no tendría otro remedio que imponerse obligaciones estrictamente prudenciales que le forzarían a resistir a sus deseos presentes o inferiores, los cuales, de ser satisfechos, le impedirían producir u obtener bienes superiores o ulteriores. Por el mero hecho de reflexionar sobre lo que es bueno para él en su situación formularía –siquiera rudimentariamente– y pondría en práctica alguna teoría sobre la naturaleza de lo bueno, que lo hiciera consistir, por ejemplo, en un ideal de la propia excelencia, o en la impasibilidad ante las veleidades del azar, o en el equilibrio de una vida frugal y placentera, etc. Defoe cuenta que Robinson anota las ventajas y desventajas de su condición en el debe y el haber de una rudimentaria contabilidad. Cuando deja de escudriñar el mar en espera de un barco se aplica a la tarea de acomodarse a su forma de vida y facilitársela lo más posible, convencido de que la naturaleza y la experiencia de las cosas enseñan que todas las cosas buenas de este mundo sólo lo son en la medida en que pueden ser usadas. En consecuencia, no estaba ocioso ni ahorraba esfuerzos para conseguir lo que le parecía necesario para su cómodo sustento, tomando todas las medidas que la humana prudencia podía sugerir para su propia preservación.

Kant tiene razón al señalar que las obligaciones prudenciales, precisamente por ser interesadas, no son especialmente complicadas de entender: parece evidente que quien quiere el fin quiere los medios indispensables para obtenerlo (Kant, I., 1996: 165), aunque, por más que considere analítica esta proposición, la existencia de la *akrasía* o flaqueza de voluntad plantee graves cuestiones sobre su alcance práctico. Lo que, en cambio, resulta desconcertante es llamar *obligaciones* a unas restricciones *autoimpuestas.* No sólo porque parece extraño al uso ordinario del lenguaje, sino porque –como también hace observar Kant– "si el yo *que obliga* se toma en el mismo sentido que el yo *obligado,* el deber hacia uno mismo es un concepto contradictorio": no está en absoluto obligado quien puede dispensarse a sí mismo del deber que se impone (Kant, I., 1989: 274-275). La observación sólo tiene sen-

tido si se supone, como hace Kant, que una obligación en sentido estricto sólo puede ser literalmente *indispensable* –incondicional, absoluta y categórica, es decir *moral*–, mientras que la interesada estaría condicionada a la existencia previa de un interés.

De lo dicho hasta ahora no puede, por tanto, concluirse sin más que las obligaciones interesadas sean o no obligaciones *morales* en sentido propio. Así planteada, la cuestión es incluso capciosa si da por sentado que existe algún tipo de oposición entre moralidad e interés, lo cual depende, en gran parte, de la definición misma de moralidad. Dos teorías histórica y conceptualmente muy distantes, como son el tomismo y el consecuencialismo, coinciden en negar que un acto propiamente humano, esto es, deliberado y voluntario, pueda ser indiferente desde el punto de vista moral. La filosofía tomista afirma que todo acto racional, o está ordenado al fin debido o no lo está, y por tanto necesariamente ha de ser moralmente bueno o malo. Para el consecuencialismo no tiene sentido distinguir conceptualmente entre las decisiones o acciones que plantean cuestiones morales y las que no las plantean, pues en principio toda cuestión práctica acerca de lo que es racional hacer, evitar o incluso impedir que ocurra se decide en definitiva en función de la exigencia consecuencialista de producir las mejores consecuencias y se convierte, por definición, en una cuestión moral. En ambos casos, de la definición misma de moralidad parece seguirse que toda obligación racional es *eo ipso* moral. Pero como la cuestión no es meramente semántica, poco se gana circunscribiéndola a un mero examen de las definiciones propuestas. Si se parte de la definición kantiana de la moral en términos de autonomía normativa, se sigue lógicamente que, por ejemplo, el utilitarismo, no es tanto una forma alternativa de justificación de la conducta moral cuanto una guía heterónoma de conducta que se presenta como alternativa a la moral misma. Pero la cuestión de fondo sigue siendo la definición *real* de la moralidad.

1.5. Ética y elección racional

También se ha argüido a la inversa y sostenido que una obligación moral sólo puede justificarse si se demuestra que son racio-

nales los principios en que se sustenta. Es decir, que serían adoptados como criterios de decisión por agentes racionalmente interesados en el sentido que precisa, por ejemplo, la teoría de la elección racional. En este marco conceptual hay que situar las propuestas de tres importantes filósofos contemporáneos.

Cuando John Rawls describe la racionalidad de las personas que eligen los principios de la justicia tras un "velo de ignorancia" en la posición original, advierte expresamente que invoca "el concepto de racionalidad estándar que nos resulta familiar en la teoría social" (Rawls, J., 1979: 170). La propia posición original es construida por Rawls como un caso típico de elección en condiciones de incertidumbre o ignorancia en la que los agentes aplican al decidir el criterio *maximin* tal como lo formula la teoría de la decisión y que se discutirá más adelante. En ausencia de toda información sobre la probabilidad de que acontezca un determinado suceso favorable o desfavorable –que uno ocupe una posición más o menos ventajosa en la sociedad cuyos principios de justicia están siendo elegidos– el criterio recomienda elegir la alternativa cuya peor consecuencia sea la menos mala.

John Harsanyi considera expresamente la ética como una rama de la teoría general de la elección racional junto con las teorías de la utilidad, de la decisión y de juegos, ya que se la puede fundar en axiomas que representan especializaciones de algunos de los axiomas utilizados en esas teorías. Si la teoría de juegos es una teoría de los intereses individuales en posible conflicto, la ética puede ser entendida, en la mejor tradición utilitarista, como una teoría de los intereses comunes o del bienestar general de la sociedad: la teoría de los juicios racionales de preferencia basados en criterios imparciales e impersonales cuyo objetivo es maximizar el nivel promedio de utilidad de todos los individuos de la sociedad (Harsanyi, J., 1976b: 323). Harsanyi sostiene, en contra de Rawls, que en las condiciones de ignorancia que caracterizan la posición original no hay justificación racional para asignar tan alta probabilidad al peor resultado, haciendo así depender la elección de contingencias sumamente improbables y comprometiendo la utilidad que cabría racionalmente esperar. Argumenta, en cambio, que los decisores racionales asignarían una probabilidad igual a los sucesos favorables o desfavorables y elegirían por tanto aque-

lla sociedad cuya utilidad esperada promedio fuese más alta (Harsanyi, J., 1976a: 39-48). Cuando se trata de elegir entre dos loterías con las mismas probabilidades de ganar el premio lo racional es jugar a aquella cuyo premio es mayor.

David Gauthier es igualmente explícito al declarar su propósito de "desarrollar una teoría de la moral como parte de la teoría de la elección racional que forma el núcleo de la teoría económica clásica y neoclásica y que preserva su fundamental identificación de racionalidad con maximización" (Gauthier, D., 1994, 18); y que, además, es una concepción "aceptada casi universalmente por las ciencias sociales, algo de la mayor importancia" (íd., 23). Pero, a diferencia de Rawls y Harsanyi, considera que los principios morales no son tanto *objeto de* una elección racional cuanto *criterios para* la elección racional. Ambos representan inadecuadamente la elección de los principios morales según el esquema de las decisiones paramétricas (loterías) cuando en realidad se trata de una decisión estratégica característica de determinadas situaciones de interacción (juegos) –más específicamente, las que responden al tipo del *dilema del prisionero*– que hay que analizar en términos de la teoría de la negociación (Gauthier, D., 1998a: 68-69). El dilema muestra que en tales situaciones la estrategia racional de maximización directa es lesiva incluso para los propios intereses del decisor racional. Gauthier intenta demostrar que en ellas lo racional –es decir, lo que maximiza la utilidad del agente– es precisamente decidir en función de principios cooperativos (condicionales) o morales. Una moralidad racional es, entonces, una restricción, o conjunto de restricciones, a la maximización de lo que el agente valora, tales que quienes aspiran a maximizar tal valor encontrarían racional adherirse a ellas y cumplirlas (Gauthier, D., 1998a: 93).

Sin entrar en este momento a considerar los méritos y las complejidades de cada una de estas propuestas, su misma articulación muestra hasta qué punto el desarrollo de la teoría de la elección racional ha permitido al menos plantear con inusitada claridad y precisión algunas de las cuestiones fundamentales tanto de la filosofía moral como de las ciencias morales.

2
La agencia humana

2.1. Tiempo, decisión y libertad

La condición humana en general, su experiencia de la realidad, y muy en particular, su condición de agente están indisolublemente asociadas al tiempo. La mitología y la filosofía griegas –especialmente la heraclitiana– recurrieron a la imagen del río para ilustrar una fluyente condición a la que no escapan ni siquiera los dioses, que portan incluso "nombres de corrientes" –*rheumáton*– (Platón, 1983b: 397 [402b]). El nombre de *Rhéa*, hija de Ourános y Gea y madre de Zeus, es relacionado con *rhéo*, fluir; el de *Krónos* –hermano y marido de Rhéa y castrador de su padre común con una hoz– lo es con *krounós*, el surtidor de una fuente. Es significativo que los propios griegos terminaran confundiéndolo con *Chrónos*, el Tiempo divinizado, al que la iconografía presenta también armado de una guadaña. Aunque las etimologías son discutibles, el juego de ideas es transparente. Las cosas inertes son relativamente ajenas al tiempo porque su naturaleza está dada de antemano y fijada de una vez para siempre. Los organismos vivientes, en cambio, al ser producto de una génesis, dependen en mucha mayor medida de él: la naturaleza de la especie y del individuo son la culminación de un desarrollo temporal. Pero aun en este caso el peso de la programación genética previa es muy superior al de los episodios aleatorios de interacción con el medio. Sólo la naturaleza humana es constitutivamente abierta e inacabada y necesita, más que llegar a ser, hacerse en el tiempo, permitiendo e imponiendo al hombre la perspectiva interna sobre sí mismo que lo constituye como agente.

Esta perspectiva se manifiesta en su consciencia en forma de una doble y paradójica necesidad. Por una parte, el agente se ve irremisiblemente sumergido en una corriente temporal que lo arrastra "río abajo" sin que esté en su mano invertir su dirección ni detener su incesante fluir –*fugit irreparabile tempus*–. Por otra, al no estarle su ser totalmente dado de antemano, se ve forzado a decidir en cada instante lo que ha de ser en el instante siguiente. Es, pues, consciente de no poder no decidir y, *al tiempo,* de ser forzado a hacerlo libremente. Su condición de agente es inseparable de un grado mínimo, pero irreductible, de espontaneidad y de autonomía.

Su identidad y continuidad personal resultan de la serie de decisiones que adopta y de las acciones que ejecuta; al menos hasta cierto punto, pues ni los agentes ni sus decisiones existen en el vacío. Pero, en la medida en que las acciones son intencionadas y responden a razones, el agente, más que mero causante, es concebido como responsable –capaz de dar razón– de lo que hace y de lo que llega a ser. No es en sentido estricto creador de sí mismo, ni es literalmente cierto que su existencia preceda a su esencia, pero como hasta cierto punto se hace a sí mismo por medio de sus actos no es una mera metáfora considerarlo *hijo de sus obras,* como sentencia Don Quijote a Sancho. Podría incluso interpretarse en este sentido la distinción que introduce Aristóteles entre *poiésis*, acción instrumental cuyo objetivo y resultado es un objeto *(érgon)* exterior al agente y cuyo valor se mide por su eficacia en producirlo, como es para el escultor esculpir la estatua, y *enérgeia*, aquella otra cuyo valor está en sí misma y cuyo "resultado" permanece en el propio agente, única a la que considera Aristóteles propiamente acción –*práxis*– (1959: 1 [1094 a1-18]; Gauthier, R., 1973: 38-44). El agente viene a resultar así escultor de sí mismo y responsable de su carácter *éethos* (Taylor, C., 1982: 111-126).

Lo que vale para los agentes individualmente considerados puede extenderse al resultado de las interacciones entre múltiples agentes. Para la teoría del individualismo metodológico las acciones e interacciones individuales son la materia de la que están hechas en última instancia todas las instituciones sociales, económicas y políticas (Lukes, S., 1975; Elster, J., 1984; Bhargava,

R., 1992). Mill ya hacía observar que éstas no tienen otras propiedades que aquellas que pueden derivarse de las leyes de la naturaleza del hombre individual: "los hombres, aunque en estado de sociedad, son siempre hombres; sus acciones y pasiones obedecen a layes de la naturaleza humana individual; [...] no son, cuando están reunidos, transformados en yo no sé qué otra sustancia dada de propiedades nuevas, a la manera que el oxígeno y el hidrógeno difieren del agua" (Mill, J. S., 1917: 887). No tienen, en definitiva, otras propiedades que las que se derivan de las leyes naturales del individuo y que a ellas se pueden reducir, o resolverse en ellas por medio de la composición de las causas. La sociedad se explica por –y se reduce a– el entramado infinitamente complejo de acciones individuales e interacciones recíprocas: "las acciones y los estados de espíritu de los seres humanos en sociedad son incontestablemente regidos en su totalidad por leyes psicológicas y etológicas; cualquiera que sea la acción que otras causas puedan ejercer sobre los fenómenos sociales, no la ejercen sino por el intermediario de estas leyes" (íd.: 907). Su identidad y continuidad sería por tanto el efecto –no necesariamente intentado ni deseado– de las acciones de sus miembros.

La fugacidad del tiempo y la necesidad de decidir fuerzan al agente a formularse la pregunta práctica ineludible y radical: *¿qué hacer?* y a hallar de forma perentoria una respuesta. Desde la perspectiva –externa– de la evolución histórica de la especie, esa respuesta la ofrecen los rituales, las costumbres y, en general, las normas y pautas de conducta socialmente sancionadas. Las disposiciones psicológicas del individuo, en forma de inclinaciones, sentimientos, emociones o deseos, lo predisponen a aceptar determinadas respuestas. La realidad objetiva –física, social, histórica, económica– proporciona y restringe a un tiempo las oportunidades disponibles para hacer efectivas tales respuestas.

Pero desde la perspectiva interna propia del agente ninguna de esas posibles respuestas a la pregunta desencadena automáticamente la decisión. Es imposible *atenerse* a lo establecido, *abandonarse* a sus impulsos o *dejar* que los hechos decidan por uno sin conjugar en primera persona verbos de acción, esto es, sin decidir. Una vez más: el agente, en cuanto agente, no puede ver esas cosas como algo que *le pasa* sino como algo que *él hace*. Incluso

un hipotético no decidir seguiría siendo decidir. En la conocida historia del asno de Buridán –que, enfrentado a dos montones de heno idénticos, fue incapaz de decidir entre ellos y murió de hambre– no hubo *in*decisión en sentido estricto: entre elegir cualquiera o ninguno, el asno de hecho eligió esto último y como consecuencia pereció (Sen, A., 1986b: 67-69).

Por eso las diversas doctrinas filosóficas o místicas –estoicas, budistas, fatalistas, quietistas– que coinciden en proponer como objetivo práctico la supresión del deseo presentan idéntico aire paradójico e incurren en incoherencias pragmáticas. Ponerlas *en práctica* implica querer lo que no puede ser querido, desear dejar de desear, actuar para dejar de ser activo, suspender toda deliberación y elección para abandonarse al fatum o a la voluntad divina –empeño tan vano como huir de la propia sombra: no hay fatalista consecuente– (Mill, J. S., 1917: 843).

El agente, como tal, no puede pensarse como sujeto pasivo de determinaciones externas. Lo que llama "yo" aparece siempre un paso más allá de la consciencia de sus determinaciones, pues lo que está ineludiblemente asociado al actuar no es tanto *ser* libre cuanto la *consciencia* de serlo. Como hizo ver Kant, es imposible concebirse a sí mismo como agente "de otra suerte que bajo la idea de la libertad", por lo que, a todos los efectos prácticos, es *como si* realmente se fuese libre. Actuar es por definición un ejercicio de la voluntad y "a todo ser que tiene una voluntad debemos atribuirle también necesariamente la idea de la libertad, bajo la cual obra [...] y es imposible pensar una razón que con su propia consciencia reciba respecto de sus juicios una dirección cuyo impulso proceda de otra parte, pues entonces el sujeto atribuiría, no a su razón, sino a un impulso, la determinación del juicio. Tiene que considerarse a sí misma como autora de sus principios, independientemente de ajenos influjos" (Kant, I., 1996: 227). La afirmación de Séneca *ducunt volentem fata, nolentem trahunt* (Séneca, 1953, 336 [107 § 11]; Epicteto, 1991: 117 [§ 53]) podrá o no ser *verdadera* en términos teóricos, pero ciertamente es *irrelevante* a efectos prácticos, pues es imposible extraer de ella ninguna indicación sobre qué querer o no querer. La negación por el agente de su propia libertad se revela como una imposibilidad conceptual y pragmática. A esta autoconsciencia del sujeto como

agente libre no es óbice afirmar que "en realidad", esto es, desde el punto de vista del *espectador,* el agente está determinado y que la libertad es ilusoria, pues aun en ese caso se trata de una ilusión estrictamente incorregible. En este punto la consciencia es impermeable a la realidad.

Como ejemplos recientes de algunas implicaciones de esta tesis basta citar la contradicción en que, a juicio de Popper, incurre el "historicismo" al fundar su propósito de cambiar la sociedad en el conocimiento de la ley natural que determina su movimiento (Popper, K., 1973: 55-66), así como las conocidas paradojas suscitadas por la posibilidad de predecir las decisiones, la más desconcertante de las cuales se plantea a propósito del llamado "problema de Newcomb" (Nozick, R., 1985: 107-133; 1995: 69-80).

Numerosos filósofos –Plotino, San Agustín, Kant, Bergson o Sartre entre otros– han puesto de manifiesto las aporías a que conduce todo intento de conceptualizar el tiempo, la duración, el continuo o la libertad. Si desear y decidir son concomitantes necesarios de la consciencia –o incluso de la mera vigilia– que sólo el sueño, temporal o eterno, logra extinguir, parecería seguirse que *ser,* para el agente, es un incesante *desear* y un inevitable *decidir,* hasta el punto de que *en* cada instante de su existencia decidiría *de* su propia existencia. Algo así parece dar a entender Hobbes cuando afirma que "tan imposible es que viva ya un hombre cuyos deseos han cesado como que viva aquel cuyos sentidos y cuya imaginación se han inmovilizado (pues) la inclinación general de la humanidad (es) un deseo perpetuo e incesante de más y más poder, el cual cesa tan sólo con la muerte" (Hobbes, T., 1989: I, 87).

La experiencia real se resiste a la coherencia lógica de esta conclusión: si las cosas fueran de hecho así toda actividad consciente se paralizaría y el problema práctico de decidir se anularía a sí mismo, pues realmente no habría *desde dónde* decidir. Lo cierto es que la naturaleza y la sociedad introducen en el psiquismo del agente innumerables artificios –rutinas, costumbres, disposiciones, hábitos, virtudes– que reducen el abanico de elecciones posibles y fomentan determinadas predisposiciones a elegir de manera cuasi espontánea, con el resultado de reducir el insoportable coste de tener que decidir todo a cada momento.

No es posible entrar aquí en el análisis detallado de la naturaleza del tiempo y de la consciencia, pero parece inevitable, tanto conceptual como pragmáticamente, suponer que, en el incesante fluir del tiempo, puede y tiene que darse un *instante* en el que el agente "se pare a pensar" para responder a la pregunta práctica *¿qué hacer?* Así, pues, deliberar es un proceso temporal que exige ser concebido como una suspensión del tiempo.

2.2. Teoría y práctica

El objeto y el producto de la deliberación son los criterios, principios o razones para actuar cuyo carácter más o menos abstracto y universal –o al menos general– les confiere una relativa intemporalidad, aun cuando su función práctica los ligue a la decisión *hic et nunc*. Esto es lo que permite articularlos y sistematizarlos en esa específica forma de teorizar que la tradición filosófica clásica ha llamado *filosofía práctica,* que en tiempos recientes se extiende a las teorías de la decisión y de juegos, y que de una u otra forma siempre ha incluido, como parte especialmente significativa, la *ética.*

En la medida que es reflexiva, crítica, lógicamente argumentada y referida a la experiencia, la ética posee las características formales atribuidas en general a todo intento de sistematizar en forma de *teoría* el conocimiento de algún ámbito de la realidad para explicar lo que *es* o lo que hay. No hay que olvidar que "teorizar", etimológicamente, es lo que hace quien se limita a mirar lo que tiene delante, como contempla el espectador a quienes realmente actúan –los *agonistas*– en la escena o en el estadio. En términos de un realismo ingenuo, la teoría vendría a ser como la novela para Stendhal: un espejo que se pasea a lo largo del camino, que refleja el mundo dejándolo como estaba.

Lo peculiar de la ética es que, por ser una teoría *práctica* –o, si se prefiere, *normativa*– desempeña la ulterior función de hacer que por medio de la acción algo *sea*. Ya se vio cómo, partiendo de este supuesto, Aristóteles señalaba que el propósito de su *Ética* no era, como el de sus otros tratados, la teoría sino la práctica. Lo que en definitiva interesa no es saber *qué es* lo bueno sino

llegar a serlo, "pues de lo contrario no aprovecharía nada". Se estudia la teoría –"lo relativo a las acciones"– pero no por mor del conocimiento sino de la práctica: para saber cómo actuar (Aristóteles, 1959, 20, [1103b26-30]).

Conviene entonces precisar, siquiera someramente, dos sentidos elementales y diferentes del concepto de "teoría". Aunque el término se emplea para referirse tanto a las teorías positivas o científicas como a las teorías normativas, no es inmediatamente evidente que, por ejemplo, doctrinas éticas como el Hedonismo o el Consecuencialismo sean teorías exactamente en el mismo sentido que lo son la Teoría General de la Relatividad o la Tectónica de Placas. La distinción entre ambos tipos de teorías –normativas unas, explicativas otras– no se reduce a la mera presencia o ausencia de elementos normativos. Aún sin terciar en el debate entre los positivistas que aceptan la existencia de hechos puros dados con carácter previo a toda teoría –el "mito de lo dado"– y los pragmatistas que niegan que haya hechos independientes de toda teoría, parece evidente que la formulación de las oraciones observacionales más primitivas presupone al menos determinados *criterios* de selección y de clasificación de los datos. En todo caso, por su propia estructura toda teoría, explicativa o normativa, implica algún tipo de ajuste entre los enunciados, principios o leyes y los hechos. El propósito específico al que sirve la formulación de ambos tipos de teoría endereza en cada una el ajuste en sentidos opuestos. El objetivo de las teorías positivas o científicas es proveer razones en forma de leyes para responder a la pregunta por lo que hay, lo que pasa o lo que es el caso, incluyendo retrodicciones y predicciones respecto de lo que hubo o habrá. El de las teorías normativas o prácticas, incluyendo desde la programación lineal a la ética, es proveer razones en forma de principios, normas o reglas para responder a la cuestión sobre lo que hay que hacer para que algo sea el caso por la acción de los agentes.

Ninguna teoría, positiva o normativa, está libre de compromisos ontológicos. Para cumplir su función práctica una teoría normativa ha de presuponer a su vez una teoría "objetiva" de la naturaleza en general y de la de los agentes que deliberan, deciden y actúan en particular; es decir, una teoría que defina lo que,

en esencia, *es* la racionalidad práctica. Cuando se examina el entramado conceptual de las teorías prácticas, y en especial de la ética, se observa las implicaciones recíprocas de tres tipos o familias fundamentales de conceptos que permiten al discurso práctico desempeñar su función.

En primer lugar los conceptos propiamente *teóricos* –psicológicos, antropológicos o filosóficos en general– que pretenden describir la naturaleza objetiva y las propiedades de las operaciones de los agentes en el mundo. En segundo lugar, los conceptos propiamente *prácticos,* entre los que figuran los conceptos *valorativos* o axiológicos que se refieren a lo que es valioso o preferible y con los que se juzga de la bondad de los objetivos, fines o metas alternativas de la acción, y los conceptos normativos o *deónticos* con los que se enjuicia la corrección de la decisión por referencia a reglas, normas o principios que enuncian lo que debe (o debe no, o puede, etc.) hacerse. Las oraciones que contienen estos últimos configuran los dos tipos básicos de discurso que intervienen en la deliberación práctica: las preferencias y las normas (Wright, G.v., 1967: 9; 1970: 21-35).

2.3. Preferencias y normas

Los conceptos valorativos y deónticos forman parte de las proposiciones que pueden considerarse propiamente prácticas por enunciar las razones que impelen al agente a actuar. Sus contextos de aplicación son las situaciones reales en las que los agentes eligen, deciden y actúan. Estas situaciones presentan características formales que permiten elaborar una tipología polar, y por tanto abstracta, de situaciones de elección. En un extremo se supone al agente en condiciones de decidir libremente según sus puras preferencias personales; esto es, de hacer literalmente lo que desee, sin más restricciones que las que le impongan las condiciones materiales o los costes alternativos de sus posibles decisiones. En el otro se supone que el agente reconoce que pesan sobre él restricciones –obligaciones– que le impiden, ciertamente no de forma física, dar libre curso a sus puras preferencias.

En el texto de Cicerón citado en el capítulo anterior el epicúreo Torcuato no considera digno de censura sólo al hedonista que se priva de un placer mayor o superior por su imprevisión, sino también a quien abandona sus *deberes –officia–* por molicie, es decir, por evitarse fatigas y dolores. El propio Torcuato establece una distinción "fácil y clara" entre los dos tipos de situaciones. En circunstancias desembarazadas, cuando se tiene libre facultad de elección *–soluta eligendi optio–* y nada impide hacer lo que más agrada, hay que permitirse todo placer y rechazar todo dolor. Por el contrario, en otras ocasiones, ya sea porque existen deberes que obligan o porque la fuerza de las cosas se impone *–officiis debitis aut rerum necessitatibus–* con frecuencia acaecerá que hayan de rechazarse los placeres y no rehuirse las incomodidades (Cicerón, 1987: 70 [I § 32-33]). En ambos casos el conjunto de oportunidades disponibles para el agente se ve restringido: si la fuerza de las cosas impone restricciones físicas, los deberes establecen restricciones normativas. Ya se vio con anterioridad que estos dos tipos polares de situaciones de decisión configuran los respectivos escenarios en los que la economía y la sociología colocan el tipo ideal de agente *–actor–* económico o social.

Como es evidente que lo que existe en la realidad es una infinidad de situaciones intermedias en las que las características de ambas se mezclan en proporción diversa, ambas situaciones de elección se oponen polarmente sólo como tipos puros o ideales. Por una parte, es cierto que las preferencias puramente personales, en el límite, equivaldrían a meros gustos –como entre dos helados de distinto sabor– que no se podrían justificar sino aceptar como dadas y explicar tan sólo en términos de causalidad psicológica. En pura teoría, el mercado operaría únicamente con preferencias de este tipo. Pero no es menos cierto que en la génesis de las preferencias, los deseos, los intereses y hasta de los mismos gustos intervienen factores simbólicos y culturales, que son en última instancia normativos.

Por otra parte, sin embargo, no podría entenderse la conducta de quien renuncia a ciertas preferencias en aras de alguna obligación o deber si no se supone que de una u otra forma quiere, desea o prefiere hacerlo –éste es precisamente el escollo con que tro-

pieza la negativa kantiana a admitir que sólo las inclinaciones sensibles pueden mover a la acción, y que pretende sortear por medio del concepto de interés de la razón (Kant, I., 1994: 151). Una situación normativamente definida puede dejar la decisión al libre arbitrio del agente: cuando a éste se le reconoce el derecho a hacer y a no hacer, la situación se vuelve indiferente desde el punto de vista normativo y puede ser decidida por las preferencias no-normativas del agente. Y, a la inversa, entre las puras preferencias del agente puede figurar la de ajustarse a lo normativamente dispuesto: en la "personalidad autoritaria" estudiada por Adorno se revelaría una predisposición conformista. Más aún: el agente podría tener *meta*preferencias, esto es, preferencias de segundo orden sobre sus preferencias de primer orden, que podrían ejercer sobre éstas una influencia normativa, como es el caso del que, insatisfecho consigo mismo, quiere dejar de ser como es. Todo ello, como se verá, plantea arduas cuestiones sobre la compleja naturaleza de las preferencias personales. En todo caso, la decisiva cuestión en torno a la metodología –individualista u holista– específica de las ciencias morales remite en última instancia a la articulación de ambas dimensiones –la preferencial y la normativa– en la unidad de la acción humana.

Sea de ello lo que fuere, lo cierto es que en ninguna de estas situaciones la solución al problema práctico de decidir le viene dada al agente de manera automática. Ya se ha visto que es impensable que éste pueda literalmente "abandonarse a sus impulsos" o "dejar que los hechos decidan por él", o ver sus acciones como meros acontecimientos que le sobrevienen: la perspectiva del agente se muestra tenazmente irreductible a la del espectador.

La respuesta a la pregunta práctica se alcanza al final de un proceso discursivo y deliberativo, real o virtual, que puede expresarse en proposiciones significativas y que, a través de la elección y la decisión, culmina en la propia acción. En cada una de las situaciones típicas antes descritas se recurre a ciertos tipos específicos de proposiciones.

En el caso de la elección basada en las puras preferencias del agente, decidir implica, lógica y prácticamente, poder discriminarlas, diferenciarlas y compararlas entre sí. La propia etimología de preferir –anteponer– menciona la posibilidad de juzgar de

su valor relativo. Las proposiciones empleadas en este proceso deliberativo contienen conceptos valorativos o axiológicos y pueden expresarse lógicamente en forma de comparaciones del tipo "X es mejor que Y" o "X es al menos tan bueno como Y", cuyo análisis formal corresponde a la lógica de la preferencia.

En aquellas otras situaciones en las que el agente reconoce que su abanico de preferencias viables se ve restringido por el peso de ciertas consideraciones normativas, las respuestas a la pregunta por lo que hacer se formulan en proposiciones que contienen expresiones y conceptos típicamente normativos tales como "obligatorio", "prohibido" o "permitido", que son precisamente los functores a los que la lógica deóntica atribuye la función de crear normas. Éste es el caso de aquellas actividades o prácticas gobernadas por reglas, entre las cuales ocupan un lugar central los *juegos.*

2.4. Juegos: reglas

Los modelos que la filosofía y las ciencias en general proyectan modelos sobre un determinado dominio de la realidad para comprender su estructura y funcionamiento muestran cómo ciertos elementos formales de la proyección permanecen constantes a través de ciertas transformaciones. Las tipologías antes mencionadas ponen de manifiesto precisamente las diferencias formales entre diversas situaciones de elección. Los aspectos formales de la conducta humana han suscitado en el presente siglo el interés de teóricos sociales como Georg Simmel (1858-1918), Vilfredo Pareto (1843-1923) o Kurt Lewin (1890-1947) (Caplow, T., 1974: 11-35). Pero ha sido en las últimas décadas cuando se ha visto reforzado, gracias al explícito reconocimiento del valor de los *juegos* como modelos formales de otros tipos de actividades humanas tan dispares como el uso del lenguaje, para el filósofo Ludwig Wittgenstein (1889-1951), o el comportamiento económico, para el matemático John von Neumann (1903-1957) y el economista Oskar Morgenstern (1902-1977).

El historiador neerlandés Johan Huizinga (1872-1945) fue de los primeros en llamar la atención sobre la función del juego

en la cultura en su obra llamada precisamente *Homo ludens* (1938). Huizinga entendía por juego "toda acción u ocupación libre que se desarrolla dentro de límites temporales y espaciales determinados, según reglas absolutamente obligatorias, aunque libremente aceptadas, acción que tiene su fin en sí misma y va acompañada de un sentimiento de tensión y alegría y de la conciencia de 'ser de otro modo' que en la vida corriente" (Huizinga, J., 1998: 43-44). El juego "*crea* orden, *es* orden": crea un mundo temporal dentro del mundo habitual, y dentro de ese mundo creado existe un orden propio y absoluto (íd.: 23). Las reglas de cada juego determinan lo que ha de valer en ese mundo provisional; son obligatorias y no permiten duda alguna: frente a ellas no cabe ningún escepticismo (íd.: 24). En su aspecto formal el juego es "una acción libre ejecutada 'como si' y sentida como situada fuera de la vida corriente" (íd.: 26, 33).

