# ¿QUÉ MODELO DE RACIONALIDAD?

GILBERTO GUTIÉRREZ
*Universidad Complutense de Madrid*

ABSTRACT

The Prisoner's Dilemma is the experimentum crucis for the model of rationality employed by Rational Choice Theory (RCT) in its analysis of cooperative behaviour. It can be argued that certain features of the model unduly restrict the range of possible solutions. Any single model must be formally and internally consistent but in order to achieve its explanatory and predictive goals more is required. It cannot be self-contained nor empty of contents. If the epistemic and heuristic value of a model depends on its isomorphism with reality it should embody a plausible psychological hypothesis about its material interpretation. Much of the appeal of the RCT model of rationality is due to the intuitive plausibility of the primitive notions of utility and preference. It is precisely on account of its focusing on typically consequentialist, forward-looking reasons for action that the model runs into a deadlock. It is suggested that the exolusion of other, backward-looking kind of reasons is not sufficiently warranted and misrepresents important features of practical rationality.

## 1

Con su lección inaugural de 1 954 –a sólo diez años de la publicación de la obra de Von Neumann y Morgenstern– Richard Braithwaite es pionero en adoptar «la nueva disciplina matemáti-

Τέλος Vol. VII, n.° 1. Junio 1998, págs. 51-60. © S.I.E.U.

ca llamada Teoría de Juegos[1]» como herramienta para el filósofo moral. Es interesante observar qué características de la teoría determinan esta temprana recepción. Braithwaite no la concibe como una herramienta meramente analítica, sólo apta para describir ciertas estructuras de interacción y predecir conductas, sino como un instrumento dotado de eficacia práctica que permite «al moralista filosófico 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»[2].

El espíritu de Braithwaite alienta aún en la propuesta de Julia Barragán: como todo modelo conceptual, los que presenta la Teoría de Juegos –por ejemplo, el Dilema del Prisionero– son susceptibles «de dos interpretaciones (y usos) diferentes: (...) como descriptivos de la estructura de determinadas relaciones, cuyo análisis permitiría elaborar categorías explicativas y predecir (su) futuro comportamiento; (...) como definidores de conductas sociales deseables (...) harían posible producir normas de comportamiento para regir las relaciones sociales»[3]. Esto es, nos dicen cómo se comportarán dos jugadores racionales en el marco de las restricciones establecidas por el modelo o, alternativamente, cómo deberían comportarse para aprovechar las ventajas de la conducta cooperativa, al hacer evidentes «las consecuencias que conlleva la aplicación de ciertos principios".

La Teoría pretende responder al problema específico que plantea la *pluralidad de objetivos* de los posibles cooperadores. Braithwaite reconoce que «la mayoría de los filósofos morales han ignorado esta cuestión, al suponer que, a menos que los individuos puedan ponerse de acuerdo como mínimo en los