La referencia misma a las reglas implica ya algún grado de organización, en contraste con el espontáneo jugar de los niños pequeños o el retozar de los cachorros de animales. Más aún: en el caso de los juegos organizados como el ajedrez, la esgrima o el fútbol– son las propias reglas las que *instituyen* el juego –ya que se trata de una actividad artificial que no *existía* antes de las reglas– y lo *constituyen* –porque jugar a ese juego *consiste* en actuar conforme a las reglas (Searle, J., 1994: 42-51). Las reglas *constitutivas* definen al mismo tiempo *qué* cuenta como ganar el juego –dar jaque al rey, introducir más balones en la portería que el rival, etc.– y *cómo* conseguirlo de modo legítimo, es decir, formalmente de acuerdo con las reglas –moviendo las fichas o impulsando el balón de determinada manera–. Hacerlo de otra manera, o bien anula la jugada *dentro* de ese juego, o bien convierte ese juego en *otro* distinto. Entrar en el juego –*in-ludere*– es compartir la *ilusión* en la que el juego consiste. Por eso cabe distinguir entre el *tramposo*, que "hace como que juega y reconoce, por lo menos en apariencia, el círculo mágico del juego", y el *aguafiestas (Spielverderber)*, que al "no entrar en el juego" lo deshace, arrebatándole precisamente la *inlusio* (Huizinga, J., 1998: 24). Cuenta la tradición que el fútbol *rugby* nació en 1823 cuando, durante un partido de fútbol en la localidad inglesa de Rugby, un jugador tomó el balón en sus manos y corrió con él, violando las reglas del jue-

go. Pero la infracción ganó adeptos y dio lugar a un *nuevo* juego de pelota que permitía emplear la mano además del pie. La práctica de jugar al fútbol, como la de cualquier tipo de juego, no puede siquiera *describirse* sin mencionar las reglas en las que consiste.

Los juegos organizados son un caso de *convención explícita* –como por cierto también lo son los *contratos*– pero no son el único tipo de conductas reguladas o normativas regidas por convenciones. Existen otras prácticas humanas que en su desarrollo se ajustan más o menos explícitamente a reglas, aunque no son instituidas ni constituidas por ellas y, por consiguiente, no *consisten* en su seguimiento. En este caso las reglas regulan la práctica desde fuera, son extrínsecas a ella. De enorme importancia para el entendimiento de la racionalidad colectiva son las *convenciones* que emergen como solución espontánea a problemas de coordinación y se aceptan en un principio de forma *implícita* aunque posteriormente se expliciten e incluso codifiquen (Lewis, D., 1969: 5-8; 36-51; 100-107).

Hume considera que las reglas de la justicia y la propiedad son ciertamente artificiales –es decir, convencionales– pero no porque se originen en algún tipo de promesa o contrato deliberados, sino porque resultan de un proceso gradual y espontáneo de coordinación implícita movido por un sentido del interés común: "así, dos hombres tiran de los remos de una barca por convención común, por interés común, sin promesa ni contrato alguno; así, el oro y la plata se hacen medidas de cambio; así, el lenguaje y las palabras son fijados por convención y acuerdo humanos" (Hume, D., 1993: 198 [§125]). Aunque muchas convenciones se originan en promesa y contratos éstos a su vez *presuponen* ya ciertas convenciones: prometer implica el acuerdo previo entre quien hace y quien recibe la promesa para atribuir una específica trascendencia práctica a determinadas palabras pronunciadas en circunstancias muy precisas. El prometer y las promesas son *ya* un *juego,* una *institución,* un sistema de reglas (Hare, R., 1974: 171-187; Searle, J., 1974: 151-170). Por eso no pueden ser el fundamento absoluto de las reglas de la justicia: el motivo para actuar conforme a las reglas de la justicia ha de ser previo a, y distinto de, el deber de acatarlas (Hume, D., 1992: 645 [478-479).

En otros casos las convenciones se adoptan desde el principio de forma *explícita*. Así ocurre, por ejemplo, con las regulaciones de tráfico. La práctica de conducir vehículos responde al propósito de los conductores de alcanzar el objetivo material de trasladarse de un sitio a otro por motivos diversos y por medios materiales más o menos eficaces. Tanto la propia actividad como los objetivos de los conductores y los medios para alcanzarlos existen *in rerum natura* antes, y con absoluta independencia, de las normas de tráfico. A diferencia de los juegos en sentido estricto, es formal y materialmente posible *describirlos* sin mencionar para nada las reglas de tráfico. La regulación responde más bien al objetivo del legislador de lograr que la actividad se desarrolle de forma eficiente y compatible con los intereses de todos cuando la mera coordinación espontánea resulta insuficiente para conseguirlo. La utilidad del reglamento de tráfico es función de su acatamiento por todos o una parte significativa del conjunto de los conductores. Sólo a través de la eficacia del sistema recoge cada conductor individual el beneficio que le compensa del coste de su acatamiento individual. Su aportación personal al sistema en forma de cumplimiento de la norma equivale a la del ladrillo que se añade en la construcción de una bóveda en la que "cada ladrillo por sí solo caería al suelo, y donde la construcción sólo se mantiene gracias a la combinación y asistencia mutua de sus partes correspondientes. Su beneficio únicamente se deriva de la observancia de la regla general" (Hume, D., 1993: 196-197 [§ 124]).

De hecho el utilitarismo –y el consecuencialismo en general– considera que una regla de conducta es útil en la medida en que conduce a la práctica de actos cuyos efectos acumulados son beneficiosos para todos y cada uno de los miembros de la colectividad. El utilitarismo por su propia naturaleza ha de plantearse estos problemas de matemáticas morales, pues parte del supuesto metafísico y moral de que los efectos marginales de las acciones siempre incrementan la utilidad total y de que es posible asociar con cada una de ellas un número real que representa la utilidad de hecho o esperada del estado de cosas producido por ella (Regan, D., 1980: 64-65). Es posible por tanto cometer "errores en las matemáticas morales" si se ignoran los efectos muy pequeños o incluso imperceptibles de las acciones sobre cantidades muy gran-

des de personas (Parfit, D., 1985: 67-86; 1991: 9-19). Pero, por otra parte, si en principio se pueden calcular esos efectos nada excluye que haya casos en los que, desde el punto de vista estricto de la *utilidad*, no es necesario que *todos* cumplan *siempre* las reglas. Otra cosa bien distinta es que lo sea por razones de *equidad*, sobre todo si ésta no pudiera justificarse a su vez por consideraciones de eficacia (Gutiérrez, G., 1990a: 159-170).

Cuando la mera persuasión no basta para conseguir que la práctica se ajuste a las normas, las sanciones que las acompañan buscan disuadir al posible infractor haciendo disminuir la utilidad que puede esperar de su infracción. Pero, en todo caso, para cada conductor sigue siendo formal y materialmente posible conseguir sus objetivos personales –por ejemplo, llegar antes a su destino– al margen de las reglas de tráfico. Éstas *regulan* pero no *constituyen* la práctica de conducir.

Existen asimismo muchas otras prácticas gobernadas por reglas en las cuales, a diferencia de los juegos formales o el tráfico, éstas son seguidas de forma menos explícita y deliberada aunque no menos real. Sociólogos como Ralf Dahrendorf, psicólogos sociales como Erving Goffman, psiquiatras como Eric Berne, antropólogos como Edward Hall y estudiosos de la comunicación no verbal como Ray Birdwhistell han puesto en evidencia cómo muchas conductas humanas responden a un implícito reparto de papeles en una especie de escenario virtual mucho más reglado de lo que podría parecer. Muchas interacciones humanas –¿todas?– se desarrollan en marcos muy estructurados y siguen pautas inequívocamente rituales. Dominar un juego, como hablar una lengua, es haber interiorizado sus reglas en forma de una disposición regular a hacer cosas similares en ocasiones similares o recurrentes sin detenerse a pensarlo. La conducta pautada "contiene" reglas que se dejan captar por quien aprende a jugar viendo a otros jugar o jugando él mismo. De ahí que determinadas conductas individuales y sociales en apariencia inexplicables se iluminen cuando el observador descubre y explicita "a qué juegan" los agentes, que en este caso son literalmente *actores*.

En el sentido más literal el hombre puede ser concebido como un animal que sigue reglas. Lo que parece definir más radicalmente la condición humana es su capacidad para el lenguaje. Pero

el lenguaje natural, que ciertamente *es* una institución, no es sólo *una* institución más, sino la raíz misma de toda posible institución. El concepto de juego revela su gran potencial heurístico cuando Wittgenstein lo aplica a la teoría lingüística que asimila las palabras a herramientas: de igual modo que una barra de hierro únicamente se convierte en una palanca cuando se la usa así, los nombres sólo tienen referencia en el contexto de la proposición según el uso que le asignan determinadas reglas. Su primera mención expresa de la analogía de los lenguajes formales con el juego de ajedrez, que data de 1930, es deudora de Frege y preludia el ulterior desarrollo de la noción general de *juego de lenguaje* en la *Gramática filosófica* y las *Investigaciones filosóficas*. Wittgenstein hace observar que entre los distintos juegos no existe tanto un elemento común cuanto una complicada red de semejanzas. El concepto de juego se extiende como la soga que ata el barco al muelle, que no debe su resistencia a ninguna fibra individual, pues ninguna la recorre de un extremo al otro, sino al trenzado sucesivo de un gran número de ellas. Los diversos tipos de juegos y las reglas que los regulan mantienen entre sí intrincadas relaciones de parentesco.

Existen diversos grados de completud y determinación entre las reglas de distintos tipos de juegos: no son igualmente completas ni rígidas en su aplicación las reglas del lenguaje artificial de un cálculo que las de la gramática del lenguaje natural. Las reglas gramaticales son más flexibles e indeterminadas que las de juegos como el ajedrez o de lenguajes artificiales. Su plasticidad se manifiesta en la textura abierta –*porosa*– de los conceptos del lenguaje natural y fundamenta la crítica que dirige Wittgenstein a quienes conciben el proceso de entender o usar el lenguaje como una operación de cálculo con reglas bien definidas. Pero ni siquiera las reglas de los lenguajes artificiales son completas en sentido estricto. Un "reglamento" completo mencionaría y regularía explícitamente todas las actividades en las que consistiese el juego. Cada caso particular sería *decidible* y no quedaría resquicio alguno para la perplejidad –para *decidir* en el sentido ordinario del término– ya que las reglas, en gráfica frase de Wittgenstein "le taponan todos los agujeros" (Wittgenstein, L., 1988: 104-105 [§ 84]).

2.5. Juegos: estrategias

Al definir los movimientos permitidos, prohibidos y obligatorios del juego las reglas establecen el entorno fijo y constante *dentro* del cual los jugadores pueden tomar las decisiones estratégicas que favorecen sus objetivos tal y como el propio juego los define. Aparte de las funciones básicas de instituir y constituir formalmente los juegos y de fijar sus objetivos y premios, las reglas cumplen asimismo la función eminentemente práctica y material de servir de instrumentos para jugar –y *ganar*– el juego. Desde el momento en que los jugadores deciden libremente aceptar esas reglas como obligatorias, las convenciones artificiales del juego restringen su conjunto de oportunidades de forma no menos rotunda que las limitaciones impuestas por la naturaleza, son *comme des choses.* Pero lo que cuenta como triunfo *es* lo que se consigue del modo especificado por las reglas. Si en el fútbol se tratase únicamente de introducir más balones que el rival, sería poco razonable atenerse a formalismos reglamentarios: existen métodos mucho más eficaces para lograrlo. Como los hay para llegar antes al destino: excediendo los límites legales de velocidad. Sin embargo, hay importantes diferencias en términos de eficacia en el logro de los objetivos. En el fútbol no *sirve* cualquier medio porque no todo *vale:* meter balones en la portería contraria con un comando de operaciones especiales no cuenta como victoria. En la carretera tampoco *vale* todo en términos legales o incluso morales pero, en términos de eficacia, todo *sirve:* el que llega a su destino cuando quería, ha alcanzado sus objetivos. Como en la guerra. Y tal vez también en el amor: el precepto agustiniano *ama et fac quod vis* exhorta a no someter el querer a normas, en el supuesto implícito de que, en cuanto actividad real, el amor ni consiste en convenciones ni necesita de regulaciones para asegurar su eficacia. Como *norma sui* es literalmente infalible: *nec falli nec fallere potest.*

Precisamente el caso de la guerra ilustra ciertas características estructurales de los juegos que han permitido extender el modelo a actividades en las que el propio concepto de reglas parece superfluo. Reducida a puro esquema, la guerra, como el certamen, es el prototipo de un conflicto de intereses –*agón*– directamente con-

trarios en el que cada contendiente –*agonistés*– pugna por hacer valer los propios a costa de los del adversario –*antagonistés*– y en el que, en consecuencia, sólo puede haber un ganador, pues lo que uno gana lo pierde el otro –lo que los teóricos llaman un juego de *suma cero* o *constante*–. De hecho, en las guerras y en otras situaciones reales el asunto es infinitamente más complejo como lo son los propios intereses, por lo que existen juegos tanto de conflicto puro como mixtos de conflicto y cooperación: es posible que todos pierdan –lo que convertiría el juego en uno de *suma negativa*– o que todos en parte pierdan y en parte ganen, de forma que el juego arrojara una *suma positiva*, etc. Los juegos de conflicto puro tienen su lógica propia e inexorable, la *Realpolitik* de la lucha por el triunfo que tan vívidamente describe, entre otros, Hobbes (1989: 105-109 [I, § 13]). En el caso de las guerras *reales* las convenciones de todo tipo introducidas a lo largo de la historia, el recurso a los ideales caballerescos o al sentido del honor, la intervención de organizaciones humanitarias o el testimonio de reporteros y observadores internacionales han cumplido de muy diversas formas la misma función de restringir su desarrollo en términos de pura eficacia estratégica y minimizar así sus nefastas consecuencias (Huizinga, J., 1965: 146-167). Precisamente porque se *pueden* ganar con mayor eficacia desembarazándose de todo escrúpulo. Pero ése no es exactamente el caso de las guerras *simuladas* como son los juegos competitivos.

Éstos acotan un espacio dentro del cual puede desarrollarse el antagonismo de intereses y la competición de forma inocua. Aunque en un sentido muy cierto las reglas definen formalmente los juegos como el ajedrez o el fútbol, la actividad *material* de jugar a esos juegos está inspirada por el propósito de ganar. No se entiende que alguien juegue un juego J^1 sin desear ganar *ese* juego J^1. Se comprende que un padre quiera dejarse ganar al ajedrez por su hijo pequeño, pero *ipso facto* no está jugando *al ajedrez,* sino a *otro* juego diferente, tal vez a un *metajuego* J^2 en el que sus pérdidas en el juego de primer nivel sean sus ganancias en el del segundo –por ejemplo, en términos de beneficios en la formación de su hijo–. *Pace* Coubertin, lo importante es participar, pero a condición de desear ganar. El propósito de ganar es más básico que el de acatar las reglas y se da por supuesto en el

concepto mismo de juego. Los reglamentos del fútbol o de tráfico presuponen jugadores y conductores con motivos e intereses reales y aspiran a que los canalicen a través de sus normas, pero no les muestra el modo más eficaz de satisfacerlos. Las reglas enseñan simplemente a *jugar* el juego, no a *ganarlo*: esto último depende de las estrategias racionales e inteligentes que permiten a cada jugador sacar el mejor partido posible de cada situación y de cada jugada del rival.

No es casualidad que en muy diversas culturas y desde las épocas más remotas se hayan diseñado juegos y certámenes –juegos olímpicos, campeonatos, torneos, regatas, etc.– que, al recrear en una escala manejable situaciones de conflicto real, han cumplido una función catártica para participantes y espectadores, aunque como efecto perverso hayan generado a su vez nuevos motivos de conflicto. Otros juegos han añadido a la función puramente catártica la más teórica de servir de modelos que *simulan* situaciones en las que los jugadores persiguen sus intereses antagónicos mediante estrategias racionales. De China o India procede el más antiguo, el *go,* que parece remontarse al tercer milenio a. de C, y el mejor analizado, el *ajedrez,* que en su forma más o menos actual data del siglo VI; en las academias militares de la Prusia del siglo XVIII se desarrolló el llamado *Kriegspiel* que en las sucesivas guerras del siglo XIX fue obsesivamente analizado por los estados mayores de los ejércitos de media Europa y Estados Unidos hasta bien entrado el XX. Se ha señalado la importancia que tuvo para su fundador, John von Neumann, el conocimiento del *Kriegspiel* en sus años de juventud en Budapest. Aunque "oficialmente la teoría de juegos fue inspirada por el póquer" (Poundstone, W., 1995: 66): el antiguo alumno de Wittgenstein y a la sazón profesor de filosofía moral en la Universidad de Cambridge, Richard Braithwaite (1900-1990), concluía su lección inaugural de 1954 sobre *La teoría de juegos como herramienta para el filósofo moral* con la esperanza de que en el futuro esta disciplina "floreciera al calor de una teoría cuyo prototipo se había alumbrado en torno a las mesas de póquer de Princeton" (Braithwaite, R., 1955: 55).

3
Razón y maximización

3.1. El modelo de la Teoría de Juegos

La condición de agente racional suscita una compleja red de cuestiones fronterizas entre materias tan diversas como la filosofía de la acción y de la mente, la inteligencia artificial, la teoría del conocimiento, la filosofía del lenguaje o la epistemología. Las ciencias morales, por su parte, abordan bajo distintas perspectivas y denominaciones –interacción, intercambio, cooperación, etc.– la acción humana considerada como foco central de una constelación de procesos genéricamente racionales que la anteceden –deliberación, elección, decisión, etc.– y de efectos –previstos o imprevistos, deseados o perversos– que se siguen de ella. Sería muy difícil entender el desarrollo contemporáneo de muchas ciencias sociales, e incluso de una parte importante de la ética, sin tener en cuenta la función catalizadora de disciplinas como la Teoría de Juegos en el entendimiento de la *agencia* humana.

Desde los primeros momentos de su presentación formal por John von Neumann y Oskar Morgenstern en 1944, la Teoría de Juegos puso de manifiesto sus virtualidades para el análisis de la conducta racional. Su modelo de decisión racional encontró rápida acogida en áreas del saber ajenas a los economistas profesionales como la sociología, la ciencia política o la filosofía moral. Esta última en concreto advirtió tempranamente su excepcional valor para analizar y solucionar cuestiones morales no sólo de aplicación práctica, sino de índole teórica. El citado Richard Braithwaite proponía ya en 1954 utilizar la Teoría de Juegos "como herramienta para el filósofo moral", como instrumento práctico,

moralmente neutral, para "aconsejar a personas que se proponen objetivos diferentes sobre la forma de colaborar en tareas comunes para obtener la máxima satisfacción compatible con una distribución equitativa". Poseedor de una amplia formación matemática y abierto a preocupaciones filosóficas marcadamente interdisciplinares, Braithwaite continuaba la tradición de pensadores como Pascal o Edgeworth que aspiraban, en palabras de Condorcet, a "iluminar las ciencias morales y políticas con la antorcha del álgebra".

A lo largo de medio siglo la teoría ha evolucionado hacia niveles de complejidad más refinados y ha extendido su aplicación a áreas de conducta humana –e incluso subhumana (Axelrod, R., 1986: 89-105)– cada vez más alejadas del comportamiento estrictamente económico de sus aplicaciones originales, transformando a su vez profundamente el propio enfoque económico. Buena prueba de ello es que un número apreciable de los Nobel de Economía desde su creación en 1969 hayan premiado la contribución de los galardonados, desde perspectivas diversas pero estrechamente conectadas, a la comprensión de la racionalidad humana en sentido amplio. Lo recibieron Kenneth Arrow (1972) por su estudio de las decisiones colectivas y el bienestar social; Gunnar Myrdal y Friedrich von Hayek (1974) por su análisis de la interdependencia de los fenómenos sociales, institucionales y económicos; Herbert Simon (1978) por su diseño de modelos de racionalidad limitada; James Buchanan (1986) por el desarrollo de la teoría de la elección pública que conecta la teoría política con la económica; Ronald Coase (1991) por aplicar el análisis económico a las relaciones jurídicas; Gary Becker (1992) por extender las posibilidades metodológicas del enfoque económico allende los límites de la ciencia económica en sentido estricto; John Harsanyi, John Nash y Reinhard Selten (1994) por su análisis de los conceptos de negociación, regateo y equilibrio racional; y, el más reciente, Amartya Sen (1998) por su contribución al estudio de las dimensiones y limitaciones éticas de ciertas interpretaciones de la racionalidad económica.

Precisamente John Harsanyi (1920-), al ilustrar en 1976 a un auditorio de especialistas interesados en los problemas de fundamentación de las ciencias "especiales" sobre los avances logrados

hasta entonces en la comprensión de la conducta racional, proponía sistematizar las distintas disciplinas formales de la acción racional en una única Teoría General de la Conducta Racional que incluiría la Teoría de la Utilidad, la Teoría de la Decisión, la Teoría de Juegos y hasta la propia Ética (Harsanyi, J., 1976b: 322-324). En épocas más recientes esta teoría integrada ha terminado siendo conocida como Teoría de la Elección Racional –*Rational Choice Theory*– y se ha desarrollado hasta el punto de presentarse como un auténtico paradigma de la racionalidad práctica.

3.2. El enfoque económico

En cierto sentido la teoría intenta sistematizar y generalizar el punto de vista del ya mencionado "enfoque económico" –*economic approach*– sobre la conducta humana. En sentido estricto no es sino una extensión del modelo del que se sirve la ciencia económica para entender y explicar la conducta de todo agente que, en condiciones bien definidas, elige, bien entre los medios más eficaces para alcanzar fines dados, o bien entre fines alternativos en función de sus preferencias y de las oportunidades disponibles. Max Weber (1864-1920) consideraba que son precisamente las conductas de este tipo las que resultan inteligibles en grado máximo, colmando así la aspiración a la evidencia que caracteriza a toda ciencia. Aunque la noción weberiana de *tipo ideal* plantea complejos problemas epistemológicos, puede decirse de forma simplificada que el tipo ideal del *homo economicus* se construye simplificando, estilizando o acentuando unilateralmente ciertas características de los actores reales que protagonizan las actividades mercantiles como compradores, vendedores, consumidores, empresarios, inversores, etc. La función de límite ideal que cumple el *homo economicus* como modelo de agente perfectamente racional sería análoga a la que desempeña en la física la noción de ausencia de fricción para formular las ecuaciones de la mecánica de fluidos en las condiciones ideales que en el mundo real tienden a cumplir los superfluidos o los superconductores.

El *homo economicus* es por tanto un modelo abstracto que representa los procesos de toma de decisiones de un hipotético agente racional en las condiciones formalmente definidas por la teoría. Para valorar la plausibilidad teórica del modelo ha de tenerse primariamente en cuenta su simplicidad y coherencia internas; las proposiciones deducidas de aquéllas son formalmente válidas con independencia de que existan o no en la realidad agentes que sean racionales en el sentido definido por la teoría. La hidrodinámica sigue siendo una teoría coherente aunque es imposible o altamente improbable que en el mundo físico existan fluidos perfectos. Puede discutirse si los axiomas de la teoría son meras tautologías o si, por el contrario, expresan verdades necesarias (Hollis, M. y Nell, E., 1975: 170-181), pero en la medida en que aspira a integrar los datos que ofrece la experiencia no es posible desentenderse de su grado de realismo. Precisamente por su condición de instrumento de comprensión y explicación científica, esta propiedad es la piedra de toque del rendimiento heurístico y predictivo del modelo, pero al mismo tiempo, como se verá, lo hace acreedor al reproche de reducir toda motivación humana al propio interés.

Lo peculiar del *enfoque económico* asociado a los nombres de James Buchanan (1919-), Gary Becker (1930-) o Ronald Coase (1910-) es la extensión de dicho modelo a actividades que en el sentido habitual del término no pueden considerarse como mercantiles, como las políticas o familiares. Como lo que permite tal extensión es el elemento de decisión individual presente en *toda* actividad humana, parece necesario reformular el propio objeto de estudio de la ciencia económica de forma que incluya actividades ajenas al mercado. Confinar la economía al estudio de los procesos producción, distribución y consumo de riqueza, o de los mecanismos de asignación de bienes materiales –tierra, trabajo, capital, maquinaria, dinero, etc.– para satisfacer necesidades materiales restringe en exceso el potencial campo de aplicación de su metodología. Incluso resulta restrictivo confinarla al estudio del mercado entendido en el sentido técnico, aunque abstracto, de lugar de intercambio de bienes y servicios entre compradores y vendedores.

Si se aísla el elemento verdaderamente común a todas las manifestaciones materiales de la actividad económica es posible preci-

sar en qué consiste lo distintivo y específico del enfoque económico, aun asumiendo el riesgo de incurrir en un cierto formalismo. Toda decisión enfrenta al agente, en el plano objetivo, con la escasez intrínseca de cualquier recurso finito y, en el subjetivo, con la imposibilidad de satisfacer simultáneamente todas sus preferencias. En el proceso de deliberación que conduce a la resolución de cualquier problema práctico es inevitable, por tanto, suponer un momento o situación de elección entre alternativas. Lo característico del enfoque económico es considerar estas acciones en tanto que resultan de la forma muy precisa de deliberar, elegir y decidir que plantea la necesidad de asignar recursos escasos para alcanzar objetivos en competencia.

3.3. Maximización y optimización

Lo que hace racional la elección entre alternativas es la capacidad que posee el agente de *maximizar* una función objetiva definida respecto a variables bien determinadas. Es axiomático en el modelo que, al enfrentarse con una elección, el individuo –ya sea formalmente definido, o interpretado como representativo o promedio– elegirá *más* en vez de *menos*. No es casual que el modelo de agente racional se construya sobre el postulado de la maximización y que el agente racional sea necesariamente concebido como un maximizador, como un programador lineal ideal. La razón hay que buscarla en la estructura conceptual de la propia teoría.

Como toda teoría científica, la teoría de la elección racional se propone como objetivo ofrecer una explicación de alcance universal de la conducta humana. El procedimiento lógico para alcanzar ese objetivo exige considerar las acciones humanas como entidades naturales pertenecientes al único universo natural. En cierto modo puede entendérsela como culminación del proyecto que, desde el primitivo positivismo de Saint-Simon y Comte hasta el positivismo lógico del Círculo de Viena –principalmente Carnap y Neurath– aspira a constituir las ciencias morales como una *física social* en el marco de una ciencia unificada. La unidad de las ciencias se basa en el hecho de que todas ellas comparten el mis-

mo lenguaje, aplican la misma metodología y se basan, en definitiva, en las mismas leyes. Todos los términos verdaderamente científicos podrían reformularse en un conjunto de enunciados básicos o protocolares descriptivos de la experiencia inmediata o, más específicamente, reducirse a términos físicos. La metodología sería asimismo única, puesto que los procedimientos para verificar y fundamentar los enunciados científicos son básicamente los mismos para toda ciencia. Tampoco habría razones fundamentales para mantener la jerarquía piramidal de ciencias físicas, biológicas y sociales, pues en última instancia las leyes propias de cada nivel se deducirían de un mismo conjunto de leyes físicas fundamentales.

La influencia del modelo de la física en la teoría de la elección racional va más allá de una simple analogía retórica o incluso de la adopción del formalismo matemático. De hecho ha ofrecido a la economía un ideal regulativo de la explicación científica, que por extensión permite comprender la atracción que el modelo explicativo de la ciencia económica ha ejercido a su vez sobre el resto de las ciencias morales (Barry, B., 1974: 9-58, 113-142; Becker, G., 1980; Buchanan, J., 1984; Casas, J., 1984; Polinski, A., 1985; Green, D. y Shapiro, I., 1994; Friedman, J. (ed.), 1996). Pero el influjo más decisivo se manifiesta en la aceptación de presupuestos ontológicos como el determinismo, que proporciona el fundamento último de toda genuina explicación, tanto predictiva como retrodictiva, o de conceptos básicos como el de *equilibrio,* que ofrece un marco analítico para integrar todos los elementos de un sistema, y que tan decisiva función ha desempeñado en la teoría económica y en la teoría de juegos.

Para el economista inglés Alfred Marshall (1848-1924) los elementos del sistema económico se mantienen en su conjunto gracias a su contrapeso mutuo y a su acción recíproca, como en el equilibro general de un sistema copernicano. En el modelo matemático global del sistema económico elaborado por Léon Walras (1834-1910) las funciones de la oferta y la demanda para cada producto determinan un punto de equilibrio –análogo al equilibrio de fuerzas en la mecánica clásica– que es el precio satisfactorio para el productor y el consumidor, al menos en un determinado mercado. El consumidor maximiza su función de utili-

dad hasta el límite impuesto por la restricción presupuestaria. Como el equilibrio en cada mercado depende de lo que acontece en otros mercados, es preciso extender la teoría para dar cuenta del equilibrio general, que implica la determinación simultánea de todos los equilibrios parciales. En la Teoría de Juegos desempeña un papel decisivo la demostración por John Nash (1928-) de que existe una solución en equilibrio y única para los juegos cooperativos bipersonales. El "equilibrio de Nash" define un conjunto óptimo –un óptimo paretiano– de decisiones, una para cada jugador, imposible de mejorar para ambos simultáneamente.

Entendida como "física social" la teoría de la elección racional vendría, por tanto, a ser "una variante del proyecto mucho más amplio de análisis del equilibrio que aplicó la mecánica de la energía del siglo XIX, primero a la ingeniería, luego a la microeconomía, luego a la biología, a la investigación operativa y por fin a la política" (Murphy, J., 1996: 157). Un postulado fundamental del análisis del equilibrio establece como principio que los objetos de la investigación se comportan de forma que hacen *máximos* o *mínimos* el valor de ciertas variables. Este *principio de extremos* ha sido a través de los siglos uno de los principales instrumentos conceptuales de los que se ha servido la ciencia para comprender los fenómenos físicos (Hildebrandt, S. y Tromba, A., 1990: 150). Así, por ejemplo, un estado de equilibrio estable se define por hallarse en él la energía potencial en su valor mínimo respecto de cualquier configuración alternativa. Pero éste es un caso especial de aplicación del principio de entropía creciente, que enuncia en forma de principio de extremos la segunda ley de la termodinámica: el estado de equilibrio estable de un sistema físico aislado es aquel en el que la entropía alcanza el máximo valor posible.

Son numerosas las aplicaciones de tales principios en el campo de la física: Fermat (1601-1665) formula el principio según el cual la luz sigue la trayectoria más rápida entre dos puntos fijos en un conjunto de lentes o espejos; Hamilton (1805-1865) aquel que establece que una partícula en movimiento entre dos puntos sigue la trayectoria que minimiza la acción, etc. En las matemáticas del equilibrio es axiomático que todo movimiento es ópti-

mamente eficiente. Ello se da por sentado hasta tal punto que, si los datos empíricos de una variable divergen de las predicciones de la teoría, no se concluye que el sistema no esté de hecho maximizando el valor de dicha variable, sino que se han especificado incorrectamente las restricciones con las que opera. El principio es sumamente resistente –inmune incluso– a la revisión hasta por razones pragmáticas: el coste de oportunidad que implicaría su abandono podría resultar inasumible.

Existe una teleología latente en la idea misma de una naturaleza regida por un principio universal de optimización y de economía de medios que se refleja en la concepción leibniziana del mundo existente como solución a un problema de optimización restringida: el de cómo hacer composibles la omnipotencia y la infinita bondad divinas. Maupertuis (1698-1759) formula un esquema global del universo cuya ley general es la de la mínima acción o del mínimo esfuerzo, y según la cual la naturaleza opera siempre con la mayor economía posible. Euler (1707-1783) logra demostrar de modo más riguroso que el principio de mínima acción permite describir el movimiento de una masa puntual en un campo de fuerzas homogéneo y conservativo.

Pero la idea de que todo fenómeno observable obedece la regla según la cual una magnitud se hace máxima o mínima es mucho más que un mero ideal regulativo, pues va acompañada de un poderoso instrumento matemático preciso para describir y hallar objetos óptimos –el *cálculo de variaciones*–. Basado en el cálculo infinitesimal creado por Newton y Leibniz, desarrollado por Jakob (1655-1705) y Johann Bernoulli (1667-1748), Euler y Lagrange (1736-1813), nace originalmente como solución al problema de hallar la forma de la curva que seguiría en su trayectoria una cuenta que cayera libremente y sin fricción por un fino alambre para alcanzar el punto más bajo en el mínimo tiempo. Problema que, generalizado, equivale al de hallar una función para la cual el valor de determinada integral sea el mayor o el menor posible, o al de localizar picos y valles en una cordillera matemáticamente construida. Es precisamente el cálculo de máximos y mínimos el que hace posible describir los equilibrios físicos.

La teleología subyacente de la mecánica clásica se traslada, aparentemente sin mayores problemas, a los agentes racionales:

si las partículas físicas "eligen" trayectorias óptimas, las elecciones de los agentes exigen ser concebidas asimismo como optimizadoras. En los términos de la teoría económica los agentes individuales, como partículas que se mueven en un espacio de bienes, siguen una trayectoria definida por la maximización, sometida a restricciones, de su utilidad. La función teórica del mercado es coordinar y hacer mutuamente consistentes las elecciones de los distintos individuos en forma de un equilibrio óptimo, dado que los precios restringen objetivamente las posibilidades que tienen los agentes de satisfacer sus preferencias. Si se supone, además, que éstas son relativamente estables en el tiempo y suficientemente similares entre los diversos agentes que interactúan, se comprueba que los objetos de elección no son tanto los bienes y servicios *materiales* ofrecidos en el mercado cuanto la *utilidad* de los objetos subyacentes que pueden producirse por medio de ellos. Éstos representan valores básicos de la vida que no siempre guardan una relación unívoca con los bienes y servicios del mercado. Los precios, sean o no monetarios, miden el coste de oportunidad de la asignación de recursos escasos a objetivos alternativos. La eficacia del modelo se mide por su capacidad de predecir un mismo *tipo* de respuesta por parte del agente tanto en situaciones de mercado propiamente dichas como en aquellas otras en las que, por ejemplo, ha de elegir pareja, decidir el número de hijos, establecer prioridades en la resolución de problemas científicos o distribuir el tiempo libre entre formas de ocio alternativas, según su escala de preferencias. Se entiende así la definición informal que Bernard Shaw daba de la ciencia económica: "el arte de sacar el mejor partido de la vida".

El núcleo central de la perspectiva económica sobre las decisiones humanas lo forma, por tanto, una teoría de la *elección* racional, ya que la elección sigue siendo el elemento común a todas ellas. Dicho de otra manera: si la teoría de la elección racional aspira a ofrecer un conocimiento universalmente válido de la acción humana, la ciencia económica es de hecho su rama mejor elaborada (Mises, L. v., 1986: 122). Al aislar el elemento de elección individual presente en *toda* actividad humana y considerarlo como el *locus* en el que se manifiesta la racionalidad práctica, la teoría económica construye un modelo refinado y potente de

decisión racional que se propone capturar la esencia misma de la racionalidad práctica, redefiniendo para ello su propio objeto de estudio para extenderlo a la acción humana en general, sin restringirlo a las actividades características del mercado.

El economista austríaco Ludwig von Mises (1883-1953) ofrece una versión radical de esta manera de entender la ciencia económica. Para él la economía es, en sentido estricto, *la* ciencia de la acción humana o *praxeología* (Mises, L. v., 1986: 125; Kaufmann, A., 1967: *passim;* Gasparski, W. y Pszczolowski, T. (eds.), 1983: *passim*). Y no precisamente por ser una ciencia empírica basada en el método *comprensivo* que caracteriza las ciencias morales, sino por ofrecer una axiomática universal y apriorística de la conducta humana; es, en palabras del propio Mises, *la lógica de la acción y de los hechos.* Como la lógica y la matemática, elabora construcciones formales que permiten captar las categorías fundamentales de la acción, que no proceden de la experiencia sino que son presupuestas por ella. El principio de acción equivale al de causalidad porque como éste, está presente y se da por supuesto en el reconocimiento de una conducta que sea algo más que la mera reacción a un estímulo. En la perspectiva de Mises el hombre es ante todo un agente, un ser que actúa, y por ello no es posible pensarlo sin presuponer la categoría de la acción –según el evangelio praxeológico, *"en el principio era la acción"*– (Mises, L. v., 1986: 124). Pero tampoco puede pensarse esta categoría sin hacer referencia a los prerrequisitos universales de la acción humana. Mises cree incluso posible construir, mediante el método axiomático, una praxeología universal válida para cualquier mundo posible, al modo de la geometría hilbertiana. Los teoremas de la economía no se derivan de la experiencia, sino que se deducen de la categoría fundamental de la acción, que podría formularse con certeza apodíctica como "el principio económico, es decir, la necesidad de economizar" (íd.: 129), versión económica del principio físico de la mínima acción que, a su vez, puede interpretarse como la versión física del principio económico: en realidad ambos son interpretaciones del mismo principio universal.

Toda decisión presupone una elección, y no necesariamente entre diversos bienes materiales. Pero el mero hecho de poder y tener que elegir entre alternativas parece implicar que éstas han

de ser comparables entre sí con arreglo a una medida común: "todos los fines y todos los medios –las aspiraciones espirituales y las materiales, lo sublime y lo despreciable, lo noble y lo vil– ofrécense al hombre a idéntico nivel para que elija prefiriendo unos y repudiando otros. Nada de cuanto los hombres ansían o repugnan queda fuera de esta clasificación en escala única de grado y preferencia" (Mises, L. v., 1968: 25-26). La *completitud* es el requisito formal de un conjunto de preferencias que permite su ordenación, es decir, la jerarquización de todas las alternativas. Por ejemplo, la relación "al menos tan buena como" debe ser completa: para cualquier par de alternativas x e y, o x es al menos tan buena como y, o y es al menos tan buena como x (o ambas): un individuo con una relación de preferencia completa conoce su decisión a la hora de elegir entre dos alternativas cualesquiera (Sen, A., 1976: 15). Las preferencias de un agente racional son por definición homogéneas e integran un conjunto completo y consistente que no puede no estar ordenado: todas las alternativas han de ser comparables en términos de una unidad de medida común –la utilidad– formal.

3.4. Utilidad, utilitarismo y optimidad

Pero si el agente es necesariamente concebido como maximizador ¿qué es lo que maximiza? A primera vista podría parecer que requerir la comparabilidad universal de todos los valores implica medirlos todos por el rasero de la utilidad personal substantivamente entendida y dar por sentado que el egoísmo es el único móvil de la conducta racional. Para entender los presupuestos y el verdadero alcance de la teoría es preciso disipar ese malentendido superficial, mostrando cómo el desarrollo de la teoría económica en la Gran Bretaña de la segunda mitad del siglo XIX es concebido, en buena medida, como la solución a un problema de optimización planteado por la filosofía moral utilitarista. No hay que olvidar que para Jeremy Bentham el utilitarismo no es una teoría de la moralidad privada o individual sino una ética para los agentes con responsabilidades públicas: legisladores y gobernantes. Es ante todo un proyecto de reforma radical que

abarca desde la legislación penal y las instituciones penitenciarias hasta la medicina preventiva pública pasando por la justicia gratuita para los pobres o el sufragio femenino, cuya influencia social ulterior sólo es comparable a la de Marx.

La Ilustración británica del siglo XVIII puede ser llamada a justo título tanto *era de la razón* como *era de las pasiones* por su simultánea insistencia en realzar el papel de la razón en el conocimiento teórico o científico pero potenciar en cambio el de las emociones, afectos y sentimientos en la conducta humana. Bentham parte del supuesto, común a teólogos y filósofos escépticos de su siglo –presente ya en el pesimismo jansenista de los fabulistas franceses del *grand siècle* que tanto influyeron en Mandeville– de que los hombres son gobernados por los dos principios soberanos del placer y del dolor. Entiende la felicidad como una suma de placeres y dolores; no sólo sensoriales sino también espirituales, ni únicamente centrados en uno mismo, pues por asociación de ideas es posible simpatizar con los placeres ajenos. Por eso, el que los hombres sólo actúen por interés no implica *per se* que nada más se interesen por sí mismos, pues es habitual que les interese la felicidad ajena y, en casos excepcionales, son incluso capaces de experimentar una benevolencia universal que les lleva a interesarse por la felicidad general de la humanidad –*la mayor felicidad del mayor número,* en la conocida formulación atribuida a Joseph Priestley, a Francis Hutcheson y a Cesare Beccaria, virtud de los liberales que al Conde Mosca en *La cartuja de Parma* le parecía una inocentada.

El Principio de Utilidad aprueba las acciones en la medida en que contribuyen a aumentar la felicidad de los agentes implicados, entendiendo por *utilidad,* de manera laxa, la propiedad que posee cualquier objeto de producir beneficio, ventaja, placer, bien o felicidad –todo lo cual, según palabras de Bentham, "viene a ser lo mismo"–. Esa propiedad puede ser medida y sumada de forma unitaria, y el primer cometido de la razón es proporcionar la técnica que permita comparar entre sí las pasiones de cada individuo y las de los distintos individuos de un conjunto para permitir la elección óptima: por eso todo problema de elección racional es un ejercicio de maximización (Hollis, M. y Sugden, R., 1993: 5). Bentham llamó a esa técnica de varias formas: "termómetro moral",

"aritmética moral" o "cálculo felicífico". Mucho se ha escrito sobre la naturaleza ordinal o cardinal de tales mediciones y sobre la posibilidad misma de realizar comparaciones interpersonales de utilidad, pero aunque no logró elaborar una escala numérica satisfactoria, Bentham jamás renunció a su convicción de que la utilidad era, por principio, mensurable (Viner, J., 1973: 195-221).

Al mismo tiempo, su insistencia en la eficacia de las acciones para *producir* los mejores resultados revela que la moral utilitaria es una de las interpretaciones posibles del principio ético del consecuencialismo, que hace depender la corrección moral de una acción de sus buenos o malos efectos. Como teoría esencialmente *optimífica* propone como objetivo último de la acción alcanzar los mejores resultados posibles e impone a cada agente la obligación de hacer todo lo que produzca un resultado mejor. Si, por tanto, alguien hace lo que cree que producirá un resultado peor, actúa de forma objetivamente inmoral, pues produce una cantidad de bien menor de la que está a su alcance y es axiomáticamente verdadero que el bien es moralmente preferible al mal (Parfit, D., 1984: 24). Negarlo parecería implicar que la moralidad prohibe al agente hacer todo el bien posible o le manda hacer menos del que sería capaz (Scheffler, S., 1988: 1).

Como no es posible afirmar, sin incurrir en una regresión infinita, que todo lo moralmente valioso lo es en virtud de sus consecuencias, es lógicamente necesario afirmar que hay "estados de cosas" que poseen valor intrínseco y que, como metas finales y productos terminales de la actividad racional, no se ordenan a la producción ulterior de ningún otro estado de cosas. Para el utilitarismo clásico lo único absolutamente bueno o deseable –objetivo, por tanto, de todo agente racional– es la máxima felicidad universal. Lo que hay de formalmente consecuencialista en esta propuesta no es el contenido de lo que se ha de maximizar –la felicidad– sino la exigencia misma de hacer máximo un valor total para un conjunto máximo de individuos. Esta indeterminación formal es la que ha permitido a la teoría de la elección racional desprenderse de las interpretaciones hedonistas del utilitarismo benthamiano y evolucionar hacia un análisis cuasi lógico de las decisiones sin alterar el esquema consecuencialista básico de la acción humana.

Pero en Bentham es el hedonismo psicológico en concreto el que proporciona la hipótesis psicológica que permite elaborar una teoría maximizadora de la utilidad: al actuar, los individuos maximizan lo que se les presenta como su interés. Y aquí aparece el segundo cometido de la razón en relación con las pasiones, que consiste en modificar inteligentemente la función de utilidad de los individuos. Dentro de la lógica de esta hipótesis, de la que a pesar de todo no logra prescindir, se explica que Bentham asigne al legislador y al gobernante la tarea de coaccionar o incentivar a los agentes para que, aun por motivos egoístas, actúen de hecho en consonancia con el principio utilitarista –propuesta que evoca el diestro manejo por los gobernantes del egoísmo humano a que aludía Mandeville casi un siglo antes.

Se debe a Henry Sidgwick (1838-1900), antecesor, por cierto, de Richard Braithwaite en la Cátedra Knightbridge de Filosofía Moral en la Universidad de Cambridge, la primera formulación sistemática de la ética utilitarista, fundamentalmente en su obra *The Methods of Ethics*, publicada en 1874. Por "método de la ética" entiende Sidgwick "cualquier procedimiento que permita determinar qué 'deben' hacer los seres humanos individuales, o qué es 'correcto' que hagan o busquen realizar mediante la acción voluntaria". De los tres métodos que analiza –el egoísmo hedonista, el utilitarismo y el intuicionismo– el primero adopta la máxima felicidad propia como fin último de las acciones de cada individuo. El supuesto básico, no sólo del método empírico del hedonismo egoísta, sino de la propia concepción de "máxima felicidad" es la conmensurabilidad cuantitativa de los placeres y los dolores que permite disponerlos en una escala, condición necesaria para que puedan entrar como elementos de un conjunto que se busca hacer lo mayor posible.

Por la lógica interna de su filosofía utilitarista, los intereses intelectuales de Sidgwick no se limitaban a la filosofía moral, sino que se extendían a cuestiones fundamentales de la economía política y de la metodología de la ciencia económica, sobre las cuales publica sendos tratados en 1883 y 1885. Análogas razones llevaron a numerosos economistas de la época a interesarse asimismo por las cuestiones éticas fundamentales, planteadas en términos utilitaristas. En 1876 el economista y estadístico irlandés Francis

Edgeworth (1852-1926), que siempre tuvo a Sidgwick por autoridad indudable en cuestiones morales, publica su primera obra, una breve nota en la revista *Mind,* en donde introduce métodos matemáticos avanzados con el propósito de examinar las implicaciones de la maximización utilitarista para la distribución *óptima* de los ingresos y de la felicidad. Acepta el supuesto, enunciado entre otros por Joseph Butler (1692-1752) o David Hume (1711-1776) de que deber e interés coinciden perfectamente, en gran parte en este mundo, pero de forma plena y perfecta si se tiene en cuenta el futuro y se adopta una perspectiva global y, en consecuencia, de que egoísmo y utilitarismo pueden ser subsumidos en un mismo principio.

En 1877, sólo tres años después de aparecer la obra de Sidgwick publica Edgeworth *New and Old Methods of Ethics,* en la que, como contribución al mismo problema de hallar la óptima distribución de la utilidad, intenta aclarar el significado de "máxima felicidad" en términos de la función de utilidad sirviéndose para ello del cálculo de variaciones. En su obra más conocida, que lleva el revelador título de *Mathematical Psychics* (1881), insiste en la analogía de la economía con las ciencias físicas, debida al papel que desempeña en ambas el principio de la mínima acción. Este principio máximo, único y bifronte, constituye el pináculo supremo de ambas ciencias así como de la ciencia moral. Edgeworth desarrolla sus implicaciones de mediante los instrumentos matemáticos de la función de utilidad, de las curvas de indiferencia y de contrato, de capital importancia para el análisis económico de la elección racional en términos de maximización y equilibrio.

3.5. Lo sustantivo y lo formal en la elección racional

Para Edgeworth, el supuesto básico de la ciencia económica es que todo agente obra por interés propio. El supuesto no es realista si se entiende como una afirmación de que el único motivo de cualquier acción es el egoísmo, y menos aún si se lo concibe de forma substantiva como hedonismo psicológico. Tampoco lo es si se define el interés propio como interés exclusivo por sí mis-

mo. En este último sentido, en cambio, sí puede darse por descontado que en determinado tipo de situaciones –la guerra, el mercado, las relaciones contractuales– no es razonable suponer que los agentes actúen motivados por algo distinto del propio interés; estas situaciones específicas permiten y hasta exigen que el individuo base sus decisiones en un estricto cálculo maximizador de su personal utilidad.

Aunque el *homo economicus* es habitualmente concebido como un agente motivado únicamente por su propio interés y, a su vez, este interés es entendido de forma substantiva como interés egoísta por uno mismo, lo cierto es que la teoría ha evolucionado hacia una progresiva formalización de los procesos de decisión –una verdadera lógica de la elección– con abandono de presupuestos substantivos de índole psicológica. Entre los primeros en avanzar en esa dirección se cuenta Vilfredo Pareto, que se vale de la noción de curva de indiferencia introducida por Edgeworth para representar la *elección* a la que se enfrenta un agente que es indiferente entre diversas combinaciones de dos bienes cualesquiera. Elegir una combinación implica trocar –*substituir*– unidades de uno de los bienes por unidades del otro, costear con la disminución de uno el incremento del otro. La forma clásica de la curva, descendente de izquierda a derecha y convexa hacia el origen de coordenadas, refleja simplemente la tasa de substitución marginal entre ambos bienes, la medida de la disposición del agente a asumir el coste del trueque, sin que sea ya necesario referirse al placer que produce o al deseo que se experimenta por ninguno de ellos.

Si las preferencias del agente forman un conjunto ordenado y completo podrán representarse en un mapa en el que las curvas de indiferencia indicarán los diversos niveles en un gradiente de preferibilidad, utilidad o –con el término que Pareto propone para evitar las indeseables connotaciones de este último– *ofelimidad.*

La función matemática que asigna índices numéricos a cada curva de indiferencia representa las preferencias de la persona en un índice de utilidad, de forma que la elección del agente, según sus preferencias y dentro de las limitaciones que le impone la situación real –el conjunto de *oportunidades* representado en el mapa por la recta de presupuesto– puede describirse como maxi-

mizadora de su utilidad. La elección racional, en efecto, ocurre en el punto en el que la curva de indiferencia *más alta* es tangente a la recta de presupuesto que representa su techo de posibilidades. Pareto considera que la teoría económica, al deducir sus resultados de la experiencia sin necesidad de introducir ninguna entidad metafísica, adquiere el rigor de la mecánica racional.

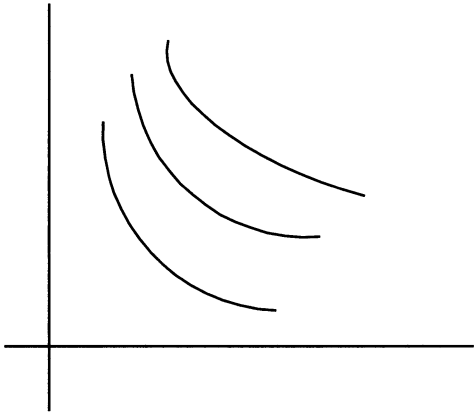


Figura 3.1. Mapa de indiferencia.

La teoría no parece, por tanto, necesitar de ninguna interpretación substantiva de los deseos del agente. El contenido concreto de los deseos, *lo que* de hecho se desea, aportaría simplemente el material sobre el que se ejerce la elección formal entre alternativas parcialmente excluyentes y diversamente preferidas, pero quedaría como tal fuera del modelo. Puede hablarse entonces de dos interpretaciones distintas de la racionalidad del *homo economicus*. En un primer sentido –que suele denominarse *broad* (Elster, J., 1987) o *thick* (Friedman, J., 1996b) esto es, "amplio" o "espeso"– el agente es racional en la medida en que sus preferencias son substantivamente egoístas y poseen además la propiedad formal de ser completas, ordenadas y coherentes. En un segundo sentido –*thin*, "delgado" o "tenue"– lo es por el simple hecho de que sus preferencias poseen tal propiedad formal, con independencia de que sean egoístas o altruistas. Las preferencias

en sí mismas no son racionales o irracionales: lo que es racional es su consistencia en un conjunto. Aunque no responde estrictamente a esta distinción es imposible no recordar que Hume no considera contrario a la razón preferir "la destrucción del mundo entero a sufrir un rasguño en el dedo, [ni...] la ruina total con tal de evitar el menor sufrimiento a un indio o a cualquier persona totalmente desconocida, [ni...] un bien pequeño, aunque se reconozca menor, a otro mayor, y tener una afección más ardiente por el primero que por el segundo. Un bien trivial puede, por ciertas circunstancias, producir un deseo superior al surgido del goce más intenso y valioso; no hay en esto nada más extraordinario que lo hay en mecánica, cuando vemos que una libra de peso levanta otras cien gracias a lo ventajoso de su situación" (Hume, D., 1992: 562-563 [416]).

La teoría de la elección racional entendida en este segundo sentido caracteriza al agente con independencia del entorno: el agente tiene preferencias y se caracteriza por ellas: a los efectos de la teoría el agente *es* su función de utilidad. El dominio de sus preferencias es el conjunto de *consecuencias* de sus posibles acciones: dado el conjunto de acciones disponibles, el agente elige racionalmente si no hay ninguna otra acción disponible cuya consecuencia prefiera a la de la acción elegida (Hahn, F. y Hollis, M., 1986: 14).

Al suponer tan sólo que los agentes persiguen sus fines, cualesquiera sean, mediante un mismo tipo de comportamiento estratégico y maximizador, el modelo permite en principio cualquier interpretación material. Daría lo mismo que las preferencias fuesen de hecho las de un santo o las de un pecador. El agente racional, en cuanto maximizador, elegirá siempre la mejor alternativa o, en todo caso, rechazará las alternativas inferiores. La *ratio boni* bajo la cual actúa necesariamente todo agente racional es entendida aquí como *ratio optimi* porque resulta de la *comparación* entre alternativas y la elección de la mejor. Un altruista racional descartará las opciones inferiores para lograr sus propósitos desinteresados, no menos que un egoísta para obtener los suyos. Al cristiano se le exhorta a ser perfecto; el hombre hobbesiano se ve obligado a maximizar su poder en cada elección; para Moore el deber de todo agente es preferir lo mejor a lo peor si tiene que

escoger entre un par de acciones de las cuales éstos fueran los únicos efectos (Moore, G., 1989: 119). Sin embargo, como habrá ocasión de ver, los dilemas y paradojas a que se enfrenta la elección racional en este modelo no dependen precisamente del contenido de las preferencias, sino "del hecho mismo de preferir y dar prioridad, siendo igual que se dé a uno mismo o a otros" (Parfit, D., 1991: 33, 37).

La distinción entre las interpretaciones *thick* y *thin* tiene importantes consecuencias, no sólo para la definición del concepto mismo de racionalidad, sino para las aplicaciones tanto normativas como explicativas del modelo. Así, por ejemplo, una de las más importantes extensiones del enfoque económico al análisis de las decisiones colectivas, la llamada Escuela de Virginia de la *Elección Pública –Public Choice–* parte del supuesto de que los agentes públicos tienen la misma propensión que los agentes privados a perseguir sus propios intereses materiales, lo que permitiría explicar y predecir su conducta con más éxito que si se los supone movidos por un supuesto interés público. En principio la hipótesis valdría asimismo para los agentes que participan en la política "privadamente", por ejemplo, votando en las elecciones. La teoría demuestra que un agente racional –más aún si es egoísta– carece de motivos racionales para votar en comicios en los que es improbable que su voto sea decisivo y, por tanto, predice que no votará. Si de hecho vota, se sigue lógicamente, o bien que su conducta es irracional pues no maximiza su utilidad, o bien que no se han especificado suficientemente las condiciones reales en las que se produce, lo que ha impedido poner de manifiesto su racionalidad latente.

En todo caso hay que recurrir a explicaciones que no impliquen renunciar a la propia definición de conducta racional, aduciendo, por ejemplo, que el agente valora la eficacia instrumental del acto mismo de votar no sólo en tanto que asegura los resultados de la votación, sino en cuanto refuerza la propia institución democrática del voto ciudadano; o que, más que su eficacia, valora su capacidad simbólica o expresiva de un sentimiento cívico, del deber, o de pertenencia a la comunidad; sentimiento que, en su versión más crudamente conductista, se interpreta como una comezón que se alivia votando; o que ha sido condi-

cionado por el proceso educativo para responder a ciertos estímulos sociales; o que posee dos conjuntos de preferencias, egoístas y "éticas", y elige estas últimas para el caso de la votación, etc.

Amartya Sen observa con agudeza que una de las razones fundamentales para concebir al hombre como un egoísta interesado es que "siempre es posible definir los intereses de alguien de tal forma que haga lo que haga puede considerarse que promueve sus propios intereses en cada acto aislado de elección" (Sen, A., 1986a: 180). Una conducta en apariencia desconcertante sólo puede atribuirse bien a la inconsistencia o bien a un cambio de las preferencias del agente. Si el agente racional maximiza por definición su utilidad subjetiva y ésta es, también por definición, la medida de sus preferencias de hecho, sólo puede *evitar* maximizar siendo inconsistente (Hollis, M. y Sugden, R., 1993: 6). En efecto, la definición formal de *utilidad* es inevitablemente circular pues, por una parte, designa la cualidad, cualquiera que sea, que hace que algo sea deseable; pero, por otra, el hecho mismo de ser deseado indica que algo posee utilidad (Robinson, J., 1973: 48; Mill, J. S., 1984: 90). En consecuencia, el agente, si es consistente, siempre aparece maximizando su utilidad "en este mundo encantado de definiciones" (Sen, A., 1986a: 181). Como ocurría con el principio físico de la mínima acción mencionado más arriba, el principio económico de la maximización de la utilidad es también sumamente resistente a la refutación, pero lo es al precio de convertirse en cuasi-tautológico. Como hace notar Parfit, cuando se afirma que lo que cada uno hace es, por definición, lo mejor para él –en frase de los economistas, 'maximizará su utilidad' –, como se trata de una mera definición no puede ser falso, pero puede ser irrelevante si de lo que se trata es de los intereses a largo plazo de la persona (Parfit, D., 1991: 18).

Al reproche de tautologismo formulado a la teoría de la elección racional puede responderse distinguiendo entre el contenido de un enunciado particular, como puede ser el de una ley fundamental, dentro de una teoría, que *puede* ser tautológico, y el contenido de la teoría misma –que implica la conjunción de la ley fundamental, de leyes específicas, restricciones, condiciones *ceteris paribus*, etc. –y no *tiene* por qué serlo. La Segunda Ley de Newton es en sentido estricto tautológica, pero la mecánica clá-

sica de partículas no lo es. Son las leyes adicionales tanto generales como específicas las que permiten aplicar la teoría en un dominio especificado. El supuesto del maximizador racional, aun siendo tautológico, cumpliría la misma función que la Segunda Ley: enlazar y sistematizar regularidades observables aisladas. Su contenido empírico lo proporcionaría el supuesto adicional de que los agentes actúan por unos motivos específicos que permanecen relativamente constantes en marcos institucionales muy variados (Diermeier, D., 1996: 67-68).

En cierta forma el riesgo de tautologismo está implícito en la formulación por el economista norteamericano Paul Samuelson (1915-) del concepto de *preferencia revelada* y en la función que le asigna en la teoría del consumidor. Por –y en– el mismo hecho de elegir, los agentes *revelan* sus preferencias. La teoría de la preferencia revelada se construye a partir de un conjunto de axiomas referidos a la elección que se reducen a la ya mencionada exigencia de consistencia entre las elecciones y, en definitiva, entre las preferencias subyacentes a éstas. Así, por ejemplo, en el caso de una elección entre dos alternativas, X e Y, si el agente revela su preferencia por X entonces no puede revelar también una preferencia por Y. En el de una elección entre tres alternativas, X, Y, Z, por el axioma de transitividad si prefiere X a Y, y prefiere Y a Z, entonces no puede preferir también Z a X.

Las preferencias, a su vez, son conocidas únicamente a través de la conducta efectiva –observable– del agente: de acuerdo con la metodología conductista implícitamente aceptada, no hay ninguna manera de conocer las actitudes de alguien respecto a las alternativas que sea independiente de sus elecciones reales. No sólo las preferencias y las actitudes: los principios morales que verdaderamente sostiene una persona sólo podrían ser conocidos a través de su puesta en práctica: "si nos preguntamos a propósito de alguien '¿cuáles son sus principios morales?', la forma más segura de acertar con la respuesta verdadera sería estudiar lo que hizo" (Hare, R., 1975: 13). Si se observa que un agente elige la alternativa X y rechaza la alternativa Y se entiende que *revela* su preferencia por X frente a Y. Su *utilidad* personal no es entonces más que la representación numérica de su preferencia, que asigna una utilidad mayor a la alternativa preferida.

El consumidor es simplemente un caso particular de agente racional que maximiza la utilidad derivada de la adquisición de bienes alternativos dentro de sus posibilidades presupuestarias. Sería poco realista suponer que los consumidores reales calculan en toda elección, de forma deliberada y consciente, su estrategia maximizadora. Pero tampoco la partícula física elige deliberadamente la trayectoria más corta y sin embargo su comportamiento revela que ésa es de hecho la que sigue; es *como si,* metafóricamente hablando, aplicase a su conducta el principio de la mínima acción. De la misma forma, para explicar la conducta humana es útil suponer que el arquero que acierta en el blanco actúa *como si* conociese los principios de la balística. Y, en el caso del consumidor, dar por supuesto que actúa *como si* fuera un maximizador racional ayuda a predecir una gran variedad de fenómenos del mercado. De hecho, sin el supuesto de la maximización muchos de los conceptos más básicos de la economía como los de coste de oportunidad, recta de presupuesto o curvas de indiferencia simplemente carecerían de sentido. Si en el mercado los precios revelan las preferencias de los consumidores expresadas mediante votos monetarios, en la política los individuos revelan mediante su voto sus funciones de demanda. En este sentido la afirmación de que los pueblos tienen el gobierno que merecen sería analíticamente verdadera, si se elimina de "merecer" la dimensión moral del "merecer ser feliz" kantiano y se lo entiende meramente como la inevitable consecuencia de "habérselo buscado" con su propia elección.

En el proceso de formalización creciente de la teoría de la elección racional, el análisis lógico de las decisiones en condiciones de riesgo e incertidumbre se ha beneficiado de las aportaciones de la estadística y de la teoría de la probabilidad. Las curvas de indiferencia de Edgeworth y Pareto permitían representar las preferencias de un agente en términos de su disposición a trocar un bien *cierto* por otro. La Teoría de Juegos elaborada por Von Neumann y Morgenstern permite medirlas en términos de su disposición a asumir *riesgos* para satisfacerlas en el caso de elección entre bienes *inciertos.* En este caso la medida de sus preferencias no es la simple utilidad asociada a los resultados ciertos, sino es la utilidad afectada por la probabilidad de los resultados inciertos –la *utilidad esperada.*

En *The Foundations of Statistics* (1951) Leonard Savage propone como conceptos primitivos de la teoría los de preferencia, estados del mundo y consecuencias, y la racionalidad sigue siendo cuestión de consistencia de preferencias. Una *lotería* es una elección entre alternativas de resultados inciertos. Un ejemplo típico es el del participante en un concurso de televisión que ha de elegir entre retirarse con lo ganado o seguir jugando y correr el riesgo de aumentar sus ganancias o perder incluso lo ganado. La *incertidumbre* se define en términos de un conjunto de estados del mundo mutuamente excluyentes –"ganar" y "perder"– de los cuales sólo uno se realizará. Una *consecuencia* es cualquier cosa que le ocurra al agente, por ejemplo, las ganancias o pérdidas asociadas al concurso. Un *acto* es una lista de las posibles consecuencias, una para cada estado del mundo: (A) se retira con lo ganado; (B) continúa y (a) gana o (b) pierde. El agente elige entre los actos (A) y (B). La *preferencia* del agente por los actos se define en términos de su elección entre ellos. Son postulados fundamentales de la teoría de la elección racional: *a)* que el agente posee un conjunto ordenado y completo de preferencias entre todos los actos posibles; *b)* que las preferencias entre cualquier par ordenado de actos están condicionadas a la realización efectiva de sus consecuencias; *c)* que es posible definir la relación de "subjetivamente más probable que" entre acontecimientos en la medida en que el agente *prefiere apostar* por uno en vez de otro.

En tales condiciones es posible demostrar que el agente cuyas preferencias satisfacen tales postulados elegirá *como si* estuviese maximizando su utilidad esperada. Será posible entonces asignar a cada acontecimiento un número entre 0 y 1 que representará las probabilidades subjetivas de dicho acontecimiento. Será igualmente posible asignar a cada consecuencia un número, de tal forma que la ordenación numérica de las *consecuencias* sea la misma que la ordenación de *preferencias* del agente, interpretando tales números como las *utilidades* de las respectivas consecuencias. Y, por último, será posible elegir tales números de forma que, para cada par de *actos,* su ordenación en términos de su *utilidad esperada* se corresponda con su ordenación de preferencias (Hollis, M. y Sugden, R., 1993: 7).

Nada depende del objeto ni del contenido de los deseos del agente. Lo único que se requiere es la consistencia interna de sus preferencias *dadas*, tal como las revelan sus elecciones. Si son consistentes, su conducta puede ser interpretada *como si* respondiera a complejos cálculos utilitarios en los que midiera la utilidad y la probabilidad de los actos entre los que elige para estimar su respectiva utilidad esperada. Pero estas medidas no explican ni justifican las elecciones, sino al contrario: las medidas se derivan de las elecciones *dadas* que las revelan. Si el concursante del ejemplo ha elegido seguir jugando, eso *significa* que ha asignado a ese acto una utilidad esperada mayor que al de retirarse. En la última fase de su proceso de formalización la teoría de la elección racional conserva el aparato matemático de la economía utilitarista pero ha abandonado "no sólo la psicología utilitarista, sino también la teoría de la razón práctica genéricamente humeana en la que estas matemáticas se apoyaban anteriormente: la noción de utilidad ha sido blanqueada de todo contenido psicológico y sólo queda un esquema abstracto de relaciones cuasi-lógicas" (íd.: 7).

4
Decisiones paramétricas

4.1. Deseos y creencias

La situación de elección a la que se enfrenta el agente es el punto en el que desemboca el proceso de deliberación que se ha desarrollado en diversas etapas, próximas unas y remotas otras en el tiempo e incluso en la historia. Las preferencias que de hecho posee son el complejo resultado de una génesis en la que intervienen no sólo su historia individual, sino la de infinitos grupos y colectividades que le han precedido. Es un lugar común del análisis de la subjetividad humana, en el que coinciden autores tan dispares como San Juan de la Cruz, Mandeville, Kant, Nietzsche, Marx o Freud –y cuantos aplican la que Ricoeur llama "hermenéutica de la sospecha"– que los agentes no son siempre ni necesariamente los mejores jueces de lo que hacen, de por qué lo hacen o de las consecuencias, individuales o colectivas, de lo que hacen. Las ciencias de la conducta han puesto de manifiesto los múltiples y refinados mecanismos, conscientes o inconscientes, que modelan y configuran las preferencias: desde la cruda manipulación exógena hasta los oscuros procesos endógenos de racionalización, sublimación, adaptación, auto-engaño, reducción de la disonancia cognitiva, etc.

Siendo ello cierto, a los estrictos efectos de la decisión del agente, todos estos factores se toman no obstante como *dados*. Por hipótesis se hallan ya representados en su función de utilidad. La teoría de la elección racional obedece a la misma lógica del análisis del equilibrio característico de la ciencia física. Para la mecánica clásica la trayectoria de la partícula, al minimizar algu-

na magnitud, revela la optimidad de la naturaleza; la predicción es posible porque abstrae las relaciones causales de la trayectoria previa de la partícula. El cálculo diferencial permite hallar los máximos y mínimos de una función y medir la velocidad y la aceleración de una partícula en un punto dado abstrayendo de su "historia" precedente. De manera análoga, la conducta del agente racional es concebida como maximizadora de su función de utilidad tal y como está dada en el instante de su decisión presente; en él quedan subsumidas toda su historia anterior y toda la información subjetivamente disponible. Desde la perspectiva de su decisión todo comienza en el presente: su historia es sólo futura.

La teoría de la elección racional reconoce la remota inspiración hobbesiana del modelo que considera que el *primum movens* de la acción humana es el deseo o el temor de estados de cosas futuros que el agente puede obtener o evitar (Gauthier, D., 1994: 25). Para Hobbes el hombre es parte de la naturaleza y, como cualquier otro animal, obedece a dos tipos peculiares de movimientos, vitales y voluntarios. Estos últimos dependen de una representación de la imaginación, que es el primer principio interno de todo movimiento voluntario. Este "pequeño principio" de movimiento en el interior del cuerpo –*endeavour*– precede a las acciones visibles y se manifiesta básicamente como deseo y aversión. El hombre llama *bueno* a cualquier objeto de deseo, y *malo* a cualquier objeto de aversión. La conducta voluntaria del hombre, por muy compleja que sea, puede explicarse en definitiva como un intento de obtener lo que desea.

Las mismas cosas suscitan en la mente de forma alternativa deseos y aversiones, esperanzas y temores, y Hobbes llama *deliberación* al proceso de calcularlos que prosigue hasta que, o bien se actúa o bien se concluye que es imposible hacerlo. En la deliberación los apetitos y aversiones se suscitan por la previsión –*foresight*– de las *consecuencias* de la acción sobre la cual deliberamos. No se delibera respecto de cosas pasadas o que se cree imposibles, algo que también Aristóteles daba por descontado (Aristóteles, 1959: 92 [1140a30-b2]); lo prueba el hecho de que al deliberar se emplea el modo subjuntivo, hipotético o condicional que expresa la relación de los supuestos con sus consecuencias –"si se hace esto, enton-

ces se seguirá esto otro" (Hobbes, 1989: 57 [I, § 6]). La acción es el vector resultante de los deseos y de las creencias de un agente que, a su vez, mira exclusivamente hacia el futuro, pues su elección se orienta tan sólo "por las *expectativas* sobre el comportamiento tanto de objetos del mundo exterior como de otros hombres" (Weber, M., 1964: 20). En este modelo de la acción racional el agente, inmerso en la corriente del tiempo, es *stricto sensu* prudente, o lo que es lo mismo, *previsor –prudens* procede de *providens–*, pues la *prudencia* no es más que "una presunción del futuro adquirida por la experiencia del tiempo pasado".

Considerar que el agente racional actúa exclusivamente por razones prudenciales –en la terminología habitual, *forward-looking–* tiene importantes consecuencias para la extensión de la teoría a la moralidad pero también para la consistencia interna de la propia teoría. Así, por ejemplo, en el caso de las promesas y los acuerdos, el mero hecho de haber prometido o suscrito un pacto en el *pasado* no es en sentido estricto una razón para cumplirlo en el *futuro,* salvo que pueda traducirse en razones de utilidad esperada *presente.* Las razones para actuar sólo pueden ser de tipo prudencial o consecuencial pues en otro caso serían incapaces de mover a la acción. Los agentes racionales no son de poco fiar por el mero hecho de ser egoístas o amorales, pero ciertamente no cumplen su palabra por el mero hecho de haberla dado: necesitan razones que puedan representarse en su función de utilidad presente –como pueden ser el malestar producido por la mala conciencia, o la mala reputación resultante.

Como ya se ha visto, el objeto material de los deseos no es directamente relevante para la lógica de la decisión racional. La noción de deseo está relacionada con la de lo *bueno.* Desear algo es tender, ser atraído o inclinarse por *un* objeto *sub ratione boni.* Preferir, sin embargo, presupone comparar al menos *dos* objetos e implica por tanto la noción de *mejor que.* Por eso la teoría –que lo es precisamente de la *elección* racional– ni siquiera emplea el concepto de *deseo* sino que lo sustituye por el de *preferencia,* que permite en principio abstraer de lo que de hecho se desea y atender tan sólo a la necesidad misma de *elegir* entre las alternativas de satisfacción de deseos concurrentes y en conflicto. Weber caracteriza al agente estrictamente *zweckrational* por su capacidad de situar

estos deseos subjetivos "en una escala de urgencias consecuentemente establecida, orientando por ella su acción, de tal manera que, en lo posible, queden satisfechos en el orden de esa escala (principio de la utilidad marginal)" (Weber, M., 1964: 21). La racionalidad no se predica de los deseos concretos sino de la consistencia del conjunto ordenado de las preferencias *entre* ellos. En un sentido más amplio, sin embargo, la racionalidad implica una relación entre las preferencias, las acciones y sus consecuencias.

Cabe describir la acción, intuitiva y trivialmente, como el resultado de lo que se quiere dentro de los límites de lo que se puede; o también, a la inversa, como lo que resulta posible de entre todo lo que se quiere. La teoría de la elección racional formula con mayor precisión técnica esta misma idea cuando sitúa la elección racional en la intersección de dos conjuntos, el de *oportunidades* y el de *preferencias.* Es posible entonces explicar la acción como el resultado final de dos operaciones sucesivas de filtrado. El primer filtro contiene las *restricciones* de toda índole –físicas, económicas, legales, psicológicas– a las que se enfrenta el agente, y su producto es el llamado *conjunto de oportunidades.* Este último término hace referencia a un concepto económico elemental. Cuando es necesario destinar un recurso escaso a producir un determinado bien se prescinde de la oportunidad de producir con él otro bien alternativo. El sacrificio de una determinada alternativa en beneficio de otra representa un coste real –el *coste de oportunidad*–. El segundo filtro contiene el *conjunto de preferencias* entre las alternativas disponibles dentro del conjunto de oportunidades y el mecanismo para seleccionar la alternativa preferida (Elster, J., 1991c: 23-30).

Describir la acción como la resultante de los deseos y las creencias o, más refinadamente, de las preferencias y las oportunidades implica que el agente es capaz de procesar racionalmente la información de dos conjuntos de datos. El primero de ellos contiene los datos *internos* representados por su función de utilidad, es decir, los que se refieren a lo que desea y al orden en que lo desea. Su primera obligación, interesada, es hacer patentes sus preferencias latentes, lo que equivale a ordenarlas –en lenguaje coloquial, "aclararse" sobre lo que realmente quiere, ponerse de acuerdo consigo mismo.

4.2. El conjunto de preferencias

La noción ordinaria de preferencia es compleja pues existen variados tipos de preferencias que plantean distintos problemas teóricos y prácticos. Suele llamarse preferencias *extrínsecas* a las que se basan en razones e *intrínsecas* a las que equivalen a puros gustos: se puede preferir las naranjas a los plátanos por su mayor contenido vitamínico o simplemente porque saben mejor (Arrow, K., 1974: 78-79; Wright, G. v., 1967: 17-18; Hollis, M. y Nell, E., 1975: 132-133; March, J., 1986: 150-165; Sen, A., 1986b: 63; Hollis, M., 1988: 23-24). A los efectos de la teoría moral es importante distinguir entre las preferencias interesadas que *de hecho* tienen los individuos, que pueden ser tanto egoístas como altruistas y son expresadas por su función de utilidad, y las preferencias *hipotéticas* que tendrían si hubieran de juzgar imparcialmente los intereses de todos incluido los propios (Sen, A., 1976: 20; Harsanyi, J., 1976c: 6-7, 15-20; 1997: 60-66). A una necesidad semejante responde la distinción que se establece entre los meros *gustos*, preferencias dirigidas al consumo directo del individuo, y los *valores*, preferencias que incorporan criterios generales de equidad (Arrow, K., 1974: 78). De gran alcance para la teoría de la democracia liberal es la distinción entre preferencias *personales*, cuyo dominio son los bienes u oportunidades de consumo propio –por ejemplo no leer uno literatura pornográfica–, y preferencias *externas*, que se refieren a la asignación de bienes u oportunidades para el consumo de otros –que nadie la lea (Dworkin, R., 1984: c.9; Sen, A., 1986c: 251-262). Y, en general, es posible hablar de preferencias temporales, irrelevantes, distorsionadas, e incluso de preferencias sobre las preferencias o meta-preferencias.

Aunque en todo caso las preferencias son siempre *individuales*, las colectividades toman decisiones conjuntas mediante algún "mecanismo de elección", que formalmente es un procedimiento de agregación de las preferencias individuales en una preferencia *colectiva*, cumpliendo determinadas condiciones. El economista norteamericano Kenneth Arrow (1921-) demostró el teorema según el cual es imposible garantizar que una ordenación de preferencias colectiva entre al menos tres alternativas

corresponda a las ordenaciones individuales de preferencia, satisfaciendo al mismo tiempo determinados supuestos de racionalidad e igualdad (Arrow, K., 1974: 125-149).

Las preferencias exhiben ciertas características lógicas que permiten sistematizarlas en una teoría formal, una *lógica de la preferencia* o de la elección –*proairética*, empleando el feliz término propuesto, aunque con escaso éxito, por Von Wright (1967: 25)–. La comparación entre los estados de cosas x, y, z pertenecientes a un conjunto establece distintas relaciones diádicas o binarias B entre ellos. Estas relaciones pueden satisfacer ciertas propiedades formales para integrar un conjunto ordenado y consistente, en especial las de:

Reflexividad: $x\ B\ x$
Completud: $x \neq y \rightarrow (x\ B\ y) \vee (y\ B\ x)$
Transitividad: $(x\ B\ y)\ \&\ (y\ B\ z) \rightarrow x\ B\ z$
Antisimetría: $(x\ B\ y\ \&\ y\ B\ x \rightarrow x = y$
Asimetría: $x\ B\ y \rightarrow \neg\ (y\ B\ x)$
Simetría: $x\ B\ y \rightarrow y\ B\ x$

Llamando P a la *preferencia estricta* (x es preferible a y), R a la *preferencia débil* (x es al menos tan preferible como y) e I a la *indiferencia* (x es tan preferible como y), se obtienen las siguientes definiciones:

Preferencia estricta: $x\ P\ y \leftrightarrow [(x\ R\ y)\ \&\ \neg(y\ R\ x)]$. Es una ordenación no reflexiva, asimétrica y transitiva.
Preferencia débil: $x\ R\ y \leftrightarrow [(x\ R\ y) \vee (y\ R\ x)]$. Es una ordenación reflexiva, asimétrica y transitiva.
Indiferencia: $x\ I\ y \leftrightarrow [(x\ R\ y)\ \&\ (y\ R\ x)]$. Es una ordenación reflexiva, simétrica y transitiva (Sen, A., 1976: 23).

El objeto de las preferencias son los *estados de cosas,* esto es, configuraciones o estados del mundo específicamente definidas por (o por sus efectos en) los estados –"mentales" en el sentido más amplio del término– de las personas. Preferir la salud a la enfermedad es preferir un estado de cosas en las que el propio agente, u otras personas, están sanos a otro estado de cosas en el que el agente, u otras personas, están enfermos. Los estados de

cosas objeto de las preferencias son el resultado de pares de factores: por una parte, la propia decisión del agente –que depende de él– y, por otra, un estado aleatorio del mundo –que *puede* ser independiente de él–. Para el que juega a la lotería, por ejemplo, el estado de cosas "ganar el premio" se da si y sólo si se conjugan su decisión de apostar por un determinado numero *y* el estado de cosas del mundo en el que sale ese número. Se supone que el conjunto de *estados del mundo* relevantes para una decisión particular forman una lista exhaustiva y mutuamente excluyente, y que uno de ellos es *verdadero* en el momento de la elección, sea o no conocido por el agente (Luce, R. y Raiffa, H., 1957: 276).

4.3. El conjunto de oportunidades

El segundo conjunto contiene los datos del mundo *exterior* al agente; más específicamente: sobre las probabilidades de que se den las condiciones aleatorias de las que depende la efectiva realización de los estados de cosas.

Existen dos tipos de situaciones o entornos de elección radicalmente diferentes según que la probabilidad de que se den esos estados del mundo –y, consecuentemente, los estados de cosas objeto de las preferencias– sea *independiente* o *dependiente* de la elección del agente. Si de una caja que contiene b bolas blancas y n bolas negras se sacan sucesivamente dos bolas (acontecimientos x e y), la probabilidad de que la primera sea blanca (px) es b / (b + n); la probabilidad de sacar a continuación otra blanca, *condicionada* a que se haya sacado previamente una blanca (py|x) es (b – 1) / (b + n – 1). Si, por el contrario, se repone la bola blanca que salió en primer lugar, la probabilidad de sacar una blanca en segundo lugar permanece inalterable (py|x) = py. Dos acontecimientos x e y son, por tanto, estocásticamente *independientes* si la probabilidad de que ocurra y no varía porque (se sepa que) haya ocurrido x.

Una elección en un entorno *paramétrico* se caracteriza porque la probabilidad de que se den las condiciones aleatorias es *independiente* de la decisión del agente pues está ya dada de antemano. El agente ha de considerarla como variable independiente en su proceso de decisión, como una constante que define los pará-

metros fijos dentro de los cuales tiene que decidir. La probabilidad de que salga un determinado número en una lotería es independiente de la decisión de quien apuesta por él. Un entorno *estratégico,* por el contrario, es aquel en el que al menos dos agentes toman sus decisiones en función de las decisiones que cada uno espera que tome el otro. Las razones para emplear el término son claras: es el razonamiento propio del *strategós,* del general que busca adelantarse a la estrategia del general enemigo en las guerras reales o simuladas como el ajedrez. La probabilidad de que uno de los jugadores de una partida de ajedrez mueva el alfil *depende* de (su estimación de) la probabilidad de que el otro jugador decida mover la torre; la de que éste, a su vez, decida tal cosa *depende* asimismo de (su propia estimación de) la probabilidad de que el otro decida mover el alfil, y así sucesivamente. Son precisamente la distinción cualitativa y la radical discontinuidad entre ambos tipos de entornos lo que enfrenta a la racionalidad concebida según el modelo del *homo economicus* con importantes dilemas y paradojas que afectan a su consistencia interna, a su propósito de representar adecuadamente el razonamiento práctico y a su pretensión de ofrecer un fundamento normativo a la moralidad.

4.4. La decisión en condiciones de certeza

En los entornos paramétricos el conocimiento que el decisor racional posee del grado de probabilidad de los factores independientes de su decisión puede encontrarse en cualquier punto entre la certeza absoluta ($p = 1$) y la (cuasi) total incertidumbre ($1 > p > 0$). El caso límite ideal es la decisión en condiciones de *certeza.* En entornos altamente determinísticos y a efectos prácticos el agente puede dar por ciertas las consecuencias de sus actos. La probabilidad de que una moneda lanzada al aire vuelva a caer a tierra (px) es 1; ciertamente caerá y es imposible que no caiga ($p\neg x = 0$). *Ceteris paribus* –y la naturaleza o la experiencia inveterada aseguran que las "otras cosas" se mantienen de hecho "iguales"– la noche sigue al día, las luces se encienden al accionar los interruptores, los camareros sirven las bebidas pedidas y, tarde o temprano, todos muertos. En este caso extremadamente simple,

el decisor da por descontado que su conjunto de oportunidades está compuesto de un único elemento, que permanece constante sea cual sea su decisión; si su conjunto de preferencias está ordenado su elección *no puede ser* otra que la de la alternativa más preferida que es, por definición, la que maximiza su utilidad.

Los elementos que intervienen en el proceso de decisión en condiciones de *certeza* pueden representarse formalmente en las tablas o matrices siguientes, que servirán asimismo para esquematizar en lo sucesivo otras situaciones de decisión.

Los *actos* (A_1, A_2) en el estado del mundo E determinan estados de cosas o *consecuencias* (x, y):

Cuadro 4.1. *Decisiones paramétricas en condiciones de certeza: matriz de consecuencias*

Actos	*Estado E*
A_1	consecuencia x
A_2	consecuencia y

Las consecuencias x,y están determinadas de manera *necesaria y suficiente* por los actos:

Cuadro 4.2. *Decisiones paramétricas en condiciones de certeza: matriz de probabilidades*

Actos	*Probabilidades*
A_1	$p_x = 1$
A_2	$p_y = 1$

Las consecuencias proporcionan al agente utilidades diversas (u_x, u_y), y es el caso que $u_x > u_y$:

Cuadro 4.3. *Decisiones paramétricas en condiciones de certeza: matriz de utilidades*

Actos	*Utilidades*
A_1	u_x
A_2	u_y

Supóngase el caso trivial de un lector racional en una habitación mal iluminada que le impide leer. Contempla por hipótesis dos estados de cosas. En uno (x), la habitación se ilumina. En otro (y), la habitación continúa mal iluminada. Como lector prefiere x a y, es decir que x le proporciona una utilidad mayor que y ($u_x > u_y$). Tiene a su disposición dos acciones: accionar el interruptor (A_1) o no accionarlo (A_2). Sabe que en el estado del mundo E, la probabilidad de que al accionar el interruptor se ilumine la habitación es 1, y la de que al no accionarlo se ilumine es 0.

Dado que la probabilidad de los efectos de sus acciones es por hipótesis la misma, tanto si acciona el interruptor como si no, el agente racional elegirá accionarlo por ser el acto (A_1) que maximiza su utilidad ($u_x > u_y$). En esta circunstancia ésa es la acción máximamente inteligente. Un observador que conozca los parámetros que definen el conjunto de oportunidades del agente y posea alguna hipótesis sobre las preferencias del lector hallará dicha acción máximamente inteligible –es lo que hay que hacer en tales circunstancias–. Si viera que el lector *no* lo acciona, tendría que modificar su hipótesis, o poner en duda la exactitud de sus datos, o suponer que el agente no es racional y que su "acción" en realidad no ha sido tal sino un descuido, un acto fallido, un impulso irreflexivo, un acto reflejo, etc. Si la acción no es inteligible ¿cómo considerar inteligente al agente? Porque para el propio agente ¿qué *significaría* no llevar a cabo la acción que ciertamente le proporcionará el resultado que prefiere? En las condiciones de certeza se despliega por entero la dimensión optimizadora de la racionalidad concebida según el modelo del *homo economicus* o de la programación lineal.

La programación lineal es, en efecto, una técnica matemática que permite elaborar modelos cuantitativos para tomar decisiones óptimas en la planificación empresarial, en la asignación de recursos en la producción industrial o incluso en la prescripción de una dieta médica. En todos estos casos el problema es común: elegir una estrategia (por ejemplo, una determinada dieta) que satisfaga las necesidades de la nutrición al mínimo coste. Un problema típico de programación lineal consta de (1) *actos,* cada uno de los cuales consiste en la elección específica de n números reales, por ejemplo, dietas; (2) las *condiciones de factibilidad,* que son igualdades o desigualdades lineales que restringen los posibles actos, por ejemplo, los mínimos nutricionales; (3) un *índice* asociado a cada acto que es el promedio ponderado de los n números que constituyen el acto, por ejemplo, la función de coste. El problema es hallar un acto que satisfaga (1) y minimice (2) (Luce, R. y Raiffa, H., 1957: 17-19). La programación lineal se sirve para ello de ecuaciones o desigualdades que limitan el problema y permiten hallar el valor óptimo (máximo o mínimo según el caso) de una función objetiva, que constituye su solución.

El supuesto implícito es que el agente racional hace lo que quiere, quiere lo que hace, y hace y quiere lo que considera mejor. Esto es lo que parece poner en cuestión el clásico problema de la *akrasía* –incontinencia, debilidad de la voluntad o flaqueza de ánimo– suscitado por la negativa socrática a reconocer que alguien pueda elegir a sabiendas un bien menor. El *akratés* parece actuar libre y deliberadamente en contra de su mejor juicio. Cree que *debe* –moral o prudencialmente– hacer algo, pero no lo hace. Si el problema que plantea la *akrasía* es real, es decir, si ésta es realmente posible parece inevitable suponer una distinción real y efectiva entre el nivel de las preferencias que se tienen de hecho en el momento de la elección akrásica y el metanivel desde el que se juzgan –y lamentan– las anteriores: el de las preferencias que cabría denominar reales, verdaderas, a largo plazo.

4.5. La decisión en condiciones de riesgo e incertidumbre

La situación se hace más compleja cuando en los parámetros de decisión del agente intervienen variables estocásticas, esto es,

cuando los resultados dependen en mayor o menor medida de factores aleatorios. En el caso de la moneda lanzada al aire, la probabilidad p del acontecimiento x "caer al suelo" es 1, y la del acontecimiento y "permanecer en el aire" es de 1 – p = 0; pero como *tiene* que caer o cara o cruz –descartando, por simplificar, que caiga de canto– y *puede* caer tanto cara (x) como cruz (y), las probabilidades de caer cara (p) o cruz (1 – p) están en algún punto intermedio de forma que $0 < p < 1$.

Hobbes hace notar que hay placeres y pesares que son suscitados por la expectativa de las consecuencias previstas y que reciben diversos nombres según "la opinión que albergan los hombres de conseguir lo que desean". El deseo junto con la expectativa de alcanzar lo deseado es llamado *esperanza;* la aversión acompañada de la previsión del daño procedente del objeto aborrecido es llamada *temor* (Hobbes, T., 1989: 53 [I, § 6]). Espinosa la define como "la alegría inconstante que brota de la idea de una cosa futura o pretérita de cuya efectividad dudamos de algún modo" (Spinoza, B, 1987: 248 [III, prop. LIX, def. XII]). También Hume considera evidente "que el mismo acontecimiento que por su certeza produciría pesar o alegría, suscita siempre temor o esperanza cuando es sólo probable o incierto" (Hume, D., 1992: 592 [439-440]). El temor y la esperanza se combinan en proporciones que varían de forma continua en función del grado de probabilidad que se atribuye al acontecimiento. Un mal, concebido como meramente *posible*, produce a veces temor, sobre todo si es muy grande: "su exigua probabilidad es compensada por su magnitud" [444].

Estas observaciones permiten captar de manera intuitiva las peculiaridades de la decisión en aquellas situaciones que, a diferencia de las de certeza, implican un riesgo y un coste para el agente, y afectan radicalmente el propio concepto de utilidad. Aunque las decisiones arriesgadas lo son porque el agente se halla en la incertidumbre respecto de las consecuencias de sus acciones, la teoría atribuye significados precisos a ambos términos. Son de *riesgo* aquellas situaciones en las que se conocen las probabilidades "objetivas" de los acontecimientos aleatorios –1/2 para cada cara de una moneda, 1/6 para cada cara de un dado, etc.– mientras que se reserva la de *incertidumbre* para aquellas en las que se

desconocen –por ejemplo, las de que llueva dentro de un año en determinada localidad o las de acertar el número de acertantes del conjunto de resultados de una jornada deportiva–. En este segundo tipo de casos el agente asigna probabilidades "subjetivas" –estimaciones subjetivas de probabilidad– dependiendo de los riesgos que esté dispuesto a asumir según su personal disposición adversa o propicia al riesgo. La cual es a su vez otro de los factores *dados* en la situación de elección. Se reafirma con ello la concepción de la racionalidad como coherencia interna o subjetiva de un sistema de deseos y creencias dados.

4.5.1. Riesgo

Es imposible entrar aquí en la discusión de los problemas teóricos que plantea el concepto clásico o laplaceano de probabilidad *objetiva* incluso en sus interpretaciones frecuentistas o propensivistas. Es evidente, sin embargo, que el problema práctico de decidir *hic et nunc* implica las probabilidades "reales" de realización de un acontecimiento o estado de cosas singular, y respecto de ellas el concepto de probabilidad objetiva ofrece todo lo más un enunciado estadístico válido *ceteris paribus* y *ut in pluribus,* magro consuelo para quien arriesga sus utilidades concretas e individuales. De ahí que importantes teóricos de la probabilidad –Frank Ramsey y los ya mencionados Savage, Von Neumann y Morgenstern– hayan adoptado un enfoque subjetivo de la probabilidad que ponga en relación las creencias, preferencias y utilidades concretas de los agentes. Es significativo que estos últimos basen su teoría de la utilidad precisamente "en la idea de que se puede medir la fuerza de las preferencias de una persona por una cosa según los riesgos que esa persona está dispuesta a correr para obtenerla" (Resnik, M., 1998: 153).

La misma intuición de fondo se refleja en refranes del tipo "más vale pájaro en mano que ciento volando". La formulación aparentemente paradójica sólo tiene sentido a condición de suponer, marginalismos al margen, que la utilidad de los pájaros cazados –por ejemplo, para un *Vogelfänger* como el Papageno de *La Flauta Mágica*– es una función lineal de su cantidad: mientras

más mejor. Por lo tanto, que uno valga más que cien sólo puede deberse al hecho de que la utilidad de la unidad de pájaro se ve afectada por la probabilidad de hacerse con ella. El estado de cosas "pájaro en mano" es cierto, y su probabilidad es de 1. El de los "pájaros volando que podrían ser cazados" es incierto, y su probabilidad es por definición < 1. Por ello la utilidad de un "pájaro volando" como tal es siempre y necesariamente menor que la de "pájaro en mano", al multiplicarse por una magnitud menor que 1. No es una utilidad presente y cierta, sino incierta por futura: es una utilidad *esperada.*

La *pasión* de la esperanza introduce en la teoría una noción que permite asociar sistemáticamente utilidad y probabilidad. Pierre Simon de Laplace (1749-1827) afirma que "la probabilidad de los acontecimientos sirve para determinar la esperanza o el temor de las personas interesadas en su existencia". En términos generales por "esperanza" se entiende "la *ventaja* del que espera un bien cualquiera dentro de suposiciones que son sólo probables". En la teoría "del azar" el concepto más preciso de *esperanza matemática* se expresa por "el producto de la suma esperada por la probabilidad de obtenerla", y equivale a la "suma parcial que ha de ser restituida cuando no se quieren correr los riesgos del evento", suponiendo un reparto equitativo en el que a igual probabilidad se tenga un derecho igual sobre la suma esperada (Laplace, P., 1985: 45-46). Equivaldría también al precio "justo" –matemáticamente exacto *y* moralmente equitativo, el ideal de una moral racional– que habría que pagar por correr el *riesgo* de participar en una lotería.

Las *loterías* poseen características formales que las convierten en modelos muy adecuados para representar las decisiones inevitablemente arriesgadas que se adoptan en condiciones paramétricas de incertidumbre, como se verá que ocurre con los *juegos* en el caso de las decisiones estratégicas. La idea general es que, enfrentados a decisiones de este tipo, los agentes pueden revelar sus preferencias y actitudes personales ante el riesgo de forma consistente con determinados axiomas de la elección racional. Ello permite inferir tanto la *función de utilidad* del agente, que mide el *valor* que poseen para él cada uno de los posibles estados de cosas, como su *distribución de probabilidad subjetiva*, que expre-

sa sus *creencias* sobre dichos estados de cosas. La decisión racional –o sea, *óptima*– es la que maximiza su utilidad esperada respecto a la probabilidad subjetiva.

Volviendo al ejemplo de los pájaros: Papageno, por su decisión de abandonar la caza de los cien pájaros volanderos al tener uno seguro en la mano, *revela* que sus preferencias son las expresadas en el conocido refrán. Al reconstruir analíticamente su proceso de elección se ve que el pajarero se enfrenta a *dos* alternativas: 1) conservar el pájaro en mano, que le proporciona una utilidad *cierta,* y 2) soltar el pájaro en mano y continuar la caza de pájaros volando en la esperanza de que la captura de más de cien pájaros con una probabilidad p, en el mejor de los casos, le compense del riesgo de no cazar ninguno con una probabilidad 1 – p en el peor. Existen, por tanto *tres* estados de cosas sometidos a la deliberación del pajarero: en x gana más de cien pájaros; en y gana un pájaro; en z gana cero pájaros. Considerados en sí mismos –esto es, abstrayendo de su probabilidad– son preferidos en ese orden (x P y P z). Papageno ha de elegir entre asegurarse la satisfacción cierta de su *segunda* preferencia o apostar en la lotería con la esperanza de satisfacer la *primera* pero con el temor de lograr sólo la *tercera.* Elegir es inevitablemente costoso: optar por una alternativa supone incurrir al menos en el *coste de oportunidad* de la alternativa desechada.

Si en una situación de elección tal como la descrita las preferencias del pajarero, además de las condiciones formales antes mencionadas, satisfacen la de ser *continuas,* entonces tiene que haber un punto en el que se declare *indiferente* entre y, y una lotería cuyos premios sean x y z. La idea es intuitivamente clara: en el extremo del continuo de la utilidad en el que las probabilidades de ganar la lotería son muy elevadas, al agente le compensará correr el pequeño riesgo asociado con tamaña ganancia; en el otro extremo, en el que las probabilidades tienden a cero y la utilidad es mínima, preferirá ir a lo seguro y no se arriesgará a apostar. Pero al recorrer Papageno el continuo de un extremo a otro tiene que atravesar un punto *a partir del cual* se invierten sus preferencias, mas *en el cual* es *indiferente* entre ambas alternativas. O, lo que es lo mismo, un punto en el que la utilidad cierta que le proporciona el pájaro en mano es igual a la utilidad esperada de la lotería que le ofrece la

probabilidad p de hacerse con n pájaros. Por lo que el refrán revela de sus preferencias, es evidente que n >100 –cien pájaros están aún por debajo del umbral necesario para inclinar la balanza.

Generalizando el ejemplo anterior, y dado que las probabilidades de cada uno de los posibles estados de cosas varían de manera continua entre 1 y 0, puede suponerse arbitrariamente que el punto de indiferencia se encuentra, por ejemplo, en 2/3. Si se asocia, de forma igualmente arbitraria, el número 1 a la alternativa x y el número 0 a la alternativa z, es evidente que habrá que asignar a la alternativa y el número 2/3. En esa situación, la utilidad de la opción cierta es igual a la utilidad esperada de la lotería que ofrece al agente el premio de 1 con probabilidad de 2/3 y el premio de 0 con probabilidad de 1/3:

$$u_x p + u_z p = u_x$$

$$1(2/3) + 0(1/3) = 2/3$$

En este caso los números 1, 2/3 y 0 proporcionan una escala de las utilidades que procuran al agente las distintas alternativas tal como sus preferencias las revelan. Pero cualquier otra tríada numérica contendría la misma información con tal que fuera el resultado de una transformación lineal positiva (x' = ax + b); para el caso concreto, que tuviera la forma

$$a + b;\ 2/3a + b;\ b$$

en la que a sea un número positivo.

Si el agente posee un conjunto ordenado de preferencias entre loterías y si a cada lotería L se asigna un número u(L) de modo que las magnitudes de dichos números reflejen las preferencias, es decir, que

$$u(L)\ P\ u(L') \leftrightarrow L\ P\ L''$$

entonces es posible decir que existe una *función de utilidad u* sobre el conjunto de loterías. Si, además, esta función de utilidad satisface la condición

$$u[pL, (1 - p)L'] = pu(L) + (1 - p)u(L')$$

–esto es, que la utilidad de una lotería es igual a la utilidad esperada de los premios que la integran– para todas las probabilidades p y todas las loterías L y L', puede decirse que la función de utilidad u es *lineal*. Si las preferencias de un agente pueden representarse por una función de utilidad lineal, entonces puede demostrarse que éste actúa *como si fuese un maximizador de los valores de utilidad esperados* (Luce, R. y Raiffa, H., 1957: 21-31). Éste es el resultado del teorema de la maximización de la utilidad esperada.

Las preferencias entre estados de cosas pueden expresarse en términos de preferencias entre las loterías cuyos premios son los estados de cosas. Una importante propiedad que poseen estas últimas es la de *monotonicidad*. Entre dos loterías que ofrecen diversas probabilidades de obtener premios igualmente preferidos es preferible la que ofrece la probabilidad:

$$L, \text{ que ofrece } pu_x, (1 - p)u_z$$

$$L', \text{ que ofrece } p'u_x, (1 - p')u_z$$

$$L \; P \; L' \leftrightarrow p > p' \text{ (condición de las mejores probabilidades)}$$

Igualmente, entre dos loterías L y L' que ofrecen igual probabilidad de obtener premios más o menos preferidos, es preferible aquella que ofrece el premio más preferido:

$$L, \text{ que ofrece } pu_x, (1 - p)u_z$$

$$L', \text{ que ofrece } pu'_x, (1 - p)u'_z$$

$$L \; P \; L' \leftrightarrow u_x > u'_x \text{ (condición de los mejores premios)}$$

Estas condiciones son expresadas por el principio de la "cosa segura": si una lotería L' proporciona a un decisor un premio y a condición de que se produzca un acontecimiento E –que se dé el estado del mundo E–, por ejemplo que determinado caballo

gane la carrera, y otra lotería L le proporciona otro premio x que el decisor prefiere a y si se produce ese mismo acontecimiento, y no hay ninguna otra diferencia entre ambas loterías, entonces el decisor considerará la lotería L al menos tan preferible como L'. A menos que asigne una probabilidad de cero al acontecimiento E, la considerará incluso estrictamente preferible, pues le proporciona el premio más atractivo (Harsanyi, J., 1976b: 320-322).

Son infinitas las situaciones cotidianas en las que los agentes adoptan decisiones arriesgadas. Virtualmente *todas* son de este tipo, pues en sentido estricto incluso las decisiones tomadas en condiciones de certeza son decisiones de riesgo en las que el agente atribuye subjetivamente una probabilidad de cero a uno de los estados de cosas. De forma que las decisiones en condiciones de certeza pueden ser consideradas como un caso especial de las decisiones arriesgadas. El hecho de que en gran parte de las situaciones reales la elección sea entre *dos* o más loterías no altera el esquema formal.

Resumiendo toda la información anterior: los *actos* A_1, A_2 en los estados del mundo E_1, E_2 determinan los estados de cosas o *consecuencias* w, x, y, z:

Cuadro 4.4. *Decisiones paramétricas en condiciones de riesgo: matriz de consecuencias*

	Estados del mundo	
Actos	E_1	E_2
A_1	consecuencia (w)	consecuencia (x)
A_2	consecuencia (y)	consecuencia (z)

(El agente cree que) los estados del mundo se darán con *probabilidades* p y 1 – p:

Cuadro 4.5. *Decisiones paramétricas en condiciones de riesgo: matriz de probabilidades*

	Probabilidades	
Actos	E_1	E_2
A_1	$p_w = p$	$p_x = 1 - p$
A_2	$p_x = p$	$p_z = 1 - p$

Las consecuencias proporcionan al agente diversas *utilidades esperadas* (ue_w, ue_x, ue_y, ue_z):

Cuadro 4.6. *Decisiones paramétricas en condiciones de riesgo: matriz de utilidades esperadas*

Actos	*Utilidades*	
A_1	$ue_w = pu_w$	$ue_x = (1 - p)u_x$
A_2	$ue_y = pu_y$	$ue_z = (1 - p)u_z$

La teoría proporciona al observador un procedimiento para entender la conducta de un decisor inteligente. Supóngase que en una mañana en la que amenaza con llover un individuo llega al trabajo provisto de paraguas. Partiendo de este dato el observador considera inteligible su conducta al reconstruir el proceso de decisión de la manera siguiente. Antes de salir de su casa y ante el pronóstico incierto del tiempo que haría a la salida del trabajo horas más tarde el individuo se ha planteado si llevar o no paraguas. Los posibles estados del mundo que ha tenido en cuenta son: que llueva (E_1) con probabilidad p, o que no llueva (E_2) con probabilidad 1 – p. Sus elecciones alternativas han sido llevar el paraguas (A_1) o no llevarlo (A_2). Los posibles estados de cosas resultantes de la concurrencia de su elección y del azar serían: w

(llueve y lleva paraguas), x (no llueve y lleva paraguas), y (llueve y no lleva paraguas) y z (no llueve y no lleva paraguas). El agente ha estimado la utilidad de cada uno de esos posibles estados de cosas (u_w, u_x, u_y, u_z) atendiendo a variados criterios: la molestia de cargar con el paraguas sin provecho si no llueve; el temor a perderlo, por su carácter olvidadizo; el temor a resfriarse, por su constitución enfermiza; los reproches que se haría si lloviese y lo hubiese dejado en casa, etc. Es muy posible que el individuo admita que es una desgracia ser como es y vivir en una ciudad de clima tan incierto, donde la mera posibilidad de que llueva es ya ominosa. En ese caso sus utilidades pueden ser consideradas "negativas" sin que ello altere su posición relativa en la escala: puede considerar que, *dada* esa desgracia, elegir en esas condiciones el mal menor es igualmente maximizar la utilidad esperada. Nada hay tan malo que no pueda empeorar, y lo peor es enemigo de lo malo. También ha estimado la probabilidad de que llueva o no en función de su experiencia o su habilidad personal para conjeturar a la vista del cielo, de la confianza que le merece el pronóstico meteorológico, etc. De cada uno de los cuatro posibles estados de cosas ha *esperado* por tanto una determinada utilidad (ue_w, ue_x, ue_y, ue_z). Su decisión de llevar (A_1) o no (A_2) el paraguas en previsión de que llueva o no ha sido el resultado de una elección entre dos loterías cuyos respectivos premios son la suma de las utilidades esperadas de los estados de cosas que se materializarán si llueve o no. Su conducta es la evidencia de que la utilidad esperada de la lotería "llevar el paraguas" era superior a la de la lotería "no llevar el paraguas". El agente ha actuado *como si* maximizase su utilidad esperada.

Quod erat exspectandum. Una vez más puede resultar sorprendente que no quepan sorpresas respecto a lo que el modelo predice sobre la conducta de los agentes. Propiamente hablando no se trataría siquiera de predicciones en sentido estricto, sino de implicaciones lógicas que se siguen de ciertos supuestos en condiciones *ceteris paribus.* Todo lo que se ajusta al modelo se conduce como el modelo predice o, si se prefiere, *define.* Lo que justifica el recurso a las utilidades esperadas en el caso de las decisiones arriesgadas es que el agente, al elegir el acto cuya utilidad esperada es máxima, hace simplemente lo que quiere hacer (Res-

nik, M., 1998: 170). Pero si es literalmente así parece ponerse en cuestión la utilidad del modelo no ya sólo para explicar y predecir conductas reales, sino para ofrecer prescripciones para la decisión.

Respecto de lo primero conviene recordar que el modelo no pretende tanto *explicar* cuanto *representar* la conducta observable. El teorema de la utilidad esperada es "un teorema de representación, que muestra que es posible representar numéricamente una estructura no numérica" (íd.: 169). Lo *dado* son las elecciones reales que revelan las preferencias subyacentes, que pueden representarse numéricamente en una escala de utilidad. Las preferencias del agente no se *explican* por las propiedades de sus escalas de utilidad; es más bien a la inversa: las escalas de utilidades poseen determinadas propiedades *porque* los agentes tienen las preferencias que tienen (íd.: 170).

Respecto de lo segundo, si la conducta de los agentes no responde a lo que *predice* la teoría es culpa de los agentes, no de la teoría. El supuesto material de la teoría precisamente en cuanto *práctica* es que actuar de manera racional *beneficia* al agente pues hace máximo el valor de lo que más valora. En consecuencia, por su propio interés, éste *debe* ajustar sus preferencias a las condiciones formales del teorema de la utilidad esperada, construir sus funciones personales de utilidad por extrapolación y "suavización" de curvas a partir de dominios relativamente restringidos de alternativas, etc. (íd.: 99-100).

El agente racional es *metron* y *kanon* ideal de la conducta real porque "lo que él *haría* es lo que el hombre corriente *debería hacer*" (Hahn, F. y Hollis, M., 1986: 31). La teoría da su razón de ser a todo asesoramiento, sea económico, jurídico, fiscal, médico, psicológico o incluso espiritual. El ideal de todo asesor de elecciones racionales sería el programador lineal que pone su técnica de optimización racional al servicio de las preferencias de su cliente, pero que no se limita a ser un mero mecanismo maximizador de preferencias dadas y presentes sino que extiende su asesoramiento al dominio de las metapreferencias y se permite aconsejar racionalmente sobre las preferencias que el cliente *debería* adquirir por su propio interés "verdadero", "a largo plazo", "ilustrado", etc.

4.5.2. *Incertidumbre o ignorancia*

Luce y Raiffa (1957: 276-278) toman de Savage un ejemplo que ilustra las características de la elección en aquellas situaciones en las que el decisor ignora las probabilidades de los estados del mundo de los que depende la utilidad relativa de las consecuencias de sus actos o en las que ni siquiera tiene sentido hablar de probabilidades. Un individuo se ofrece a terminar de preparar una tortilla cuando su mujer ya ha batido cinco huevos en un cuenco. Queda un sexto huevo por abrir, que puede estar sano o podrido, y ha de elegir entre los siguientes actos alternativos: 1) echarlo directamente con los otros cinco; 2) echarlo antes en una taza; 3) tirarlo a la basura sin más. Dependiendo del estado verdadero del mundo –de que el huevo esté sano o podrido– cada uno de los actos tendrá una consecuencia de diversa utilidad para él.

Cuadro 4.7. *Decisión paramétrica en condiciones de ignorancia*

		Estados	
		sano	*podrido*
Actos	echar en el cuenco	tortilla de 6 huevos	sin tortilla y cinco huevos perdidos
	echar en la taza	tortilla de 6 huevos y una taza que lavar	tortilla de 5 huevos y una taza que lavar
	tirar a la basura	tortilla de 5 huevos y un huevo desperdiciado	tortilla de 5 huevos

Si las preferencias del decisor cumplen las condiciones que permiten representarlas por una función de utilidad, el problema de elección en condiciones de ignorancia puede esquematizarse como sigue:

Cuadro 4.8. *Decisión paramétrica en condiciones de ignorancia: esquema general*

		Estados					
		E_1	E_2	...	E_j	...	E_n
Actos	A_1	u_{11}	u_{12}	...	u_{1j}	...	u_{1n}
	A_2	u_{21}	u_{22}	...	u_{2j}	...	u_{2n}
	...	...	...	...	...	...	...
	A_i	u_{i1}	u_{i2}	...	u_{ij}	...	u_{in}
	...	...	...	...	...	...	...
	A_m	u_{m1}	u_{m2}	...	u_{mj}	...	u_{mn}

En la matriz, como es habitual, u_{ij} es la utilidad asociada con el par integrado por el estado E_j y el acto A_i. El problema se reduce entonces a elegir una fila –un acto– cuya utilidad sea máxima. En el caso del riesgo el individuo en cuestión conoce la distribución de la probabilidad a priori en el conjunto de estados posibles del mundo: un técnico avícola "sabría" que en una muestra aleatoria de seis huevos la probabilidad condicional de que el sexto esté podrido cuando los otros cinco están sanos es, por ejemplo, de 0,008. Esto le permitiría considerar la opción de echarlo directamente en el cuenco como una lotería cuyos premios son la tortilla de seis huevos con una probabilidad de 0,992 y quedarse sin tortilla echando a perder cinco con una probabilidad de 0,008. En el caso de la incertidumbre el individuo ignora por completo qué estado del mundo se realizará: puede que sólo haya visto huevos ya fritos, o que el sexto huevo sea moreno y de mayor tamaño que los otros y no tenga la menor idea de lo que pueda esperar. Pero si a pesar de todo *tiene* que tomar una decisión y no parece que cualquiera valga, es necesario analizar si existen criterios precisos para decidir en condiciones de

ignorancia. Un criterio podrá considerarse bien definido "si y sólo si prescribe un algoritmo preciso que en tales condiciones elija sin ambigüedad el (los) acto(s) definidos tautológicamente como 'óptimos según el criterio'" (Luce, R. y Raiffa, H., 1957: 278).

El más conocido de los criterios es el llamado *maximin,* que asocia a cada acto un nivel de seguridad o garantía correspondiente a la utilidad *mínima* que puede procurar al agente. Prescribe a éste que compare las utilidades mínimas que le reporta cada uno de los actos disponibles y elija de entre ellos aquel cuya utilidad mínima sea *máxima* –el *máxi*mo de los *mín*imos (Davis, M., 1986: 31-74; Resnik, M., 1998: 56-58). Si se invierte la escala considerando las pérdidas como utilidades negativas, puede asimismo llamarse *minimax* al criterio que busca hacer mínimas las pérdidas máximas. Se trata de un principio de actuación muy conservador pues implica actuar arriesgando lo mínimo bajo el supuesto de que ocurrirá lo peor, como si se jugase un juego de suma cero contra una naturaleza diabólicamente hostil. Aunque hay casos en los que se justifica adoptar ese criterio –situaciones del tipo "ruleta rusa" en las que las potenciales pérdidas son desproporcionadamente grandes– existen otros en los que el criterio prohibe aprovecharse de la ocasión de obtener grandes ganancias al riesgo de pequeñas pérdidas, como por ejemplo (suponiendo que la utilidad es una función lineal del dinero):

Cuadro 4.9. *Decisión paramétrica en condiciones de ignorancia: criterio maximin*

	E_1	E_2
A_1	1.500 ptas.	1.500 ptas.
A_2	1.400 ptas.	10.000 ptas.

A_1 es el acto óptimo según el criterio *maximin.* Pero elegirlo implica no arriesgarse a perder siquiera 100 ptas. por la opción

de ganar 10.000, lo cual parece intuitivamente poco razonable. Cuando el criterio no selecciona un único acto puede romperse el empate recurriendo a una extensión del principio: el criterio *maximin léxico* o *leximin,* que elige el siguiente menor mínimo.

El criterio *maximin* opera con funciones *ordinales* de utilidad, es decir con la información elemental que proporciona la mera ordenación de los estados en función de las preferencias. El orden relativo entre los elementos de una escala permanece invariable si se somete esa escala a una simple transformación ordinal. Es evidente que tres individuos ordenados por edades decrecientes en 1999 conservan ese mismo orden cinco años más tarde, con independencia de sus edades concretas y, en consecuencia, de los *intervalos* entre ellas. El orden de preferencia de los actos dentro de una misma columna permanece invariable si, por ejemplo, se añade una utilidad constante a todos ellos. De forma general, una transformación ordinal t de una escala es por definición cualquier transformación que conserva el orden de los elementos de esa escala:

$$ua_x \geq ua_y \leftrightarrow t(ua_x) \geq t(ua_y) \text{ para todos los } a_x \text{ y } a_y \text{ en la escala}$$

o en términos de preferencias:

$$a_x \, P \, a_y \leftrightarrow t(a_x) \, P \, t(a_y)$$

Si se somete la tabla anterior a una transformación ordinal arbitraria que cumpla las condiciones estipuladas se obtiene:

Cuadro 4.10. *Criterio maximin. Transformación ordinal*

	E_1	E_2
A_1	1.510 ptas.	9.999 ptas.
A_2	1.401 ptas.	10.000 ptas.

El criterio maximin sigue seleccionando en esta tabla el mismo acto A_1 que antes de la transformación. La información sobre las preferencias del agente que se necesita para aplicar el criterio *maximin* es muy somera.

No ocurre lo mismo con otro importante criterio propuesto por Savage, el diversamente llamado de *riesgo minimax* (Luce, R. y Raiffa, H., 1957: 280-282) o de *arrepentimiento minimax* (Resnik, M., 1998: 29-32). La idea de *arrepentimiento –regret–* carece de toda connotación moral en este contexto: es el puro pesar por la ocasión perdida, como mera lamentación o "tristeza acompañada por la idea de algo que creemos haber hecho por libre decisión del alma" (Spinoza, B., 1987: 251 [III, definición 27]). No nace de la razón ni es una virtud: el que se arrepiente de lo hecho es dos veces miserable e impotente (IV, prop. LIV). El concepto de arrepentimiento es importante para la teoría de juegos: el equilibrio que constituye la solución de un juego es el punto en el que cada agente, revisando al día siguiente su propia decisión, no tiene motivos para arrepentirse de la estrategia seguida por él *dada* la estrategia adoptada por el contrario (Poundstone, W., 1995: 163).

En el caso anterior (antes de la transformación ordinal), si se da el estado E_1 y se elige A_1 no se corre riesgo y no habrá nada que lamentar; si se elige A_2 se corre el pequeño riesgo de perder 100 ptas. y habrá poco de que arrepentirse. Si se da el estado E_2 y uno elige A_2 tampoco se corre riesgo ni habrá ocasión de arrepentirse. Pero si se elige A_1 el riesgo es enorme y se lamentará haber desperdiciado la ocasión de ganar 8.500 más. Este criterio permite tener en cuenta las ocasiones perdidas en vez de las probabilidades peores.

Si las matrices anteriores representaban las *utilidades,* a partir de ellas se puede obtener una matriz de *riesgos* o motivos de *arrepentimiento* comparando la utilidad de cada acto del conjunto de los posibles en un determinado estado del mundo con la del acto de ese conjunto que proporcionaría la máxima utilidad. En concreto, restando la utilidad de cada celda de la utilidad máxima en su columna. En el caso anterior (antes de la transformación ordinal):

Cuadro 4.11. *Matriz de arrepentimiento*

	E_1	E_2
A_1	0 ptas.	8.500 ptas.
A_2	100 ptas.	0 ptas.

El criterio de *riesgo* o *arrepentimiento minimax* selecciona el acto que minimiza el índice máximo de riesgo o arrepentimiento para cada acto, que en este caso es justamente A_2, en contra del recomendado por el criterio *maximin.*

Pero si se aplica el mismo procedimiento para obtener una matriz de arrepentimiento de la tabla resultante de la transformación ordinal:

Cuadro 4.12. *Matriz de arrepentimiento: transformación ordinal*

	E_1	E_2
A_1	0 ptas.	1 ptas.
A_2	109 ptas.	0 ptas.

el criterio de *arrepentimiento minimax* en este caso elige el acto A_1, en lo que coincide con el elegido por el criterio *maximin.* Se deduce por tanto que "la regla de arrepentimiento minimax no elige necesariamente el mismo acto cuando se aplica una transformación ordinal de utilidad a una tabla de decisión" (Resnik, M., 1998: 60-61). La razón es que esta regla introduce mayor información que la requerida en una escala ordinal. El pesar por la ocasión perdida depende de la *distancia* entre lo poco que se tiene y lo mucho que se pudo haber tenido. La analogía con la música permite entender la importancia del tipo de información contenida en las escalas ordinales y cardinales de utilidad.

La escala diatónica ordena los sonidos en grados o notas según sus frecuencias del grave al agudo, pero el número *ordinal* de cada nota no tiene una relación directa con la nota en sí: el *re* y el *mi,* que son el segundo y el tercer grados de una escala pueden ser el tercero y el cuarto, o el quinto o el sexto de otras según su relación con el que se convenga en tomar como primero en cada una de ellas. La entonación propia de cada nota es el número *cardinal* de su frecuencia; la distancia o intervalo entre dos notas es asimismo la medida cardinal de su diferencia: precisamente de ese intervalo dependen la armonía y la melodía. Al duplicar las frecuencias de las notas de una escala ésta se eleva simplemente a la octava superior, más aguda, que conserva la medida de los intervalos. Una transformación más compleja –el transporte– permite interpretar una composición musical en una tonalidad distinta de la original, pero conservando prácticamente toda la información melódica y armónica. Para *transportar* una composición de la tonalidad de *sol* menor a la de *mi* menor se someten las magnitudes cardinales que representan las frecuencias e intervalos a una operación matemática semejante a la transformación lineal.

De manera análoga las magnitudes cardinales de temperatura en un termómetro graduado no dependen en sí mismas de la escala, pues tanto el cero como la unidad de medida son arbitrariamente elegidos, por ejemplo en las escalas Celsius o Fahrenheit. Pero cualquier resultado obtenido al usar una determinada escala debe mantenerse inalterado si se selecciona un cero diferente y una unidad diferente para la escala; cualquier ley científica sobre temperaturas ha de poder expresarse indiferentemente en grados Celsius o Fahrenheit. Las únicas propiedades relevantes de los números son aquellas que permanecen invariantes en cualquier transformación lineal.

La razón de ello es que, si se representan las diversas magnitudes por medio de puntos en una línea recta, lo que una escala interválica representa son meramente las proporciones de las distancias entre los puntos, de modo que todo cuanto se diga sobre los puntos debe mantenerse sin variaciones en cualquier traslación uniforme de los puntos a lo largo de la línea, y en cualquier expansión o contracción uniforme de la línea, esto es, en cualquier cambio en el punto cero o en la unidad de la escala (Braithwaite, R., 1955: 10-11). Toda transformación meramente ordinal conserva la información contenida en

la escala de preferencias que basta para aplicar el criterio *maximin.* Pero el criterio de *arrepentimiento minimax* no opera con simples ordenamientos de preferencias, sino con magnitudes cardinales que representan los intervalos entre las utilidades de los actos posibles. Por esa razón no cualquier transformación vale, sino sólo la lineal positiva, que de hecho requiere información no sólo de *que* un acto es preferido a otro, sino de *cuánto* más, con qué intensidad, etc. Todo lo cual implica recurrir a unidades de medida que permitan efectuar comparaciones significativas, las cuales plantean serias dificultades en el caso de las comparaciones *intra*personales, pero que parecen insuperables en el de las comparaciones *inter*personales de utilidad. El asunto trasciende de su interés meramente académico pues suscita cuestiones de gran trascendencia, por ejemplo, para la ética utilitarista y para la economía del bienestar, que comparten el propósito de optimizar una magnitud común de utilidad.

Los dos criterios anteriores responden a la disposición básicamente precautoria del conservador pesimista –Casandra– que se prepara para lo peor; o más exactamente, puesto que de decisiones se trata, *apuesta* por lo peor. Un imperturbable optimista –Pangloss– aplicaría el criterio exactamente contrario de elegir el acto que proporcione el máximo de los máximos de utilidad. Como este criterio *maximax* parece en exceso imprudente, se ha propuesto elaborar un criterio mixto de la manera siguiente: sean m_i la mínima utilidad y M_i la máxima utilidad de un acto A_i; supóngase *dado* un número fijo $0 < \alpha < 1$ llamado *índice de pesimismo-optimismo;* asóciese a cada acto A_i el índice $\alpha m_i + (1 - \alpha)M_i$, llamado *índice* α de A_i, y elíjase de dos actos el que tenga un *índice* α más elevado. Si $\alpha = 0$ este procedimiento equivale al criterio *maximin;* si $\alpha = 1$, equivale al *maximax* (Luce, R. y Raiffa, H., 1957: 282-284; Resnik, M., 1998: 65-69). El índice en cuestión se *infiere* de experimentos psicológicos que "obligan" al agente a revelar sus preferencias haciendo de esa forma explícito cuán optimista es de hecho. Resulta por ello de escaso valor prescriptivo pues "no impone ninguna coherencia a las decisiones en condiciones de ignorancia adoptadas a lo largo del tiempo por un grupo de individuos y ni siquiera por un único individuo" (Resnik, M., 1998: 67).

Si el agente ha de adoptar su decisión en condiciones de genuina incertidumbre, es decir, que ignora por completo cuál de los esta-

dos del mundo E_1, E_2, ... E_n se da, realmente no existen *razones suficientes* para elegir un acto en vez de otro. En este caso el *principio de razón insuficiente* aconsejaría considerarlos a todos como *igualmente probables*, convertir el problema de decisión *incierta* en uno de decisión *arriesgada* con una distribución uniforme de probabilidad a priori, y elegir en consecuencia el acto de superior utilidad (esperada). El principio, formulado como tal por Jakob Bernoulli, "es excesivamente vago y su uso indiscriminado ha acarreado resultados insensatos" (Luce, R. y Raiffa, H., 1957: 283). Tropieza con dificultades prácticas de aplicación, como la de elaborar la lista exhaustiva y mutuamente exclusiva de estados posibles del mundo. Pero también con dificultades teóricas: una cosa es ser indiferente entre dos estados de cosas porque se cree que existen buenas razones para considerarlos equiprobables –el caso de la moneda no trucada– y otra es hallarse indeciso por carecer de cualquier razón para asignarles probabilidad alguna. Cuando se da auténtica ignorancia, parece evidente que tampoco pueden existir razones para considerarlos equiprobables como prescribe el principio.

Es comprensible la perplejidad "del pobre decisor, ahora totalmente confundido por las ventajas y los inconvenientes de tales principios. ¿Podría en su desesperación buscar un compromiso, adoptando algún tipo de compuesto arbitrario de los criterios?" (Luce, R. y Raiffa, H., 1957: 285). Pero aparte de que hay compromisos que sólo son aceptables en apariencia, los propios principios aconsejan elecciones contradictorias. En este caso podría recurrirse a tomar la decisión entre pares de actos que recomiende la mayoría de aquéllos, pero tampoco serviría de mucho: cuando las preferencias del conjunto de los votantes son intransitivas las votaciones pueden incurrir en los ciclos de mayorías que refleja la *paradoja de la votación* o de *Condorcet*. Supóngase que tres votantes V_1, V_2 y V_3 tienen que elegir entre tres alternativas x, y, z, que cada cual ordena como sigue:

$$V_1: x\ P\ y\ P\ z$$

$$V_2: y\ P\ z\ P\ x$$

$$V_3: z\ P\ x\ P\ y$$

Es fácil ver que cada una de las tres alternativas es preferida cíclicamente por una mayoría de dos a uno. Los votos de V_1 y V_2 dan el triunfo a y. Pero la mayoría (V_1 y V_3) consideraba x preferible. Ni siquiera puede considerarse la elección de y como una solución de compromiso pues lo mismo cabría decir de cada una de las alternativas. Si el compromiso consiste en establecer un determinado orden sucesivo e irreversible en la votación –por ejemplo, eliminando en una primera fase la alternativa menos preferida– éste a su vez ha de ser justificado por un criterio de decisión, con lo que se vuelve al punto de partida.

Las conclusiones a que llegan los teóricos no son precisamente optimistas. Tal vez no exista un único criterio válido para toda posible situación, sino criterios varios apropiados en situaciones determinadas. Si así fuera resultaría más provechoso "especificar las condiciones que limitan la posible aplicación de las diversas reglas en vez de buscar las que favorecen la aplicación de una regla única" (Resnik, M., 1998: 40). Pero si esto implica parcelar el universo único de las decisiones racionales es evidente que el optimismo básico del modelo queda seriamente comprometido. La transición del nivel de las elecciones paramétricas al de las estratégicas hace más ominosa aún esta perspectiva.

Robinson Crusoe solitario en su isla es el arquetipo de una situación paramétrica. Ciertamente no cabe siquiera imaginar que un Robinson aislado *ab initio*, al margen de toda interacción social y lingüística hubiese podido, no ya alcanzar el umbral de la racionalidad, sino incluso sobrevivir. De hecho se trata de un náufrago ya socializado en cuyas preferencias es difícil discernir lo que se reduce a puros gustos y lo que implica una referencia normativa. No obstante, el recurso a situaciones de este tipo –como la de Adán en el paraíso antes de la creación de Eva (Hollis, M., 1988: 15-28) y, en definitiva, la del hombre en estado de naturaleza– ha de entenderse como un *Gedankenexperiment* que permite *reconstruir* analíticamente las dimensiones individuales y colectivas de la racionalidad, aunque, como todo experimento, no esté exento de supuestos ontológicos implícitos o explícitos. Su carácter artificioso no impide que innumerables situaciones de la vida real puedan ser representadas en términos de este modelo: el individuo que se pregunta por los motivos racionales para

cooperar con sus anónimos semejantes en situaciones del tipo del dilema del prisionero indaga en última instancia qué razones tiene para abandonar el estado de naturaleza, suscribir el contrato social y atenerse a él.

A Robinson todos los elementos relevantes para plantear y resolver sus problemas de decisión le están dados de antemano y son independientes de ella. Estos datos son los parámetros o constantes del problema cuya solución es su decisión concreta. Su universo está poblado de objetos o cosas que le ignoran, pues carecen de capacidad de formarse expectativas respecto de sus decisiones y acciones, prescindiendo aquí de las cuestiones que suscitan los comportamientos innegablemente estratégicos de ciertos animales. En ausencia de esos peculiares seres que parecen poseer la condición de agentes, la hipótesis más plausible –o en todo caso más eficaz en términos pragmáticos– que Robinson puede formular sobre su entorno es que opera de forma determinista. El entorno incluye ciertamente sus capacidades y limitaciones personales como, por ejemplo, su impericia para construir toneles, las cuales "limitan la amplitud de su satisfacción; sus preferencias, en cambio, afectan al uso que hace de sus capacidades, no limitando lo que puede hacer, sino lo que cree que merece la pena hacer" (Gauthier, D., 1994: 129-131).

En el dominio de sus decisiones posibles reina la *prudencia*, que dicta siempre la elección racional a favor del mejor resultado posible en beneficio propio. Sus actos son por tanto meros medios para conseguir los objetivos que satisfacen sus preferencias. No cualesquiera, sino las ordenadas tras una deliberación racional; no las que responden a un capricho o impulso momentáneo, sino las que tienen en cuenta el beneficio a largo plazo, etc. El prudente es un optimizador que toma las decisiones que le permiten extraer la máxima utilidad de las oportunidades que le brinda un entorno *dado* con el que ha de contar como variable independiente. En su juego con(tra) la naturaleza saca el mejor partido de las cartas que le han tocado en suerte. Por eso Robinson puede considerarse el arquetipo del *homo economicus:* consume lo que produce y produce lo que prefiere. Y, "puesto que puede suspender sus esfuerzos productivos cuando el coste de una producción adicional exceda al beneficio de consumir el bien producido, sólo

se puede culpar a sí mismo si el coste marginal y los beneficios de su actividad no son iguales" (ibid.).

La irrupción de Viernes altera radicalmente los supuestos fácticos de esa hipótesis. Si atribuir intencionalidad a la naturaleza es un error relativamente inocuo, ignorarla en lo que con toda probabilidad es otro agente sigue siendo un error *teórico*, pero de graves consecuencias *prácticas*, una temeraria *imprudencia*. Le impide deliberar y actuar adecuadamente en su propio beneficio, haciéndole incapaz de predecir adecuadamente el comportamiento de un elemento de un entorno radicalmente transformado. Y, muy en particular, de uno que por poseer la singular capacidad de tomarle en cuenta y formarse expectativas respecto de su conducta puede afectar de manera imprevista su función de utilidad. La caridad bien entendida –incluso en el sentido de Davidson– empieza por uno mismo.

La presencia de Viernes obliga a Robinson a que, como mínimo y por su propio interés, no ignore los perjuicios que ésta pueda acarrearle o los beneficios que pueda procurarle. Lo mismo cabe decir de Viernes cuando descubre que la isla está habitada. O, dicho con más precisión: cuando *ambos* son mutuamente conscientes de que habitan la misma isla. El mero hecho de su coexistencia y de su rudimentaria interacción hará que ambos se formen expectativas recíprocas en función de las cuales adoptarán decisiones individuales que, aun sin cooperar expresamente ni coordinarse tácitamente, concurrirán *de hecho* a producir efectos, estados de cosas –en definitiva, *bienes*– que, en ciertas condiciones, tendrán incluso un carácter colectivo o público. La decisión de cada uno produce economías y deseconomías externas –*externalidades* en general– para el otro que es imprudente ignorar. Nada más racional, en principio, que seguir los dictados de la prudencia si se quiere obtener la máxima utilidad de las decisiones. O, al menos, no parece que hasta el presente haya razones para dudarlo.

Hasta este momento la reconstrucción hipotético-genética de los niveles progresivamente complejos de la racionalidad práctica no ha requerido introducir ninguna consideración que moviese a Robinson a imponer restricciones a su conducta por razones distintas de su propio interés. La aparición de Viernes por sí misma tampoco las exige, dado que la conducta de éste puede ser

tanto competitiva como cooperativa y corresponde a la prudencia racional de Robinson elegir la respuesta más adecuada a sus intereses en función de las probabilidades que asigne a una u otra hipótesis. No hay que desarrollar en detalle el hipotético proceso por el que tanto Robinson como Viernes deciden la estrategia más conveniente a sus respectivos intereses: el resultado es que optan por cooperar con la perspectiva de producir conjuntamente bienes –materiales y espirituales, como la amistad– que cada uno por separado no puede procurarse.

Es un lugar común en la abundante literatura contemporánea en torno a la cooperación racional (Regan, D., 1980; Axelrod, R., 1986; Taylor, M., 1987; De Jasay, A., 1989; Olson, M., 1992) que toda cooperación produce bienes comunes o públicos que en parte añaden y en parte sustraen una determinada cantidad a los beneficios propios. Si se descartan los casos de coordinación espontánea, que son de limitada eficacia y por su propia naturaleza inestables, es un hecho que organizar y sostener la cooperación implica costes de negociación y de mantenimiento y mecanismos para distribuir cargas y beneficios, etc. Todas estas circunstancias favorecen la aparición de los gorrones y los parásitos mencionados más arriba y el desarrollo de situaciones de ineficacia y de irracionalidad colectiva como las que ilustra el Dilema del Prisionero. Pero aun así parece que la propia teoría de la racionalidad prudencial podría hacer frente a las situaciones de este tipo por medio de las estrategias estrictamente interesadas que analiza la Teoría de Juegos. Bien es verdad que se hace necesario transformar determinados presupuestos de la racionalidad paramétrica en la medida necesaria para hacer frente a las situaciones estratégicas, sin renunciar a la concepción básicamente maximizadora de la racionalidad. El modelo de la elección racional reduce estas transformaciones a cambios de escala en un gradiente de complejidad creciente pero sin solución de continuidad *dentro* de un universo homogéneo de racionalidad. Como se verá, ciertos *resultados de imposibilidad* parecen exigir una mutación *de* este universo.

5

Decisiones estratégicas: moralidad y racionalidad

5.1. La interdependencia de las expectativas y las decisiones

Un joven es invitado a una cena familiar por los padres de su novia, interesados en conocerle (Jeffrey, R., 1983: 2-5). Sabe que servirán o pavo o rosbif o lubina. También sabe que tienen gustos muy convencionales respecto a la clase de vinos que deben acompañar ciertos platos, consideran pecado (gastronómico) mortal mezclar tinto con pescado o blanco con carne, y no menos pecado, aunque venial, pavo con blanco. Se considera obligado, por cortesía o por causar buena impresión, a llevar una botella de buen vino (sólo una pues su presupuesto no le da para más). Ha de elegir entre comprar vino tinto o blanco. ¿Cómo decidir?

Dependiendo de su conocimiento de los factores que intervienen en la situación el invitado puede interpretarla en principio como una elección paramétrica en condiciones de incertidumbre o de riesgo. Si no tiene la menor idea de lo que podrán servir de hecho tendrá que recurrir a alguno de los principios de decisión en situaciones de ignorancia: *maximin, maximax, riesgo minimax,* etc. Si su novia le ha proporcionado alguna información sobre lo que suelen hacer en otras ocasiones, o sobre lo que cree que han comprado esa mañana, etc., su elección equivaldrá a participar en una lotería arriesgada. Ponderará las utilidades esperadas de los *premios* asociados a su decisión de llevar vino tinto o blanco según prevea que servirán rosbif, lubina o pavo, y que abarcan desde el acierto de la combinación óptima al planchazo de la pésima. Elegirá comprar la botella que le prometa mejores resultados en promedio –la máxima utilidad esperada–. Puede

adoptar una estrategia *pura* y decidir jugar directamente pares o nones –tinto o blanco– o una estrategia *mixta* por la que somete su decisión al resultado de algún mecanismo aleatorio que simule las probabilidades y utilidades de los acontecimientos relevantes: tirar monedas o dados, deshojar margaritas, o proceder según tenga o no los ojos verdes la primera mujer con la que se cruce en la acera (Davis, M., 1986: 49-62; Gauthier, D., 1994: 97-99). Pero en todo momento dará por supuesto que la decisión de sus anfitriones es probabilísticamente independiente de la suya. Es decir, que la decisión de ellos de servir alguno de los platos está tomada con anterioridad a, y con independencia de, la suya propia de llevar un determinado vino. El invitado está justificado al considerar a sus anfitriones como parte de la naturaleza: eso es exactamente lo que significa tomar decisiones en una situación paramétrica. Aunque sin duda *son* personas –*otros yo*– que a su vez toman decisiones, a los efectos de la suya propia puede considerarlos, a ellos y a sus decisiones, *como si* fueran meros acontecimientos naturales. Son éstos los casos en los que con toda naturalidad se adopta el punto de vista externo y objetivo del científico natural para considerar como *acontecimientos* lo que el propio científico como persona sabe que son *acciones.*

Ya se vio anteriormente que la *ratio cognoscendi* de la distinción entre acontecimientos y acciones se basa en las características observables de determinados fenómenos de la experiencia, que hacen plausible conjeturar que responden a la intención o el propósito de alguien. Interpretar ciertas incisiones en una piedra como inscripciones depende en gran parte del conjunto de la información contextual de que se dispone. Si aquella distinción proporcionaba a las ciencias morales su razón de ser, su trascendencia es aún mayor para la teoría de la decisión. Supóngase ahora que el invitado adquiere nueva información sobre los datos de la situación. Por ejemplo, su novia le transmite su sospecha de que sus padres buscan poner en evidencia su bisoñez gastronómica, porque le son hostiles o porque tienen un morboso sentido del humor, o ambas cosas. Y aguardan la ocasión de servir el plato que peor vaya con el vino que lleve. Desde el momento mismo en que dispone de esa información, e *ipso facto,* sus anfitriones dejan de ser naturaleza. La naturaleza no es hostil ni propi-

cia; los agentes –por suerte o por desgracia, pero en todo caso por naturaleza– *pueden* serlo. Mirada desde su propia perspectiva de decisor, la probabilidad de que sus anfitriones decidan poner lubina no es ya la de un estado del mundo aleatorio e independiente, sino que aumenta en función de su propia decisión de llevar vino tinto. Pero la información de que dispone incluye la de la perspectiva de sus anfitriones. Si lo suponen al tanto de su hostilidad saben que decidirá llevar el vino que mejor case con el plato que sirvan: para ellos tampoco él es naturaleza, ni la probabilidad de que lleve tal o cual vino es independiente de la decisión que ellos tomen.

El modelo de racionalidad paramétrica no puede dar adecuada cuenta del juego de espejos de estas expectativas recíprocas sin alterar radicalmente sus presupuestos. Es preciso extender o incluso transformar el concepto mismo de racionalidad para incluir las situaciones estratégicas que por su propia naturaleza son *colectivas*. Si la interdependencia de las decisiones caracteriza formalmente las situaciones estratégicas, la producción de efectos *conjuntos* y de utilidades *distintas* lo hace en sentido material. En las situaciones paramétricas –*loterías*– los estados de cosas –*premios*– son el efecto conjunto de *un* acto –(no) comprar determinado número– y un estado del mundo –(no) salir ese número–. En las estratégicas –*juegos*– los estados de cosas –vino tinto con lubina– son el efecto conjunto de (al menos) *dos* actos –llevar uno y servir otra–, más, por supuesto, un determinado estado del mundo –que funcione el horno, que no se rompa la botella por el camino, etc.–. En las paramétricas, las utilidades asociadas a los premios son únicas para el único agente. En las estratégicas son distintas para cada jugador al menos en el sentido obvio de que cada uno valora el resultado según su propia escala de utilidad, con independencia de que ésta pueda o no compararse con la del otro. Cuando el invitado comprueba que ha traído el vino gastronómicamente incorrecto, la *misma* situación que a él le resulta embarazosa es *schadenfreudig* para sus malévolos anfitriones.

Lo que caracteriza formalmente la lógica de la decisión en las situaciones estratégicas es el hecho mismo de que las decisiones son interdependientes. La perspectiva individual de cada agente contiene o refleja las perspectivas individuales de los demás agen-

tes. Cada uno, como una mónada leibniziana, es un *miroir vivant* del universo de los decisores implicados. En términos formales de la Teoría de Juegos, se parte del supuesto de que los jugadores comparten en común el conocimiento de la estructura del juego. Una determinada proposición P que expresa la creencia de un determinado jugador es objeto de conocimiento común –*common knowledge*– si cada jugador tiene razones para creer que P, para creer que el otro jugador tiene razones para creer que P, etc. Y, en un ulterior nivel de complejidad lógica, se supone asimismo, en primer lugar, que es conocimiento común que cada jugador es un decisor perfectamente racional que considera las estrategias disponibles de su adversario *como si* se tratase de acontecimientos y les asigna probabilidades subjetivas; y, en segundo lugar, que cada jugador conoce la verdad de cualquier teorema acerca del juego que pueda ser demostrado (Luce, R. y Raiffa, H., 1957: 47-51; Hollis, M. y Sugden, R, 1993: 8-9).

Si se cumplen estas premisas, los jugadores resultarían ser, en principio y a todos los efectos, *transparentes:* es imposible que engañen o que sean engañados (Parfit, D., 1985: 18-19; Gauthier, D., 1994: 231-239). Esta condición plantea serias dificultades a la propia teoría de juegos no sólo como modelo formal de interacción sino como modelo plausible de representación de las interacciones reales. En los juegos de coordinación de la vida diaria que el modelo pretende representar –por ejemplo, la circulación por carretera– al agente que decide conducir por la derecha esperando que los demás también lo hagan no basa su decisión en la creencia de que éstos son demasiado estúpidos para darse cuenta del *regressus in infinitum* de las expectativas recíprocas, sino en la creencia de que acatarán las normas; en el juego del prometer, la razón de B para fiarse de A no se basa en suponerle incapaz de darse cuenta de que le compensaría incumplir su promesa. El supuesto de la completa transparencia lleva la teoría de la interacción estratégica a un punto muerto, lo que implica sin duda una deficiencia en el modelo (Hollis, M. y Sugden, R, 1993: 24-25).

Aunque pueda parecer sorprendente, en la vida real es la incertidumbre –en el sentido ordinario, no técnico del término– respecto de lo que los demás saben y creen la que hace posible las

relaciones e interacciones personales. La ignorancia absoluta, no menos que la total transparencia, las abortan de raíz. Como sagazmente observara Simmel en sus consideraciones sobre el secreto, "la confianza es una hipótesis sobre la conducta futura del otro, hipótesis que ofrece seguridad suficiente para fundar en ella una actividad práctica. Como hipótesis, constituye un grado intermedio entre el saber acerca de otros hombres y la ignorancia respecto de otros. El que sabe, no necesita 'confiar'; el que ignora no puede siquiera confiar" (Simmel, G., 1977: 366-367). Para dar cuenta teórica de este hecho se ha introducido la noción de *translucencia,* que permite representar de forma más ajustada la limitada capacidad de los agentes reales para conocer las verdaderas preferencias, la verdadera información o la verdadera capacidad de computación del otro, y, sobre todo, para manejar los niveles de complejidad que implican las creencias propias sobre las creencias ajenas sobre las creencias propias, etc.

En las situaciones estratégicas potencialmente cooperativas en las que intervienen agentes *translúcidos* ha de suponerse que cada agente dispuesto a cooperar a condición de que otros cooperen dispone de cierta *pericia* para identificar entre los demás agentes los potenciales cooperadores (Regan, D., 1980: 165-189). Más aún: dada la incertidumbre respecto de la disposición ajena a cooperar, y convencido de que la estrategia estrictamente egoísta en situaciones del tipo del dilema del prisionero conducen a un punto muerto incluso desde la perspectiva de la maximización de la propia utilidad, el agente adoptará la firme resolución –*resolute choice*– de *convertirse* en cooperador, adquiriendo las disposiciones necesarias para lograrlo (Mcclennen, E., 1988: 101-118; Sayre Mc-Cord, G., 1991: 187-195; Smith, H. 1991: 235-238). Como el propio Gauthier reconoce, "quizás el mejor modo de cosechar las ventajas de la cooperación consista en ser un cooperador genuino" (Gauthier, D., 1998a: 108), ya que "cada uno gana disponiéndose a cooperar condicionalmente con los otros"; en consecuencia, lo que un egoísta debe –*should*– hacer es "convertirse en un cooperador y aceptar la moralidad" (íd.: 117). El argumento presenta sorprendentes analogías con el de la *apuesta* de Pascal y parece implicar que el supuesto básico del modelo de agentes movidos únicamente por razones interesadas –pruden-

ciales o *forward-looking*– hace racionalmente imposible (también de explicar) la moralidad entendida como cooperación en los términos estrictos de la Teoría de Juegos.

5.2. Modelos de interacción estratégica

Con las correspondientes modificaciones, los mismos esquemas antes utilizados permiten representar las alternativas de elección de los agentes implicados en una situación estratégica. Las decisiones o estrategias de los agentes pueden representarse formalmente de dos maneras. La primera de ellas es la llamada forma *extendida* que representa en forma de *grafo* las *sucesivas* jugadas o decisiones alternativas a las que el agente se enfrenta en cada disyuntiva.

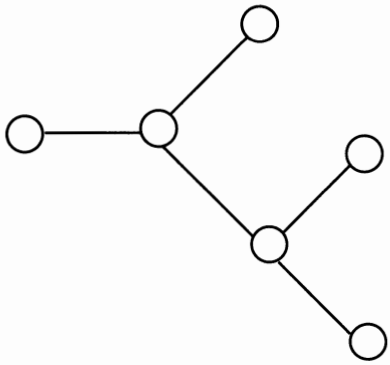


Figura 5.1. Forma extendida: el árbol de un juego.

Un *grafo conexo* consiste en una colección de puntos (llamados *nodos* o vértices) y de ramas entre ciertos pares de nodos, de forma tal que puede trazarse un camino desde cada punto a cada uno de los demás puntos, que en algunos casos son cerrados en forma de ciclos o bucles. Cuando éstos no existen el grafo recibe el nombre de *árbol del juego*. Esta representación permite tener en cuenta la *historia* del juego. Se supone que el árbol es finito en

el sentido de tener un número finito de nodos y ramas y una regla definida para terminar el juego, lo que no excluye que ese número alcance proporciones desmesuradas y que, en consecuencia, sea imposible en la práctica representar de esta forma la totalidad del juego (Luce, R., Raiffa, H., 1957: 39-41). Una manera de simplificar esta estructura consiste en especificar de antemano para cada jugador lo que decidiría en cada una de las situaciones que pudieran surgir en el desarrollo del juego, es decir, cuál sería su *estrategia pura* (íd.: 51). Cada jugador elige por tanto exactamente *una* de las alternativas a las que se enfrenta en cada nodo. Todo juego posible puede reducirse así a una forma simple estándar llamada la *forma normal* de un juego, que representa un mismo esquema formal: los diversos jugadores tienen un control limitado de las variables que determinan lo que obtendrá del juego, que cada uno de ellos desea maximizar (íd.: 53).

Reducido por tanto a su esquema más simple (2 × 2, esto es, dos jugadores y dos estrategias) la *forma normal* del juego representa en una tabla o matriz los dos *jugadores* (A, B) cada uno los cuales puede hacer (adoptar la *estrategia)* C o D. El resultado *conjunto* de sus estrategias determina *consecuencias* (w, x, y, z):

Cuadro 5.1. *Decisiones estratégicas: matriz de consecuencias*

		B	
		hace C	hace D
A	hace C	w	x
	hace D	y	z

Cada consecuencia proporciona *utilidades (pagos)* a cada jugador (uA_w, uB_w; uA_x, uB_x...). Por convención en las casillas se representan las utilidades de cada uno separadas por comas, de forma que las de la izquierda corresponden al jugador de las filas, y las de la derecha al de las columnas.

Cuadro 5.2. *Decisiones estratégicas: matriz de pagos*

		B	
		hace C	hace D
A	hace C	uA_w, uB_w	uA_x, uB_x
	hace D	uA_y, uB_y	uA_z, uB_z

En las decisiones reales de los jugadores intervienen cálculos de utilidades cardinales y estimaciones de probabilidad subjetiva, pero en términos formales puede prescindirse de este tipo de consideraciones y limitarse tan sólo a las propiedades meramente *ordinales* de las preferencias de los jugadores por los estados de cosas que son el *resultado* conjunto de los *pares de estrategias*. Esta perspectiva formal permite elaborar una taxonomía de los distintos tipos de juegos (Guyer, M. y Rapaport, A., 1966: 203-14; Poundstone, W., 1995: 321-332) cuyo enorme rendimiento heurístico explica la acogida brindada a la teoría de juegos tanto por las ciencias morales como por la filosofía moral. Gracias a ella han podido identificarse cuatro modelos formales de *juegos* que representan la lógica de la elección racional en cuatro tipos de situaciones, las cuales plantean complejos dilemas teóricos y prácticos: el juego del *punto muerto (deadlock)*, el de las *garantías mutuas (assurance)*, el del *gallina (chicken)* y, el más dramático y perverso, el *dilema del prisionero*.

Para entender las características de cada uno de estos juegos hay que tener en cuenta que los *resultados* pueden considerarse desde una doble perspectiva: por una parte –*ex ante*– desde la perspectiva del conjunto de *estrategias* cuyo producto son; por otra –*ex post*– como tales productos que proporcionan *pagos* o utilidades a los jugadores. Desde la perspectiva de las estrategias que pueden adoptar los agentes, un resultado es un *equilibrio de Nash* si y sólo si cada estrategia maximiza su valor para el agente *dadas las restantes estrategias*. Desde la perspectiva de su valor como pago, un resultado es *óptimo de Pareto* si y sólo si cada pago maxi-

miza el valor de quien lo recibe *dados los restantes pagos* (Gauthier, D., 1998a: 96). En términos ideales, el modelo de la elección racional *debería* garantizar que las estrategias óptimas garantizan los resultados óptimos.

Un caso relativamente sencillo de juego de coordinación (Lewis, D., 1969: 9-11) permite entender las relaciones entre los conceptos de equilibrio y de optimalidad. Dos individuos (F, C) que han acudido juntos a un gran centro comercial se pierden mutuamente de vista de forma involuntaria y desean volver a encontrarse en alguno de tres posibles puntos de encuentro. Sus posibles estrategias (1, 2, 3) vienen reflejadas en las siguientes matrices:

Cuadro 5.3. *Indiferencia entre equilibrios*

	C_1	C_2	C_3
F_1	1 encuentro 1	0 0	0 0
F_2	0 0	1 encuentro 0	0 1
F_3	0 0	0 0	1 encuentro 1

Las estrategias F_1C_1, F_2C_2 y F_3C_3 están en equilibrio: cada una de ellas es la mejor respuesta a la estrategia del otro individuo. Por ejemplo, el resultado de la estrategia F_2C_2 es un equili-

brio porque F lo prefiere a F_1C_2 y a F_3C_2, y porque C lo prefiere a F_2C_1 y a F_2C_3. Representan las tres combinaciones de estrategias que dan como resultado que ambos van al mismo lugar y se encuentran. Los resultados son óptimos en el sentido antes indicado y equivalentes para cada uno: ambos prefieren encontrarse a no encontrarse y les da lo mismo dónde.

Cuadro 5.4. *Preferencias comunes entre equilibrios*

	C_1	C_2	C_3
F_1	1,5 encuentro 1,5	0,2 0,5	0 0,5
F_2	0,5 0,2	1,2 encuentro 1,2	0 0,2
F_3	0,5 0	0,2 0	1 encuentro 1

Los equilibrios son los mismos que en el caso anterior. Los individuos siguen prefiriendo encontrarse a no encontrarse; pero, aunque no les resulta indiferente dónde hacerlo, coinciden en sus preferencias: F_1C_1 es el resultado óptimo; F_2C_2 y F_3C_3 son subóptimos, pues pueden ser mejorados para ambos sin perjuicio para ninguno.

Los dos casos precedentes ilustran la ausencia de conflicto entre individuos que poseen intereses comunes. Ambos juegan a un juego de pura coordinación en el que el interés por encon-

trarse, o bien les hace indiferentes al sitio donde lo hagan, o bien suma las ventajas de hacerlo en el sitio predilecto de ambos. El caso siguiente introduce una notable variante en lo que hace a la optimidad de los resultados.

Cuadro 5.5. *Preferencias diversas entre equilibrios (La batalla de los sexos)*

	C_1	C_2	C_3
F_1	1 encuentro 1,5	0,2 0,5	0 0,5
F_2	0 0,2	1,2 encuentro 1,2	0,5 0,2
F_3	0 0	0,2 0	1,5 encuentro 1

Los equilibrios siguen siendo los mismos que en los casos anteriores y exactamente en el mismo sentido: cada estrategia es la mejor respuesta a la estrategia ajena. Pero los resultados producidos no tienen el mismo valor para cada uno. Ambos siguen prefiriendo encontrarse a no hacerlo –éste es el aspecto cooperativo del juego– pero sus preferencias respecto del sitio no coinciden –y éste es su aspecto conflictivo–. F_1C_1 es el mejor resultado para F y el peor para C; F_2C_2 es el segundo mejor para cada uno; F_3C_3 es el mejor resultado para C, pero el peor para F. Los tres resultados son técnicamente *óptimos de Pareto* en tanto no pueden ser mejorados

para ambos simultáneamente. Pero no existe ninguno que sea óptimo *para ambos.* Este juego mixto de conflicto y cooperación representa situaciones muy comunes en la vida real, como el regateo entre un comprador y un vendedor que prefieren realizar la transacción a frustrarla, pero que desean que se lleve a cabo en el punto de la *curva de contrato* más favorable a sus respectivos –y encontrados– intereses, y cuya ilustración formal es la llamada *caja de Edgeworth* (Buchanan, J. y Tullock, G., 1980: 126-131). La *batalla de los sexos* es el nombre con que se conoce en la teoría de juegos al conocido problema de coordinación o de negociación que se ilustra con variaciones de la historia de una pareja que prefieren cenar juntos a hacerlo cada uno por su lado, pero él prefiere ir a un restaurante mexicano y ella a uno francés.

El *dilema del prisionero* pone en evidencia la existencia de un conflicto, al parecer insoluble, precisamente entre estrategia y pago, es decir, entre *racionalidad estratégica* y *optimalidad.* La estrategia recomendada por la prudencia resulta literalmente *contraproducente –self-defeating–* (íd.: 76, 94-102; Parfit, D., 1991: *passim*), lo que constituye un escollo demasiado serio para una teoría plausible de la racionalidad.

5.3. Ordenaciones de preferencias y estrategias

Teniendo todo esto en cuenta, y llamando C y D a las estrategias o decisiones entre las que puede elegir cada uno de los dos jugadores A y B, existen cuatro *pares* de decisiones posibles:

— CC (cooperar A y B).
— CD (cooperar A y desertar B).
— DD (desertar A y B).
— DC (desertar A y cooperar B).

Si se limitan los casos posibles a aquellos en los que las preferencias de los jugadores por los resultados son simétricas existen *veinticuatro* ordenaciones posibles de resultados, de las cuales no todas son dilemas. Los dilemas se plantean cuando existe algún tipo de conflicto entre estrategias y pagos (Gauthier, D., 1998a: 98-104)

y por lo tanto entre cooperar y desertar para agentes maximizadores de su utilidad. En *doce* de los casos se prefiere el resultado CC al CD, es decir que a cada agente le va mejor si ambos cooperan que si sólo él lo hace y el otro deserta. En los otros *doce*, se prefiere el resultado DC al DD, esto es, que cada agente prefiere que el otro coopere aunque él mismo deserte. Y, por último, en *seis* de los casos coinciden ambas preferencias (CC > CD) y (DC > DD) –adoptando, por claridad tipográfica, el símbolo ">" en vez de "P", introducido en el capítulo 4 para representar la preferencia estricta–. Si los números representan magnitudes arbitrarias con el único propósito de indicar el orden de preferencias (3 > 2 > 1 > 0), se dan los siguientes tipos de juegos de estrategias en su forma normal:

Cuadro 5.6. *1. CC > CD > DC > DD (sin nombre específico)*

		B	
		C	D
A	C	3, 3	2, 1
	D	1, 2	0, 0

No es difícil advertir que esta situación es trivial y escasamente problemática. A ambos les compensa en todo caso cooperar, haga el otro lo que haga, pero en el caso de producirse una deserción cada uno prefiere que sea la del otro (CD > DC): que por él no quede.

Cuadro 5.7. *2. CC > DC > CD > DD (sin nombre específico)*

		B	
		C	D
A	C	3, 3	1, 2
	D	2, 1	0, 0

Es una variante del caso anterior, con una simple permuta del orden la segunda y tercera preferencias: sería lamentable que se frustrase el resultado cooperativo que ambos prefieren, pero de ser así cada uno prefiere ser el primero en abandonar (DC > CD).

Cuadro 5.8. *6. DC > DD > CC > CD (Punto muerto, Callejón sin salida, Atolladero, Deadlock)*

		B	
		C	D
A	C	1, 1	0, 3
	D	3, 0	2, 2

Tampoco este caso es particularmente problemático: haga lo que haga el otro es preferible desertar, por lo que la deserción de ambos (DD) es un *equilibrio de Nash* –ninguno puede obtener ninguna ventaja modificando unilateralmente su estrategia dada la estrategia del otro– y en este sentido es análogo al *dilema del prisionero.* Pero, a diferencia de lo que ocurre en éste, el resultado de la deserción de ambos en *punto muerto* es un *óptimo paretiano,* no hay ningún otro resultado factible que sea estrictamente preferido a él por alguno de los jugadores y al menos indiferente para el otro.

Los jugadores no creen que el resultado de la cooperación *de ambos* sea realmente ventajoso *para ambos.* La variedad de nombres con que el juego es conocido en la literatura científica apunta a las conocidas características de ciertos tipos de negociaciones colectivas en las que ninguno de los participantes tiene el menor interés en llegar realmente a un acuerdo y ni siquiera lo pretende: lo que de veras les interesa es lograr que el otro coopere (Poundstone, W., 1995: 325-326). El desarrollo del juego es una forma de marear la perdiz.

Cuadro 5.9. *3. CC > DC > DD > CD (Garantías, Caza del venado, Assurance)*

		B	
		C	D
A	C	3, 3	0, 2
	D	2, 0	1, 1

El antecedente más ilustre de este esquema de decisiones estratégicas es la situación descrita por Rousseau en la segunda parte del *Discurso sobre el origen de la desigualdad* y que a su juicio ilustra "cómo pudieron los hombre adquirir alguna idea grosera de los compromisos mutuos y de la ventaja de cumplirlos, aunque sólo en la medida en que podía exigirlo el interés presente y sensible, pues la previsión no significaba nada para ellos: lejos de ocuparse del porvenir lejano, ni siquiera se preocupaban del día siguiente" (Rousseau, J. J., 1971: 207). Cuando un grupo de cazadores primitivos se apostaba para cazar un ciervo –algo que necesitaba del concurso de todos– todos se mantenían fielmente en su puesto hasta que una liebre se ponía a su alcance; en ese caso "no cabe duda de que iría en pos de ella sin escrúpulos y, una vez cazada, le importaría bien poco hacer que sus compañeros perdieran la suya".

Es el caso típico que permite ilustrar los beneficios de la cooperación condicional para procurarse ciertos bienes que no pueden producirse en cantidades variables sino en *quanta*, traspasando determinado umbral crítico. Si la contribución individual, además de implicar un coste, no basta por sí sola para producir ese bien, cada uno preferirá cooperar si el otro coopera, pero desertar si el otro deserta. Así ocurre cuando, por ejemplo, es preciso dar una señal a cuenta como reserva de un autobús para una excursión, que no se devolverá si no se formaliza el contrato; pero que haya contrato depende a su vez de que se sobrepase un número preciso de excursionistas. Todos prefieren ir de excursión a que-

darse en casa, pero si al final no se logra organizarla, es aún peor quedarse en casa habiendo perdido la señal: los duelos con pan son menos.

La estructura de preferencias en *garantías* incita a cada jugador a hacer *lo mismo* que haga el otro: cada uno contribuirá si los demás contribuyen; pero si los demás no lo hacen, él tampoco lo hará. El mismo dilema se aplica a cada uno de los individuos que sinceramente desean participar en una protesta, huelga, motín o cualquier otra actividad colectiva que implica un coste individual y que –por hipótesis– no sólo no alcanzará sus objetivos a menos que participe un número *suficiente* de personas sino que, de quedar los participantes por debajo de ese umbral, las consecuencias para ellos y hasta para la propia causa serán nefastas. En estos casos la posibilidad de la *cooperación* depende de la eficacia de la *coordinación* que *garantice* –de ahí el nombre usual del juego– el cumplimiento del acuerdo. La confianza justificada, ya sea en la cooperación voluntaria de los demás o en la eficacia de algún tipo de coacción autoritaria, permite alcanzar el resultado preferido. La clásica solución práctica –que no teórica– al dilema del prisionero consiste en convertirlo en un juego de garantías: al transformar desde fuera la matriz de pagos cada cual, por la cuenta que le trae, preferirá inequívocamente cooperar.

Cuadro 5.10. *4. DC > CC > CD > DD (Gallina, Chicken)*

		B	
		C	D
A	C	2, 2	1, 3
	D	3, 1	0, 0

"Gallina" en sentido figurado se dice del cobarde o pusilánime. El juego que responde a esta matriz de preferencias reci-

be ese nombre por la imputación de cobardía que se hace al jugador que adopta una determinada estrategia. La ilustración más conocida del juego es la peculiar especie de torneo que se desarrolla en la película *Rebelde sin causa* dirigida por Nicholas Ray en 1955 y protagonizada por James Dean (Jim). En un momento determinado Jim se ve obligado a sostener un duelo de honor con Buzz, el jefe de la pandilla local de adolescentes. Ambos suben a sendos coches en una explanada que desemboca en un acantilado, en medio de la expectación de todos los jóvenes. El juego consiste en enfilar los coches a gran velocidad hacia el precipicio y *no* ser el primero en saltar –no ser *un gallina*– antes de que el coche se despeñe. El desenlace del torneo en la película es que Jim salta a tiempo pero Buzz, a pesar de intentarlo, se precipita al mar al trabarse la manga en la manilla de la puerta. En otras versiones reales de este tipo de competiciones, los coches avanzan el uno hacia el otro por una carretera estrecha; el objetivo de cada jugador es entonces hacer que el otro se aparte antes.

A diferencia del juego de *garantías*, en *gallina* la estructura de preferencias induce a cada jugador a hacer *lo contrario* que su contrincante. "Cooperar" equivale aquí a mantener el tipo; "desertar", exactamente a lo que su nombre indica. Cada uno prefiere por encima de todo quedar por valiente y vivir para contarlo –honra con barcos– pero como no puede haber más que un valiente vivo, si tiene razones para suponer que el otro está dispuesto a mantener la honra aun sin barcos, tirará la toalla y al menos salvará los barcos. No es éste, sin duda, el esquema de preferencia del *bonus miles* al que Juvenal exhorta a creer que no hay crimen mayor que *animum praeferre pudori* y, por conservar la vida, perder lo que la hace digna de ser vivida (Kant, I., 1994: 190). Pero los dilemas de este tipo no implican necesariamente una elección de tanta solemnidad moral.

El modelo se verifica en multitud de situaciones de la vida cotidiana, como por ejemplo en la pugna entre dos niños a ver quién aguanta más debajo del agua u otras semejantes como en el caso de la cooperación (en el sentido ordinario del término) en las tareas domésticas de dos egoístas amantes de la limpieza. Mantener la casa habitable requiere un *quantum* de trabajo, que

cada uno –a diferencia de lo que ocurría en *garantías*– ciertamente puede realizar por sí solo, aunque prefiere que sea el otro quien lo haga. Pero si ninguno lo hace la casa se vuelve inhabitable. Si cada uno tiene razones para suponer que el otro ciertamente *no* lo hará –que *cooperará* en el sentido estratégico y, en este caso, contraintuitivo del término– podría preferir hacerlo él solo –*desertar*– antes que soportar las incomodidades del piso sucio. Sabiendo esto, cada cual intentará *amenazar* al otro de forma creíble, dejando patente su negativa irrevocable a *desertar*, y no sólo mediante palabras sino por medio de acciones, por ejemplo, marchándose del piso y permaneciendo ilocalizable de modo que ya no sea posible la comunicación. Una forma particularmente eficaz de *cortar la comunicación* en este tipo de situaciones estratégicas es hacer ver al otro que uno actuará *irracionalmente*, fingiéndose loco o volviéndose provisionalmente loco. En el caso del duelo automovilístico, uno de los jugadores puede sentarse al volante manifiestamente borracho, ponerse gafas negras, tirar botellas vacías de licor o incluso el propio volante por la ventanilla: "si su contrincante lo ha visto todo, habrá ganado; si no lo ha visto, se ha metido en un lío" (Poundstone, W., 1995: 318-320; Schelling, T., 1964: 199-230; Parfit, D., 1984: 12-17).

Son estas características del modelo las que permiten entender la lógica de las situaciones estratégicas –ahora en el sentido ordinario del término– como las que se suscitan en las confrontaciones bélicas, en las negociaciones con secuestradores de aviones o en los juegos de póquer. Un caso típico de juego del *gallina*, objeto de detenidos análisis, fue el de la crisis de los *misiles* cubanos en 1962 (Poundstone, W., 1995: 292-320). Cuando los norteamericanos lograron convencer a los soviéticos de que su amenaza de ir a una guerra nuclear si ellos no retiraban los misiles *iba en serio*, les indujeron a actuar en función de sus segundas preferencias y desistir del juego. Existe una extensa literatura sobre la función que desempeñan las amenazas, los *faroles (bluffs)*, los intentos de *salvar la cara*, etc. en las interacciones estratégicas (Schelling, T., 1964: 51-59, 145-153; Davis, M., 1986: 115-118; Gauthier, D., 1998b: 119-159).

Cuadro 5.11. *5. DC > CC > DD > CD (Dilema del prisionero, Prisoner's dilemma)*

		B	
		C	D
A	C	2, 2	0, 3
	D	3, 0	1, 1

Es difícil entender el desarrollo de la teoría de juegos sin mencionar el papel que desempeñó en él uno de los *think tanks* más importantes de la reciente historia de la investigación científica: la RAND –*Research and Development*– Corporation, una peculiar institución fundada en California al término de la segunda guerra mundial y en los comienzos de la guerra fría. A su plantilla pertenecieron, con distinto grado de compromiso y en distintas etapas, los más destacados teóricos de las decisiones estratégicas, entre ellos John von Neumann y John Nash (Poundstone, W., 1995: 127-154; Nasar, S., 1998: 115-122). Los teóricos de RAND analizaban conductas ordinarias o proponían ciertos tests experimentales que implicaban la resolución de dilemas prácticos con objeto de refinar la comprensión de la conducta racional. A dos de ellos en particular, Merrill Flood y Melvin Dresher, les resultaba preocupante que determinados puntos de equilibrio de Nash, que constituían la solución teórica del juego, pudieran no ser razonables.

Con el propósito de averiguar "si las personas reales, especialmente aquellas que jamás habían oído hablar de Nash o de los puntos de equilibrio, llegarían sin saberlo a la estrategia de equilibrio" (íd.: 163) diseñaron un juego bipersonal a cien partidas que admitía dos estrategias: una, *cooperar* con el adversario, procuraba mejores resultados a *ambos*, pero era la *peor* en términos estratégicos, y otra, *desertar* o *defraudar*, que era la solución de Nash, pero era peor para *ambos* en términos de resultados. De hecho los resultados del experimento arrojaron mayoritariamen-

te un resultado cooperativo, cuando la teoría exige que los jugadores deserten en todas y cada una de las cien partidas.

Peor para los hechos: el propio Nash arguyó que precisamente por eso el experimento era deficiente como *test* de su teoría del punto de equilibrio, ya que en vez de estar diseñado como una secuencia de juegos independientes como en los casos de los juegos de suma cero, venía a constituir una especie de gran juego con múltiples partidas en el que había *demasiada interacción.* Nash se mostró no obstante asombrado "por lo ineficaces que eran ambos jugadores para obtener sus pagos: *se les habría creído más racionales"* (Nasar, S., 1998: 119; Poundstone, W., 1995: 173). Pero según los estrictos términos de la teoría, cuando un grupo de agentes se ve atrapado en el *dilema del prisionero* lo peor para cada uno es sin duda ser listo; en los grupos compuestos de tontos cada uno tiene mejor suerte: aunque su propia irracionalidad es peor para él, gana aún más gracias a la irracionalidad de los demás (Parfit, D., 1991: 21).

Flood y Dresher expusieron los resultados a sus colegas –incluido el propio Von Neumann– sin que advirtieran en el experimento ninguna especial trascendencia. Sin embargo uno de ellos, el matemático de Princeton Albert Tucker, decidió servirse de él en una conferencia sobre teoría de juegos a la que fue invitado por el departamento de psicología de la Universidad de Stanford. De hecho el experimento "le había parecido interesante desde una óptica más amplia que la teoría de juegos [y] puesto que el público, formado por psicólogos, tenía poca idea de teoría de juegos [decidió] que necesitaba presentar el juego inmerso en una historia. Tucker se inventó un relato con dilema incluido [y] acuñó el nombre 'el dilema del prisionero'" (íd.: 174). Un hallazgo, si no estrictamente *serendípico* (Roberts, R., 1992: 15), sí de inesperada fortuna.

El dilema del prisionero ha proporcionado el modelo formal para tal cantidad de análisis y experimentos en la ciencia política, en la economía, en psicología social, en filosofía moral y hasta en biología, que ha merecido ser comparado por su versatilidad con los servicios prestados por *Escherichia coli* en los experimentos de ingeniería genética (Axelrod, R., 1986: 38). Aunque en sentido propio no se trata del descubrimiento de un *novum:* el dilema no ha hecho sino formalizar y poner en evidencia la ine-

xorable lógica de determinadas situaciones consustanciales a la condición humana. Es "una paradoja con la que todos hemos de convivir; darse cuenta del dilema [...] es parecido a descubrir que el aire existe; siempre nos ha acompañado y la gente siempre lo ha notado, en mayor o menor grado" (Poundstone, W., 1995: 183-184).

La historia es muy simple: en la cárcel se hallan dos presos a los que el fiscal cree culpables de haber cometido un grave delito castigado con veinte años de cárcel, pero no dispone de pruebas suficientes para condenarlos por él. Posee, en cambio, pruebas para condenarlos por un delito menor, sancionado con una pena de dos años. Con el propósito de inducirles a confesar el delito mayor, hace una oferta a cada uno por separado, dejándoles claro que hace la misma oferta a ambos. La oferta es la siguiente:

1) Si *uno confiesa*, delatando al otro, y *el otro guarda silencio*, el delator quedará libre y el delatado cumplirá los *veinte* años de condena.
2) Si *ambos confiesan*, delatándose mutuamente, se les condenará por el delito grave, pero en recompensa por su colaboración con la justicia, se les rebajará la condena a sólo *diez* años.
3) Si *ninguno confiesa*, ambos serán condenados sólo por el delito menor a *dos* años.

Cuadro 5.12. *El dilema del prisionero*

		B	
		confiesa	calla
A	confiesa	–10, –10	0, –20
	calla	–20, 0	–2, –2

Cada uno de los presos se da cuenta de que lo que le interesa verdaderamente es confesar, *haga lo que haga el otro*. Si el otro

confiesa, entonces, *confesando también él*, reducirá su condena de veinte a diez años. Si el otro *no confiesa*, entonces, *confesando él*, quedará libre en vez de ser condenado a dos años. Cada preso tiene la impresión de que, no importa lo que haga el otro, lo mejor para él siempre es confesar. Así que ambos terminan confesando guiados por el propio interés y ambos van a la cárcel durante diez años. Si, en cambio, ninguno lo hubiera hecho ambos habrían pasado tan sólo dos años en prisión. A cada uno de ellos elegir racionalmente lo que le beneficia le cuesta ocho años de más en la cárcel. Si hubieran podido ponerse de acuerdo *y fiarse* cada uno de que el otro cumpliría lo acordado y ambos efectivamente lo hubieran cumplido, habrían logrado un resultado más acorde con los intereses de ambos. La paradoja del dilema del preso consiste en el hecho de que la decisión que se presenta a cada individuo como máximamente racional aboca a un resultado colectivamente irracional cuando es adoptada por ambos.

Como en los anteriores juegos, el esquema formal se verifica en numerosas situaciones reales. Un precedente clásico es el del estado de naturaleza tal como lo describe Hobbes. De hecho el fatídico orden de preferencias DC > CC > DD > CD que define el dilema es llamado por algunos "la trampa del Leviatán" y el propio juego, "el juego del Leviatán" (Hollis, M., 1988: 36). En esa miserable situación los hombres se enfrentan permanentemente a una situación que responde a la lógica del *dilema del prisionero*. Los acuerdos aparecen como una salida razonable y deseable de esa situación, pero no es posible hacerlos vinculantes por mero acuerdo verbal: el estado de naturaleza es, por definición, precisamente la ausencia de constricciones que haga que los hombres se atengan a los pactos (Taylor, M., 1987: 133; De Jasay, A., 1989: 40-69). En el estado de naturaleza cada uno prefiere la paz a la guerra, pero la desconfianza de cada uno hacia los demás le lleva a elegir la guerra a la paz. En última instancia el estado de guerra resultante no sería preferible para nadie pues todos prefieren la paz a la guerra. Por esta razón suscribir el pacto conviene a todos. Pero el pacto suscrito es un bien público del que no es posible excluir a nadie. Por ello, en cuanto egoísta, a cada uno le conviene más violarlo una vez suscrito. Pero existen razones para desconfiar de que los demás se atengan a él: es máximamente

racional hacerlo. De ahí el recurso a la garantía *externa* que presta el soberano, gracias a su poder de coacción, de que ningún infractor saldrá mejor librado si es detectado. Pero mantener una vigilancia y una coacción semejantes es costoso, probablemente ineficaz, y no garantiza la eliminación de *gorrones,* defraudadores ocultos y demás parásitos del bien público producido por la cooperación coactiva. Pero una cooperación espontánea requiere una transformación *interna* y una redefinición de la matriz de pagos de los jugadores racionales, que sería obra de la moralidad (Parfit, D., 1991: 13; Gauthier, D., 1998a: 113-117).

No sólo la filosofía política, sino también la literatura describe situaciones cuya estructura es la típica del dilema del prisionero. El cuento de Poe *El misterio de Marie Rogêt* podría ser incluso un antecedente literal de la historia ideada por Tucker. Tras el asesinato de la joven Marie se ofrece una enorme recompensa y "el completo perdón a cualquier cómplice que se presentara a declarar contra el autor del hecho". El razonamiento del *chevalier* Dupin es que "dada la enorme recompensa ofrecida y el pleno perdón que se ofrece por toda declaración probatoria, no cabe imaginar un solo instante que algún miembro de una pandilla de miserables criminales no [traicione] a sus cómplices. En una pandilla colocada en esa situación, cada uno de sus miembros no está tan ansioso de recompensa o de impunidad, como temeroso de ser traicionado. Se apresura a delatar lo antes posible, a fin de no ser delatado a su turno" (Poe, E., 1972: 463, 507-508; Poundstone, W., 1995: 185-186). Y precisamente el hecho de que *no* haya habido delación es la mejor prueba de que no ha habido una pandilla de criminales, sino un asesino solitario. Es decir, que la ausencia de delación demuestra que la decisión de callar *no ha podido ser* estratégica sino que *ha tenido que ser* paramétrica, y, en consecuencia, el crimen *ha sido* obra de un asesino solitario. Esta forma de argumentar del detective Dupin es un caso típico del razonamiento *abductivo* mencionado en el capítulo 1, en el que ciertos datos de la experiencia son interpretados a la luz de un determinado modelo de conducta racional de forma que la historia en su conjunto cobre sentido –*make sense.*

En la ópera de Puccini, el fallido pacto entre el jefe de policía Scarpia y Floria Tosca para salvar de la ejecución al amante de

ésta, Mario Cavaradossi *–cerchiamo insieme il modo di salvarlo–*, es un caso de mutua deserción: Scarpia ofrece a Tosca salvar la vida de su amado faltando a su deber a cambio de un *misero prezzo:* un instante de placer. Este resultado es para ambos preferible al que consiste en la frustración del deseo para el barón y la muerte del pintor para la cantante. Pero para cada uno la única alternativa racional es la de traicionar al otro: Tosca apuñala a Scarpia y Cavaradossi es realmente fusilado (Rapoport, A., 1974: 227-234).

La historia narrada por Gustavo Adolfo Bécquer en su Rima *Asomaba a sus ojos una lágrima* reproduce asimismo un dilema del prisionero. El poeta evoca pesaroso el día en que él y su amada estuvieron a punto de pedirse y concederse mutua y simultáneamente perdón por un agravio anterior, superando su orgullo en nombre de su mutuo amor. Éste les hace preferir que ambos cooperen –llorar ella, perdonar él– a que ambos deserten –enjugar el llanto ella, enmudecer él–; pero aquél hace que cada uno prefiera aún más ser él quien deserte y el otro quien coopere, que él perdone sin que ella se muestre arrepentida, o a la inversa. Haga lo que haga el otro, la estrategia racional para cada uno es desertar. El resultado inexorable es que ambos pierden más de lo que habrían podido ganar cooperando. No puede interpretarse de otra forma la historia contada: su conducta real ha revelado sus preferencias.

El *dilema del prisionero* guarda semejanzas y diferencias formales con los juegos anteriormente descritos. Como en los dos primeros –el 1, representado en el cuadro 5.6, el 2 en el 5.7– y en *punto muerto* –cuadro 5.8– en *prisionero* existe una *estrategia dominante* bien definida. Se dice de una estrategia E que *domina* a otra E_1 cuando: *a)* los resultados que proporciona a quien la adopta son siempre como mínimo *iguales* a los que conseguiría con la estrategia E_1, *sea cual sea la estrategia del otro jugador,* y *b)* al menos en una alternativa proporciona un resultado *superior* (Davis, M., 1986: 38-43). En los dos primeros juegos, la estrategia dominante es cooperar; en *punto muerto* y *prisionero*, desertar.

Tanto en *gallina* como en *prisionero* existe la tentación de desertar y cada uno prefiere desertar él si el otro coopera, pero si el otro deserta, en *gallina* es preferible cooperar mientras que en

prisionero es mejor desertar también. Si se comparan entre sí los *resultados* de *gallina*, *garantías* y *prisionero*, el producido por la cooperación de ambos es en los tres casos superior al de la deserción conjunta: el resultado de la falta de cooperación es un *subóptimo de Pareto*. Pero ni en *gallina* ni en *garantías* existe una estrategia dominante: todo depende de lo que haga el otro jugador.

Sólo en cuatro de los juegos existe una estrategia dominante: en 1 y 2 favorece la cooperación sin mayores problemas y el resultado es un *óptimo de Pareto*. En *punto muerto* la estrategia dominante a favor de desertar produce asimismo un resultado *óptimo de Pareto*. El problema está en el *dilema del prisionero*.

5.4. La tragedia de la racionalidad estratégica

Desde la perspectiva de los *resultados*, el de confesar ambos *no es* un óptimo paretiano; hay una alternativa –callar ambos– que ambos valoran más. Los otros dos resultados –callar uno y confesar otro– *son* óptimos de Pareto; en ellos se cumple que las alternativas que aportan más valor a *uno* de los prisioneros, implican una disminución de valor para el *otro*. Desde la perspectiva de las *estrategias*, las que determinan que ambos confiesen están en equilibrio, pues cada una de ellas es la mejor respuesta a cualquier otra posible estrategia del otro: sólo ellas constituyen un equilibrio de Nash. Así pues, "los resultados se dividen en dos conjuntos exclusivos y exhaustivos: el conjunto de los resultados que son óptimos de Pareto, y el conjunto de los resultados derivados de estrategias en equilibrio de Nash". Lo que demuestra el *dilema del prisionero* es que el conjunto resultante de la intersección de ambos conjuntos es un conjunto *vacío*: "no existe ningún principio completo de decisión en interacción [...] que incluya en su conjunto de decisión algún miembro del conjunto de estrategias en equilibrio de Nash *y* produzca además un resultado óptimo de Pareto. Ningún principio completo puede asegurar a la vez, en todas las interacciones, el equilibrio de Nash y la optimidad de Pareto" (Gauthier, D., 1998a: 95-96). Por eso, el *dilema del prisionero* es, como el *teorema de Arrow*, un teorema de *imposibilidad*.

Hay que tener muy presente esta condición del dilema a la hora de valorar los numerosos intentos prácticos, pero sobre todo teóricos, de escapar a una conclusión tan paradójica y chocante. En cuanto conclusión última de los axiomas de la racionalidad el dilema es inmune a la refutación empírica. En términos estrictamente teóricos no hay salida del dilema. Y, en la práctica, quien haya caído en la *trampa de Leviatán* ya puede *lasciare ogni speranza:* la gravedad del dilema es tal que, como un agujero negro, no deja escapar nada de lo que cae en él, ni siquiera la luz de la razón. Argumentar que el sentido del honor aun entre criminales, o el temor a las futuras represalias en el caso de los prisioneros de la historia, o una mayor generosidad en los arrepentidos amantes de la *Rima* de Bécquer habrían resuelto el problema es incurrir en una *ignoratio elenchi.* Los ladrones que también son gente honrada, los que sin serlo saben lo que les puede costar su delación y los verdaderos enamorados por definición *no* se encuentran en un dilema del prisionero, porque su orden de preferencias es *otro.* Tal vez el hecho mismo de tener tales preferencias les proteja de caer en la trampa y, gracias a ello más que *resolver* el dilema se logra *disolverlo* en la práctica. Pero no hay que creer que el dilema es el justo castigo a la perversidad, al egoísmo o a la mezquindad. El dilema no lo plantea el *contenido* de las preferencias: si en la historia original del dilema en vez de los propios prisioneros son sus abogados los que tienen que elegir, éstos se enfrentan entonces al *dilema de los abogados de los prisioneros*; pero si ambos abogados dan prioridad a sus propios clientes –como, por otra parte es su *obligación*– el resultado para éstos será peor que si ninguno la da. En definitiva, "todo dilema prudencial genera así un dilema moral; si un grupo se enfrenta con el primero, otro grupo podría, como consecuencia, enfrentarse con el segundo. Podrá ser así si creemos que cada miembro del segundo grupo debe dar prioridad a alguno de los miembros del primero. El problema proviene del hecho mismo de dar prioridad; da igual que se la dé a uno mismo o a otros" (Parfit, D., 1991: 37).

Sin embargo, los datos de la experiencia cotidiana muestran inequívocamente que las personas obran por principios, respetan los acuerdos, cumplen las promesas, actúan de forma altruista y cooperan espontáneamente. A todas estas formas de actuar

se las engloba en el concepto genérico de *moralidad.* Si los hombres fueran por naturaleza, o se hubieran hallado en algún momento en el estado de naturaleza, de la manera descrita por Hobbes, la teoría habría predicho que jamás alcanzarían el estado presente, por las mismas razones que predecía que los jugadores racionales del experimento de Flood y Dresher *tenían que* haber desertado en cada una de las partidas del juego, algo que los resultados no confirmaron. Nash negaba que los resultados del experimento refutaran la teoría, precisamente porque el diseño del *test* introducía una dimensión temporal y sucesiva que le era ajena y que generaba un exceso de interacción. En la teoría todos los elementos pertinentes para la decisión están dados en el *instante* de elegir, y por tanto cada elección es completa y cerrada en sí misma.

En el caso particular del dilema la condición instantánea de la decisión y la ausencia de interacción han permitido poner en duda hasta qué punto puede considerarse un caso de racionalidad *estratégica.* Una de las características definitorias de las situaciones de este tipo es que las decisiones de los diversos agentes no son probabilísticamente independientes entre sí, a diferencia de lo que ocurre en las paramétricas. El que decide salir de su casa con un paraguas lo hace en función de lo que crea que va a *pasar,* pero su decisión no afecta a la probabilidad de que llueva. Más aún: según crea que puede pasar una cosa *u* otra, llevará o no el paraguas. De quien lleve *siempre* un paraguas *pase lo que pase* podría decirse que, más que tomar una decisión, obra *por principio* o *por norma.* Pero en el *dilema* cada jugador deserta *siempre, haga el otro lo que haga* y por lo tanto su decisión –si de tal puede hablarse– está irrevocablemente fijada sin posibilidad de influir o ser influida por la del otro.

Robert Nozick propone un desconcertante experimento mental que plantea de forma ingeniosa las paradojas de la predicción de la decisión (Nozick, R., 1985: 107-133; Campbell, R., 1985: 21-31; Lewis, D., 1985: 251-255; Sobel, J., 1985: 263-274; Resnik, M., 1998: 185-189). El *problema de Newcomb,* llamado así por el físico William A. Newcomb que lo formuló originariamente en 1960, plantea la siguiente situación: un ser omnisciente P (predictor), que es capaz de predecir las decisiones de un

agente D (decisor), coloca a éste ante dos cajas, C_1, transparente, que contiene \$1.000, y C_2, opaca, que puede o bien estar vacía o bien contener \$1.000.000. P propone a D –evidentemente, sin comunicarle cuál es su predicción– que analice la situación y se decida por la más ventajosa de estas dos opciones:

(1) quedarse con ambas cajas (C_1 y C_2) o
(2) quedarse sólo con C_2

haciéndole saber que, según haya previsto que D elegiría

(1) habría dejado la caja C_2 vacía;
(2) habría puesto \$1.000.000 en ella.

Cuadro 5.13. *El Problema de Newcomb*

		P	
		no ha puesto el dinero	ha puesto el dinero
D	abre C_1 y C_2	\$ 1.000	\$ 1.001.000
	abre sólo C_2	\$ 0	\$ 1.000.000

El decisor D puede razonar de una de estas dos maneras:

1. Como la predicción de P es cierta, es preferible elegir sólo la caja C_2 en la expectativa cierta de ganar \$1.000.000:

   — Si se decidiese por ambas cajas, P lo habría previsto y habría dejado vacía la caja C_2; de modo que es preferible no coger ambas.
   — Si se decidiese sólo por C_2, P también lo habría previsto, y habría puesto \$1.000.000 en ella, de modo que es preferible escogerla.

Este argumento se basa en el principio de *maximización de la utilidad esperada* y en el de la *cosa segura:* entre dos loterías equiprobables es preferible la que ofrece un premio mayor. Un espectador del experimento que tuviese esa misma información apostaría sobre seguro a que si D sólo coge la caja C_2, ganaría $1.000.000 y si coge las dos, ganaría sólo $1.000. *Por lo tanto es racional para* D *abrir sólo la caja* C_2.

2. Como la predicción de P *ya está hecha,* el contenido de la caja *está ya decidido* y no va a variar, elija D lo que elija, en todo caso le compensa coger *ambas cajas* pues:

   — Si C_2 está llena, ganará $1.000 + $1.000.000 = $1.001.000.
   — Si C_2 está vacía, al menos ganará $1.000.

   El argumento en favor de abrir ambas cajas se basa en el principio de la *estrategia dominante,* que proporciona resultados como mínimo iguales a los que conseguiría con cualquier otra estrategia propia, sea cual sea la estrategia ajena, y al menos en una alternativa proporciona un resultado superior. *Por lo tanto es racional para D abrir ambas cajas.*

El argumento no depende de que sea cierta la suposición de un predictor omnisciente en sentido estricto. Basta con suponer que sus habilidades para predecir conductas suscitan suficiente confianza, por ejemplo porque en la mayoría de los casos en los que llevó a cabo el mismo experimento los que escogieron sólo la caja C_2 consiguieron el millón de dólares, mientras que los que eligieron las dos cajas se quedaron sin nada –la avaricia rompe el saco–. Para subrayar la fuerza del argumento a favor de elegir ambas casas, Nozick supone que el predictor ha formulado su predicción, ha tomado la decisión de poner o no el millón en la caja opaca y se ha ido, pero antes ha permitido que el mejor amigo del decisor vea lo que hay de hecho en ella pero sin que pueda decírselo. Si el dinero está o no ahí no va a desaparecer o a aparecer súbitamente como consecuencia de la elección del decisor.

Si se le pide a éste que imagine lo que mentalmente le está aconsejando ese buen amigo ¿puede suponer que sea algo distinto de elegir las dos cajas? Pero tampoco es necesario suponer una *pre*dicción en sentido temporal, pues también pueden *retro*decirse acontecimientos pasados: es un mero recurso dramático describir el millón como estando o no estando *ya* en la caja *antes* de la decisión, pues lo único que importa es hacer ver que conseguir o no el millón es causalmente independiente de la decisión y, en consecuencia, de la predicción (Lewis, D., 1985: 252).

Si el *dilema del prisionero* se concibe como una situación *estratégica* –incluso con *demasiada interacción*– en la que las decisiones de ambos jugadores son probabilísticamente dependientes y si, además, se asignan utilidades (cardinales) a los diversos resultados, la elección racional es cooperar pues maximiza la utilidad esperada de cada agente condicionada a que el otro coopere. De hecho se elige entre dos resultados en *dos columnas* distintas de la matriz de decisión. Equivale en el *problema de Newcomb* a abrir sólo la caja C_2.

Si, por el contrario, se considera como una situación *paramétrica* en la que las decisiones de ambos son causal y probabilísticamente independientes, la elección se lleva a cabo entre dos resultados dentro de *una misma columna* de la matriz de decisión. En esta situación lo racional es elegir la estrategia dominante de desertar, que proporciona el mejor resultado en cualquiera de las columnas determinada por la decisión del otro. Corresponde a elegir las dos cajas en el *problema de Newcomb* (Campbell, R., 1985: 16-32).

El conflicto entre los dos principios de elección parece por tanto insoluble y *dentro* del modelo teórico que lo hace posible no cabe siquiera el recurso a un metaprincipio que arbitre entre ambos. Si el *dilema del prisionero* fuese simplemente una extravagante curiosidad de laboratorio, un experimento de praxeología recreativa sin trascendencia práctica, aunque de indudable interés teórico, como las paradojas de Aquiles y la tortuga, la del mentiroso o la del examen imprevisto, carecería de relevancia para la filosofía moral. Pero es mucho más que eso: representa formalmente la situación en que el agente racional *tal como lo define la teoría* se pregunta por las razones para actuar moralmente o

para convertirse en un agente moral –siquiera en el sentido ciertamente limitado de la moralidad entendida como cooperación o altruismo– dando por sentado que la respuesta ha de ser satisfactoria *en los términos de la teoría.* La estrategia dominante de la mutua deserción no resuelve el problema práctico, ella *es* precisamente el problema. La estrategia cooperativa o moral se justifica porque maximiza la utilidad esperada a condición de que (los) otros cooperen: se trata de hacerla plausible para el agente que ha de adoptarla. De hecho "las soluciones morales son [...] con frecuencia las mejores, y [...] las únicas asequibles. Por tanto, necesitamos motivos morales. ¿Cómo se pueden introducir? Afortunadamente, ése no es problema nuestro. Existen. Así es como resolvemos muchos Dilemas del Prisionero. Lo que necesitamos es reforzar esos motivos y difundirlos más ampliamente. Para esta tarea ayuda la teoría. El Dilema del Prisionero tiene que ser explicado. Sus soluciones morales tienen que serlo también" (Parfit, D., 1991: 16). Hay que explicar en términos de la teoría cómo es posible que los hombres naturalmente prisioneros del dilema a causa de su preferencia por la estrategia dominante hayan podido elegir racionalmente desecharla a favor de la estrategia cooperativa. Es decir, cómo han logrado convertir el *dilema del prisionero* en un juego de *garantías mutuas.*

Diversos autores aceptaron parte del veredicto de Nash sobre el experimento de Flood y Dresher. En efecto, no refutaba la teoría de la elección *concebida en términos paramétricos* de forma estática e instantánea. Pero mostraba que los agentes podrían percibir o definir la situación *en términos estratégicos* si existía suficiente interacción *a lo largo del tiempo.* No es lo mismo, argüían, decidir cooperar con un desconocido con el que se cruza uno solamente una vez en la vida que hacerlo en sucesivas ocasiones con los mismos individuos. Se aceptaba entonces que si en vez de concebir el *dilema* como un juego a una sola partida, se lo concebía como una especie de superjuego o de metajuego extendido a través de múltiples partidas en el tiempo, en el que los jugadores guardaran *memoria* de las interacciones pasadas y las reajustaran para el futuro, podía *explicarse* cómo se consolidaba la estrategia de la cooperación condicional. Anatol Rapoport (1960), Nigel Howard (1971) y, más recientemente, Robert Axelrod han inten-

tado responder, en palabras de este último, a la misma pregunta: "¿en qué casos debe una persona cooperar con otra, y en qué casos ser egoísta, en el curso de una relación que puede durar mucho tiempo?" (1986: 9) o, de forma más general "en las situaciones en las que cada uno de los individuos tenga un incentivo para ser egoísta, ¿cómo podrá llegar a desarrollarse la cooperación?" (íd.: 15). Mediante un torneo entre programas informáticos cada uno de los cuales simulaba distintos tipos de estrategias cooperativas y desertoras –desertar siempre, cooperar siempre, hacerlo al azar, etc.– Axelrod halló que la estrategia triunfadora era la más sencilla: *"donde las dan las toman" (tit for tat).* En las sucesivas interacciones jamás era la primera en desertar, pero respondía con la deserción a la primera deserción del adversario y, a continuación, pagaba con la misma moneda las ulteriores decisiones de éste, cooperando si cooperaba, desertando si desertaba (íd.: 37-74; Hofstadter, D., 1982; 1983a; 1983b; Poundstone, W., 1995: 353-373). Aunque con determinadas limitaciones puede afirmarse que se trata de una estrategia evolutiva lo suficientemente estable como para desarrollar un entorno de cooperación bastante resistente a los predadores.

Puede concederse que éste sea un buen modelo *explicativo* de la evolución de la cooperación, de la *emergencia de las normas* (Ullmann-Margalit, E., 1977). Aún quedaría por justificar en términos de la teoría por qué uno *debería* tomar la *primera* decisión cooperativa, cuando la teoría concibe toda decisión como la mejor *respuesta* a la estrategia ajena. Entre agentes que comparten el conocimiento común de la racionalidad resulta tentador suponer una especie de principio de coordinación que rezara más o menos así: "si un resultado es estrictamente preferido a cualquier otro resultado por ambos jugadores, entonces la racionalidad exige que cada jugador elija la estrategia que produce dicho resultado". Y, en efecto, así sería si fuera aplicado por *ambos* jugadores. Pero como la unidad de *agencia* de la teoría es el individuo maximizador de su utilidad esperada, la razón que *cada uno* tiene para aplicarlo es puramente instrumental. Sería racional para cada uno aplicar el principio de coordinación *si espera* que el otro también lo aplica, pero no se ve en qué se basa esa expectativa. Sería incluso deseable que el principio fuera una exigencia de la racionali-

dad, de modo que pudiera contar con que el otro lo va a aplicar. Pero "sigue teniendo sentido preguntar ¿por qué la racionalidad exige que sea *yo* el que lo aplique?". Para demostrar a un jugador que es racional *para él* aplicar el principio es necesario demostrar previamente que es racional que *el otro* lo aplique (Hollis, M. y Sugden, R., 1993: 11-12).

Otro problema al que se enfrentan las estrategias condicionalmente cooperativas se plantea cuando existe un plazo definido para desarrollar y para terminar la cooperación, aspecto de la interacción que sólo la *forma extendida* del juego permite representar de manera adecuada. Nash tenía razón al afirmar que en *cada* jugada lo racional es desertar. Si tiene algún sentido cooperar lo es en la medida en que se esperan beneficios futuros. Pero cuando llega la *última* ocasión de cooperar –el último nodo del grafo– ya no hay futuro alguno, por lo tanto lo racional es desertar en *esa* ocasión; el problema se plantea ahora en la penúltima: como la cooperación tampoco tiene ya futuro es preferible desertar también en *esa* ocasión. Retrocediendo inductivamente hasta la *primera* ocasión de cooperar –el primer nodo del grafo– se ve que no existen razones para hacerlo en *ningún* caso (Poundstone, W., 1995: 341-344; Hollis, M. y Sugden, R., 1993: 15).

Todas las estrategias condicionalmente cooperativas, orientadas por definición al futuro, parten del supuesto de que se pueden obtener beneficios a largo plazo gracias a la reputación adquirida por las anteriores decisiones cooperativas (Gauthier, D., 1998a: 108). Parte de los problemas de la cooperación condicional proceden de la dificultad de identificar a los posibles cooperadores y distinguirlos de los defraudadores ocultos: la reputación es en sí misma más un asunto de *percipi* que de *esse*, y ya Glaucón hacía ver a Sócrates que "el malo, si quiere serlo a conciencia, debe realizar con destreza sus malas acciones, de forma que nadie se dé cuenta. Y al que se deje atrapar hay que tenerle por tonto –*phaûlos*– pues el colmo de la maldad es parecer bueno y no serlo" (Platón, 1969, 62 [361a]). Gauthier es consciente de este problema y no sólo desliga su argumento a favor de la cooperación condicional de toda referencia a la reputación, sino que considera incluso que "es aplicable a un conflicto que ocurra una sola vez. Supongamos que cada persona supiera –no importa

cómo– que una (y sólo una) vez en su vida, ha de afrontar un conflicto entre estrategia y pago. Si esta persona pudiera identificar con certeza la disposición de los demás agentes en esa situación, y esperara que los demás identificaran igualmente la suya, estaría racionalmente justificado para ella adoptar la disposición condicionalmente cooperativa" (Gauthier, D., 1998a: 108).

Pero esto sigue dejando sin solucionar, *en los términos de la teoría* y para cada una de las personas, por qué debería ser *ella* quien diera el primer paso para cooperar. Para la teoría toda decisión estratégica es una *respuesta* a otra estrategia. En la medida en que la moralidad como cooperación es concebida en términos de racionalidad estratégica es imposible explicar ni justificar la decisión de tomar una iniciativa *espontánea.* Y, sin embargo, también sigue siendo cierto que sería mejor que no existiesen situaciones como las del *dilema* y, si somos prudentes, lamentaremos su existencia misma (Parfit, D., 1991: 21), porque la ordenación de preferencias que arma la *trampa de Leviatán* priva a los agentes maximizadores incluso de los beneficios reales que proporciona la cooperación con sus semejantes. Hasta tal punto –y éste es el núcleo del argumento de Gauthier (1998a: 108-117)– que al egoísta, sin necesidad de abandonar el objetivo de que le vaya lo mejor posible, le iría mejor dejando de serlo, adoptando voluntariamente una restricción *interna* de su comportamiento maximizador en la forma de un principio de decisión no maximizador, es decir *moral.* Los que *ya* son agentes morales están directamente dispuestos a cooperar, a practicar "virtudes tradicionales como decir la verdad y mantener las promesas, como la honestidad, gratitud y la benevolencia recíproca" (íd.:113), algo imposible de justificar en los propios términos del modelo de la elección racional; pero les va mejor. Por lo tanto, los que *no* lo son harían bien –para sí mismos– no ya sólo adquiriendo la *disposición* a la cooperación condicional cuyas limitaciones han quedado expuestas, sino *convirtiéndose* en agentes morales (íd.: 117).

No es éste el único caso en que una teoría práctica recomienda *adquirir* ciertos motivos o disposiciones cuya posesión habitual permitirá a los agentes actuar produciendo las mejores consecuencias, aun cuando ello implique *no* actuar en los términos estrictos de la teoría. Aunque el utilitarismo es primariamente una teoría sobre lo que hace buena o mala una acción, de forma derivada ofrece un pro-

cedimiento de decisión que incluye recomendaciones acerca de los motivos más eficaces para la promoción de los fines señalados por la teoría. Por ello el propio Sidgwick advertía contra el error de interpretar que "la doctrina según la cual la Felicidad Universal es el *criterio* último implica que la Benevolencia Universal es siempre el mejor, o el único correcto, de los *motivos* de actuación". No sólo no lo implica, sino que incluso podría implicar lo contrario, pues "si la experiencia muestra que la felicidad universal se logrará de forma más satisfactoria si los hombres obran con frecuencia por motivos diferentes de la pura filantropía universal, es obvio que esos otros motivos deben preferirse razonablemente con arreglo a principios utilitaristas" (Sidgwick, H., 1981: 413). A esta necesidad teórica y práctica responde el llamado *utilitarismo de los motivos* (Adams, R., 1993): si la obligación de cumplir las normas morales o legales se justifica por la utilidad de su cumplimiento general, puede ser útil adquirir la disposición habitual a no juzgar de su utilidad en cada caso concreto considerándolas en la práctica como categóricas, aun cuando la teoría sostenga que son realmente hipotéticas. Lo problemático es que hay que justificar en términos de la teoría la decisión de adoptar la disposición a *no* decidir en los términos propios de la teoría.

Tampoco es la única solución con tintes paradójicos. Nada impide en principio a una teoría coherente con el principio consecuencialista que atribuye valor intrínseco sólo a los estados terminales de cosas desaconsejar, en nombre de ese mismo principio, su propia adopción como moralidad pública. Sidgwick reconoce que, con criterios utilitaristas, podría ser moralmente correcto hacer y recomendar en privado lo que no sería correcto defender en público. Como la conciencia moral del hombre ordinario repudia la noción de una moral esotérica y existen razones utilitarias para mantener esta opinión común con carácter general, "la conclusión utilitarista, cuidadosamente formulada, podría ser la siguiente: que la opinión que afirma que el secreto podría hacer correcta una acción que en caso contrario no lo sería debería ser mantenida comparativamente en secreto; y, de igual modo, la doctrina que defiende la conveniencia de la moralidad esotérica debería seguir siendo esotérica. O, si es difícil mantenerla oculta, podría ser deseable que el sentido común repudie las doctrinas que es conveniente reservar para una minoría ilustrada" (Sidgwick, H., 1981: 489-490). Pero como la teoría misma

manda hacerse con los mejores motivos y disposiciones, con sus correspondientes creencias y emociones, se lograría un resultado mejor revisando las creencias morales, dejando de creer en la teoría consecuencial y adoptando en su lugar alguna versión refinada de la moralidad del sentido común. En términos coloquiales diríase que el mejor favor que se hace la teoría a sí misma es recomendar que no se la crea. Afirmaría así su propia verdad, a la inversa del principio de Berkeley, desvaneciéndose de la percepción. Una teoría que concibiese el actuar conforme a los principios de la propia teoría como un mero medio para alcanzar los objetivos de ésta, podría incluso recomendar que no se le hiciera caso, que se la olvidara o que se la sustituyera por otra. Poseería la paradójica propiedad de ser, en la terminología de Parfit, autoevanescente –*self-effacing*– (Parfit, D., 1984: 40-43). Ya advertía Ryle del "aroma de absurdidad" que exhala el propósito expreso de "olvidar la diferencia entre el bien y el mal", sobre todo cuando la única razón para hacerlo no se basa en dudas intelectuales sobre nuestras viejas creencias, sino en la convicción de que se lograrían mejores resultados si tuviésemos creencias distintas (Ryle, G., 1958: 155-159). Acechan aquí las paradojas que suscitan el autoengaño o la decisión de creer a voluntad (Parfit, D., 1984: 40-43; Williams, B., 1986).

Al hablar de *convertirse* en agente moral no puede prescindirse de las connotaciones que ya estaban presentes en el más remoto de los antecedentes conceptuales del término –el griego *metánoia*–, que implica tanto cambiar de sentimientos y de opiniones como arrepentirse –*paenitere*– y mudar de vida. No se trata de un mero juego literario. Como en el caso del argumento que propone Pascal en su conocida *apuesta*, se parte de una situación de hecho en la que es forzoso elegir: "hay que apostar; eso no es voluntario: se está ya embarcado" (Pascal, B., 1963: 550, § 418), y aunque la razón teórica no logra demostrar que Dios existe hace razonable la posibilidad de que exista; las infinitas ganancias que se derivarían de creerlo si fuera cierto compensan suficientemente los costes de vivir de acuerdo con esa fe. Apostar no es por tanto "trabajar por convencerse por el aumento de las pruebas de Dios, sino por la disminución de las pasiones". Hay que unirse a los que ya han apostado, "gentes que conocen el camino que querríais seguir y que ya están curados del mal del que querríais curar". Y comenzar por donde ellos comenzaron, "hacien-

do como si creyesen, tomando agua bendita, haciendo decir misas, etc.: naturalmente incluso eso os hará creer *et vous abêtira*". Aunque resulte incómodo desde la perspectiva de la razón teórica, "nada hay que perder". Uno se volverá "fiel, honrado, agradecido, bienhechor, amigo sincero, veraz" y, aunque no disfrute de "los placeres infectos, los deleites y la gloria, tendrá otros" (íd.: 551; Rescher, N., 1985: 1-19). El egoísta racional, perplejo ante la impotencia del modelo para armonizar estrategias y pagos y viéndose forzado a tomar decisiones prácticas ineludibles, apostaría por unirse a una comunidad de agentes morales convirtiéndose en uno de ellos.

Si ésta es la propuesta final del modelo más ambicioso y coherente de la racionalidad práctica, es preciso reconocer que guarda un sorprendente parecido con la penúltima proposición del *Tractatus* de Wittgenstein: "mis proposiciones son esclarecedoras de este modo; que quien me comprende acaba por reconocer que carecen de sentido, siempre que el que comprenda haya salido a través de ellas, en ellas, fuera de ellas. (Debe, pues, por así decirlo, tirar la escalera después de haber subido). Debe superar estas proposiciones; entonces ve correctamente el mundo" (Wittgenstein, L., 1973: 6.54).

El *dilema del prisionero* demuestra ser, sin lugar a dudas, el *experimentum crucis* del modelo de racionalidad que emplea la teoría de la elección racional para el análisis de la conducta cooperativa. En términos de la coherencia formal e interna del modelo, el que la solución en equilibrio del Dilema sea la no-cooperativa es sólo la conclusión necesaria de las premisas. Pero ello no prueba que las premisas mismas sean necesariamente plausibles. Como todo modelo es diseñado para cumplir funciones científicas, ni sus conceptos primitivos ni las premisas que los contienen pueden ser simplemente estipulados. Su valor epistémico depende de su isomorfismo con la realidad, es decir de una hipótesis plausible que permita su interpretación sustantiva y material. Gran parte del atractivo de los modelos de racionalidad que aplica la teoría de la elección racional, y en particular la teoría de juegos, procede precisamente de la aparente plausibilidad empírica de que gozan los conceptos primitivos de interés, utilidad o preferencia. Pero también puede argüirse que ciertas características de ese modelo restringen indebidamente el abanico de posibles soluciones; la principal de ellas, concebir las razones para actuar en términos consecuencialistas,

basados en expectativas futuras. La exclusión de otro tipo de razones, orientadas por la coherencia con principios, no sólo no está justificada sino que refleja inadecuadamente importantes dimensiones de la racionalidad práctica. Es plausible suponer que determinadas paradojas y dilemas de la racionalidad estratégica se deben precisamente a una muy concreta interpretación de la naturaleza de las razones que son capaces de motivar a la acción. Supuestos tan irrenunciables a la teoría como el del "conocimiento común de la racionalidad", con la consecuente transparencia (o al menos translucencia) racional de agentes movidos únicamente por razones interesadas –prudenciales o *forward-looking*– tal vez sean internamente incoherentes, pero sin duda hacen racionalmente imposible explicar y justificar la cooperación y, por extensión, la moralidad en los términos estrictos de la teoría. Un examen más detenido de la plausibilidad de los presupuestos psicológico-filosóficos del modelo permitiría comprender mejor la función específica que podrían desempeñar en la elección racional junto a las preferencias y las utilidades las razones *back-* o *inward-looking*, esto es, los principios y las normas.

El problema originario al que según Braithwaite respondía la teoría de juegos no era simplemente el paso del entorno paramétrico del Robinson solitario al entorno estratégico de la cooperación mutuamente interesada con Viernes. Ya se ha visto cuán problemático resulta para una teoría de la racionalidad prudencial hacer frente a los problemas de simple cooperación –no ya *equitativa*– por medio de estrategias, directas o indirectas, pero en todo caso estrictamente interesadas. Sin duda se hace necesario transformar los presupuestos de la racionalidad paramétrica en la medida necesaria para hacer frente a las situaciones estratégicas. Tal vez se entienda mejor ahora por qué se ha afirmado más arriba que la mutación ha de ser mucho más radical: no se trata de un mero aumento de complejidad pero sin solución de continuidad *dentro* de un universo homogéneo de racionalidad, sino de una transformación *de* ese mismo universo. En definitiva, la necesidad de dar cuenta de la realidad de la cooperación obliga a revisar, por razones meramente filosóficas y no sólo por motivos prácticos, el propio modelo de racionalidad.

Bibliografía

Abel, T. (1964): "La operación llamada 'Verstehen'", en Horowitz, I. (ed.): *Historia y elementos de la sociología del conocimiento.* Eudeba. Buenos Aires.

Adams, R. (1993): *Motive Utilitarianism.* Dartmouth. Aldershot.

Aristóteles (1959): *Ética a Nicómaco.* Centro de Estudios Constitucionales. Madrid.

Arrow, K. (1974): *Elección social y valores individuales.* Instituto de Estudios Fiscales. Madrid.

— (1986): "Valores y decisión colectiva", en Hahn, F. y Hollis, M. (eds.): *Filosofía y teoría económica.* Fondo de Cultura Económica. México.

Axelrod, R. (1986): *La evolución de la cooperación. El dilema del prisionero y la teoría de juegos.* Alianza Editorial. Madrid.

Barry, B. (1974): *Los sociólogos, los economistas y la democracia.* Amorrortu. Buenos Aires.

Bhargava, R. (1992): *Individualism in social science. Forms and limits of a methodology.* Clarendon Press. Oxford.

Becker, G. (1980): "El enfoque económico del comportamiento humano". *Información Comercial Española,* 557.

— (1983): *El capital humano. Un análisis teórico y empírico referido fundamentalmente a la educación.* Alianza Editorial. Madrid.

Benn, S. y Mortimore, G. (Eds.) (1976): *Rationality and the social sciences.* Rotuledge and Kegan Paul. Londres.

Berne, E. (1966): *Los juegos en que participamos.* Diana. México.

Bernstein, R. (1979): *Praxis y acción.* Alianza Editorial. Madrid.

Bloch, A. (1997): *La ley de Murphy.* Temas de Hoy. Madrid.

Braithwaite, R. (1955): *Theory of games as a tools for the moral philosopher.* Cambridge University Press.

Buchanan, J. (1984): *Política sin romanticismos.* Instituto de Estudios Económicos. Madrid.

— (1987): *La razón de las normas.* Unión Editorial. Madrid.

— y Tullock, G. (1980): *El cálculo del consenso. Fundamentos lógicos de una democracia constitucional.* Espasa Calpe. Madrid.
Cabrillo, F. (1996): *Matrimonio, familia y economía.* Minerva. Madrid.
Campbell, R. (1985): "Background for the uninitiated", en Campbell, R. y Sowden, L. (Eds.): *Paradoxes of rationality and cooperation: Prisoner's Dilemma and Newcomb's Problem.* University of British Columbia Press. Vancouver.
Caplow, T. (1974): *Dos contra uno: teoría de coaliciones en las tríadas.* Alianza Editorial. Madrid.
Casahuga, A. (1985): *Fundamentos normativos de la acción y la organización social.* Ariel. Barcelona.
Casas, J. (1984): *El análisis económico de lo político.* Instituto de Estudios Económicos. Madrid.
Cicerón (1950): *De fato.* Les Belles Lettres. París.
— (1987): *Del supremo bien y del supremo mal.* Gredos. Madrid.
Colomer, J. (1987): *El utilitarismo: una teoría de la elección racional.* Montesinos. Barcelona.
— (1990): *El arte de la manipulación política: votaciones y teoría de juegos en la política española.* Anagrama. Barcelona.
Cruz, M. (1997): *Acción humana.* Ariel. Barcelona.
Dahrendorf, R. (1973): *Homo sociologicus. Un ensayo sobre la historia, significado y crítica de la categoría del rol social.* Instituto de Estudios Políticos. Madrid.
Davis, M. (1986): *Introducción a la teoría de juegos.* Alianza Editorial. Madrid.
De Jasay, A. (1989): *Social contract, free ride. A study of the public goods problem.* Clarendon Press. Oxford.
Diermeier, D. (1996): "Rational Choice and the role of theory in political science". En Freidman, J. (Ed.): *The Rational Choice Controversy.* Yale University Press. New Haven.
Dilthey, W. (1986): *Introducción a las ciencias del espíritu.* Alianza Editorial. Madrid.
Downs, A. (1973): *Teoría económica de la democracia.* Aguilar. Madrid.
Dworkin, R. (1984): *Los derechos en serio.* Ariel. Barcelona.
Eco, U. y Sebeok, T. (1989): *El signo de los tres. Dupin, Holmes, Peirce.* Lumen. Madrid.
Elster, J. (1984): "Marxismo, funcionalismo y teoría de juegos. Un alegato en favor del individualismo metodológico". *Zona Abierta,* 33.
— (Ed.) (1986): *Rational choice.* Basil Blackwell. Oxford.
— (1987): *Uvas amargas.* Península. Barcelona.

— (1989): *Ulises y las sirenas. Estudios sobre racionalidad e irracionalidad.* Fondo de Cultura Económica. Mexico.
— (1990): *El cambio tecnológico. Investigaciones sobre la racionalidad y la transformación social.* Gedisa. Barcelona.
— (1991a): *Domar la suerte. la aleatoriedad en las decisiones individuales y sociales.* Paidós. Barcelona.
— (1991b): *Juicios salomónicos. Las limitaciones de la racionalidad como principio de decisión.* Gedisa. Barcelona.
— (1991c): *Tuercas y tornillos. una introducción a los conceptos básicos de las ciencias sociales.* Gedisa. Barcelona.
— (1992): *El cemento de la sociedad. Las paradojas del orden social.* Gedisa. Barcelona.
— (1994): *Lógica y sociedad. Contradicciones y mundos posibles.* Gedisa. Barcelona.
— (1995a): *Justicia local. De qué modo las instituciones distribuyen bienes escasos y cargas necesarias.* Gedisa. Barcelona.
— (1995b): *Psicología política.* Gedisa. Barcelona.
— (1996): *Egonomics. Análisis de la interacción entre racionalidad, emoción, preferencias y normas sociales en la economía de la acción individual.* Gedisa. Barcelona.
Epicteto (1991): *Enquiridión.* Anthropos. Barcelona.
Febrero, R. y Schwarz, P. (1997): *La esencia de Becker.* Ariel. Barcelona.
Ferrer, U. (1990): *Perspectivas de la acción humana.* PPU. Barcelona.
Friedman, J. (Ed.) (1996a): *The rational choice controversy.* Yale University Press. New Haven.
— (1996b): "Economic approaches to politics", en Friedman, J., *The rational choice controversity.* Yale University Press. New Haven.
Gasparski, W. y Pszczolowski, T. (Eds.) (1983): *Praxiological studies: Polish contributions to the science of efficient action.* D. Reidel. Dordrecht.
Gauthier, D. (1986): "La moral y la ventaja", en Raz, J. (Ed.): *Razonamiento práctico.* Fondo de Cultura Económica. México.
— (1994): *La moral por acuerdo.* Gedisa. Barcelona.
— (1998a): "El egoísta incompleto", en Gauthier, D., *Egoísmo, moralidad y sociedad liberal.* Paidós. Barcelona.
— (1998b): "Asegurar y amenazar", en Gauthier, D., *Egoísmo, moralidad y sociedad liberal.* Paidós. Barcelona.
Gauthier, R. (1973): *La morale d'Aristote.* Presses Univ. de France. París.
Green, D. y Shapiro, I. (1994): *Pathologies of rational choice: a critique of applications in political science.* Yale University Press. New Haven.
Gutiérrez, G. (1979): "La congruencia entre lo bueno y lo justo". *Revista de Filosofía,* 2ª época, 2, pp. 33-54.

— (1987): "La decisión moral: principios universales, reglas generales y casos particulares". *Revista de Filosofía*, 3ª época, 1, pp. 127-155.
— (1990a): "La estructura consecuencialista del utilitarismo". *Revista de Filosofía*, 3ª época, 3, pp. 141-174.
— (1990b): "La trama moral de la democracia en la vida cotidiana". *Anuario de Filosofía del Derecho*, 7, pp. 13-25.
Guyer, M. y Rapaport, A. (1966): "A taxonomy of 2 × 2 games". *General Systems*, 11, pp. 203-214.
Hahn, F. y Hollis, M. (Eds.) (1986): *Filosofía y teoría económica*. Fondo de Cultura Económica. México.
Hare, R. (1974): "El juego del prometer", en Foot, P. (Ed.): *Teorías sobre la ética*. Fondo de Cultura Económica. México.
— (1975): *El lenguaje de la moral*. Universidad Nacional Autónoma de México.
Harsanyi, J. (1976a): "Can the maximin principle serve as a basis for morality? A critique of John Rawls's theory", en Harsanyi, J.: *Essays on ethics, social behavior and scientific explanation*. D. Reidel. Dordrecht.
— (1976b): "Advances in understanding rational behavior", en Harsanyi, J.: *Essays on ethics, social behavior and scientific explanation*". D. Reidel. Dordrecht.
— (1976c): "Cardinal welfare, individualistic ethics and interpersonal comparisons of utility", en Harsanyi, J.: *Essays on ethics, social behavior and scientific explanation*. D. Reidel. Dordrecht.
— (1997): "Una teoría de los valores prudenciales y una teoría de la moralidad utilitarista de la regla". *Telos*, 6, pp. 59-82.
Hildebrandt, S. y Tromba, A., (1990): *Matemática y formas óptimas*. Prensa Científica. Barcelona.
Hobbes, T. (1989): *Leviatán. La materia, forma y poder de una sociedad eclesiástica y civil*. Alianza Editorial. Madrid.
Höffe, O. (1979): *Estrategias de lo humano*. Alfa. Buenos Aires.
Hofstadter, D. (1982): "Competiciones de astucia y estrategia". *Investigación y Ciencia*. Octubre, pp. 98-104.
— (1983a): "Juegos de conducta cooperativa". *Investigación y Ciencia*. Julio, pp. 108-115.
— (1983b): "Cooperación en un concurso egregio". *Investigación y Ciencia*. Agosto, pp. 102-107.
Hollis, M. y Nell, E. (1975): *Rational economic man. A philosophical critique of neo-classical economics*. Cambridge University Press.
— (1988): *The cunning of reason*. Cambridge University Press.
— y Sugden, R. (1993): "Rationality in action". *Mind*, 102, pp. 1-35.

Howard, N. (1971): *Paradoxes of rationality: theory of metagames and political behavior.* The M.I.T. Press. Cambridge Ma.

Huizinga, J. (1965): *El otoño de la Edad Media.* Revista de Occidente. Madrid.

— (1998): *Homo ludens.* Alianza Editorial. Madrid.

Hume, D. (1992): *Tratado de la naturaleza humana.* Tecnos. Madrid.

— (1993): *Investigación sobre los principios de la moral.* Alianza Editorial. Madrid.

Irwin, T. (1977): *Plato's moral theory. The early and middle dialogues.* Clarendon Press. Oxford.

Jarvie, I. (1982): "Comprensión y explicación en sociología y en antropología social", en Borger, R. y Cioffi, F. (Eds.): *La explicación en las ciencias de la conducta.* Alianza Editorial. Madrid.

Jeffrey, R. (1983): *The logic of decision.* University of Chicago Press.

Kant, I. (1989): *Metafísica de las Costumbres.* Tecnos. Madrid.

— (1994): *Crítica de la razón práctica.* Sígueme. Salamanca.

— (1996): *Fundamentación de la metafísica de las costumbres.* Ariel. Barcelona.

Kaufmann, A. (1967): *La ciencia y el hombre de acción: introducción a la praxeología.* Guadarrama. Madrid.

Kern, L. y Müller, H. (Eds.) (1992): *La justicia: ¿discurso o mercado? Los nuevos enfoques de la teoría contractualista.* Gedisa. Barcelona.

Laplace, P. (1985): *Ensayo filosófico sobre las probabilidades.* Alianza Editorial. Madrid.

Lewis, D. (1969): *Convention: A Philosophical Study.* Harvard University Press. Cambridge MA.

— (1985): "Prisoner's Dilemma is a Newcomb Problem", en Campbell, R y Sowden, L. (Eds.): *Paradoxes of rationality and cooperation: Prisoner's Dilemma and Newcomb's Problem.* University of British Columbia Press. Vancouver.

Luce, R. y Raiffa, H. (1957): *Games and decisions. Introductory and critical survey.* John Wiley. Nueva York.

Lukes, S. (1975): *El individualismo.* Península. Barcelona.

Mcclennen, E. (1985): "Prisoner's dilemma and resolute choice", en Campbell, R. y Sowden, L. (Eds.): *Paradoxes of rationality and cooperation.* University of British Columbia Press. Vancouver.

— (1988): "Constrained maximization and resolute choice", en Paul, E. (Ed.): *The new social contract. Essays on Gauthier.* Basil Blackwell. Oxford.

Manninen, J. y Tuomela, R. (Eds.) (1980): *Ensayos sobre explicación y comprensión.* Alianza Editorial. Madrid.

March, J. (1986): "Bounded rationality, ambiguity, and the engineering of choice", en Elster, J. (Ed.): *Rational Choice.* Basil Blackwell. Oxford.
Marsh, D. (1998): *Teoría y métodos de la ciencia política.* Alianza Editorial. Madrid.
Messick, D. (Ed.) (1974): *Matemáticas en las ciencias del comportamiento.* Alianza Editorial. Madrid.
Mill, J. S. (1917): *Sistema de la lógica inductiva y deductiva.* Daniel Jorro. Madrid.
— (1984): *Utilitarismo.* Alianza Editorial. Madrid.
Mises, L. v. (1968): *La acción humana. Tratado de economía.* Sopec. Madrid.
— (1986): "La ciencia de la acción humana", en Hahn, F. y Hollis, M. (Eds.): *Filosofía y teoría económica.* Fondo de Cultura Económica. México.
Moore, G. (1983): *Principia ethica.* Universidad Nacional Autónoma de México.
— (1989): *Ética.* Labor. Barcelona.
Mosterín, J. (1978): *Racionalidad y acción humana.* Alianza Editorial. Madrid.
Mueller, D. (1984): *Elección pública.* Alianza Editorial. Madrid.
Muguerza, J. (1977): *La razón sin esperanza.* Taurus. Madrid.
Murphy, J (1996): "Rational Choice Theory as social physics", en Friedman, J. (Ed.): *The rational choice controversy.* Yale University Press. New Haven.
Nasar, S. (1998): *A beautiful mind.* Simon & Schuster. Nueva York.
Nozick, R. (1985) "Newcomb's Problem and two principles of choice" en Campbell, R. y Sowden, L. (Ed.): *Paradoxes of rationality and cooperation: Prisoner's Dilemma and Newcomb's Problem.* University of British Columbia Press. Vancouver.
— (1995): *La naturaleza de la racionalidad.* Paidós. Barcelona.
Olson, M. (1992): *La lógica de la acción colectiva: bienes públicos y la teoría de grupos.* Limusa. México.
Parfit, D. (1985): *Reasons and Persons.* Oxford University Press.
— (1991): *Prudencia, moralidad y el dilema del prisionero.* Facultad de Filosofía. Universidad Complutense. Madrid.
Pascal, B. (1963): *Oeuvres complètes.* Éditions du Seuil. París.
Platón (1969): *La República.* Instituto de Estudios Políticos. Madrid.
— (1970): *Menón.* Instituto de Estudios Políticos. Madrid.
— (1983a): *Gorgias.* Diálogos II. Gredos. Madrid.
— (1983b): *Crátilo.* Diálogos II. Gredos. Madrid.

Poe, E. (1972): *Cuentos. I.* Alianza Editorial. Madrid.
Polinski, A. (1985): *Introducción al análisis económico del derecho.* Ariel. Barcelona.
Popper, K. (1973): *La miseria del historicismo.* Alianza. Madrid.
Poundstone, W. (1995): *El dilema del prisionero. John von Neumann, la teoría de juegos y la bomba.* Alianza Editorial. Madrid.
Prichard, H. (1912): "Does moral philosophy rest on a mistake?". *Mind,* 21.
— (1970): "Duty and interest", en Sellars, W. y Hospers, J. (Eds). *Readings in ethical theory.* Prentice Hall. Englewood Cliffs.
Putnam, H. (1991): *El significado y las ciencias morales.* UNAM. México.
Rapoport, A. (1960): *Fights, games and debates.* University of Michigan Press. Ann Arbor.
— (1974): "Uso y abuso de la teoría de juegos", en Messick, D. (ed.): *Matemáticas en las ciencias del comportamiento.* Alianza Editorial. Madrid.
Rawls, J. (1979): *Teoría de la justicia.* FCE. Madrid.
Raz, J. (1991): *Razón práctica y normas.* Centro de Estudios Constitucionales. Madrid.
— (Ed.) (1986): *Razonamiento práctico.* Fondo de Cultura Económica. México.
Regan, D. (1980): *Utilitarianism and cooperation.* Clarendon Press. Oxford.
Rescher, N. (1985): *Pascal's wager. A study of practical reasoning in philosophical theology.* University of Notre Dame Press. Indianapolis.
— (1993): *La racionalidad. Una indagación filosófica sobre la naturaleza y la justificación de la razón.* Tecnos. Madrid.
Resnik, M. (1998): *Elecciones. Una introducción a la teoría de la decisión.* Gedisa. Barcelona.
Ricoeur, P. (1981): *El discurso de la acción.* Cátedra. Madrid.
Roberts, R. (1992): *Serendipia. Descubrimientos accidentales en la ciencia.* Alianza Editorial. Madrid.
Robinson, J. (1973): *Economic philosophy.* Penguin Books. Harmondsworth.
Rousseau, J. J. (1971): *Discours sur l'origine et les fondements de l'inégalité parmi les hommes.* Garnier-Flammarion. París.
Ryle, G. (1958): "On forgetting the difference between right and wrong", en Melden, A.: *Essays in moral philosophy.* Washington University Press. Seattle.
Salcedo, D. (1994): *Elección social y desigualdad económica.* Anthropos. Barcelona.
Sartre, J. P. (1984): *El ser y la nada.* Alianza Editorial. Madrid.
Sayre-McCord, G. (1991): "Deception and reasons to be moral", en Vallentyne, P. (Ed.): *Contractarianism and rational choice.* Cambridge University Press.

Scheffler, S. (Ed.) (1988): *Consequentialism and its critics.* Oxford University Press.
Schelling, T. (1964): *La estrategia del conflicto.* Tecnos. Madrid.
Searle, J. (1974): "Cómo derivar 'debe' de 'es'" en Foot, P. (Ed.): *Teorías sobre la ética.* Fondo de Cultura Económica. México.
— (1994): *Actos de habla.* Cátedra. Madrid.
Sen, A. (1976): *Elección colectiva y bienestar social.* Alianza Editorial. Madrid.
— (1986a): "Tontos racionales", en Hahn, F. y Hollis, M. (Eds.): *Filosofía y teoría económica.* Fondo de Cultura Económica. México.
— (1986b): "Behaviour and the concept of preference", en Elster, J. (Ed.): *Rational choice.* Basil Blackwell. Oxford.
— (1986c): "La imposibilidad de un liberal paretiano", en Hahn, F. y Hollis, M. (Eds.): *Filosofía y teoría económica.* Fondo de Cultura Económica. México.
— (1989): *Sobre ética y economía.* Alianza Editorial. Madrid.
— (1997): *Bienestar, justicia y mercado.* Paidós. Barcelona.
Séneca (1953): *Cartas morales.* Vol. 2. Universidad Nacional Autónoma de México.
Sfez, Lucien (1984): *Crítica de la decisión.* Fondo de Cultura Económica. México.
— (1987): *La decisión.* Fondo de Cultura Económica. México.
Shackle, G. (1966): *Decisión, orden y tiempo en las actividades humanas.* Tecnos. Madrid.
Sidgwick, H. (1981): *The methods of ethics* (1907). Hackett. Indianapolis.
Simmel, G. (1977): *Sociología. Estudio sobre las formas de socialización.* Vol. 1. Revista de Occidente. Madrid.
Simon, H. (1986): "De la racionalidad sustantiva a la procesal", en Hahn, F. y Hollis, M. (Eds.): *Filosofía y teoría económica.* Fondo de Cultura Económica. México.
— (1989): *Naturaleza y límites de la razón humana.* Fondo de Cultura Económica. México.
Smart, J. y Williams, B. (1981): *Utilitarismo, pro y contra.* Tecnos. Madrid.
Smith, H. (1991): "Deriving morality from rationality", en Vallentyne, P. (Ed.): *Contractarianism and rational choice.* Cambridge University Press.
Sobel, J. (1985): "Not every Prisoner's Dilemma is a Newcomb Problem", en Campbell, R. y Sowden, L. (Eds.): *Paradoxes of rationality and cooperation: Prisoner's Dilemma and Newcomb's Problem.* University of British Columbia Press. Vancouver.
Spinoza, B. (1987): *Ética demostrada según el orden geométrico.* Alianza Editorial. Madrid.

Taylor, C. (1982): "Responsibility for self", en Watson, G. (Ed.) (1982): *Free will.* Oxford University Press.
Taylor, M. (1987): *The possibility of cooperation.* Cambridge University Press.
Torres, J. (1979): *Análisis económico del derecho.* Tecnos. Madrid.
Toulmin, S. (1974): "Razones y causas", en Borger, R. y Cioffi, F. (Eds.): *La explicación en las ciencias de la conducta.* Alianza Editorial. Madrid.
Tullock, G. (1979): *Los motivos del voto.* Espasa Calpe. Madrid.
Ullman-Margalit, E. (1977): *The emergence of norms.* Clarendon Press. Oxford.
Viner, J. (1973): "Bentham and Mill: the Utilitarian Background", en Phelps, E. (Ed.): *Economic Justice.* Penguin Books. Harmondsworth.
Watkins, J. (1982): "Racionalidad imperfecta", en Borger, R. y Cioffi, F. (eds.): *La explicación en las ciencias de la conducta.* Alianza Editorial. Madrid.
Watson, G. (Ed.) (1982): *Free will.* Oxford University Press.
Weber, M. (1964): *Economía y sociedad. Esbozo de sociología comprensiva.* Fondo de Cultura Económica. México.
White, A. (Ed.) (1976): *La filosofía de la acción.* Fondo de Cultura Económica. México.
White, D. (1979): *Teoría de la decisión.* Alianza Editorial. Madrid.
Williams, B. (1986): "Decidir creer", en Williams, B. (Ed.): *Problemas del yo.* Universidad Nacional Autónoma de México.
Winch, P. (1972): *Filosofía de la ciencia social.* Amorrortu. Buenos Aires.
— (1980): "Causalidad y acción", en Manninen, J. y Tuomela, R. (Eds.): *Ensayos sobre explicación y comprensión.* Alianza Editorial. Madrid.
Wittgenstein, L. (1973): *Tractatus logico-philosophicus.* Alianza Editorial. Madrid.
— (1988): *Investigaciones filosóficas.* UNAM. México.
Wright, G. v. (1967): *Lógica de la preferencia.* Eudeba. Buenos Aires.
— (1970): *Norma y acción. Una investigación lógica.* Tecnos. Madrid.
— (1979): *Explicación y comprensión.* Alianza Editorial. Madrid.
Zintl, R. (1995): *Comportamiento político y elección racional.* Gedisa. Barcelona.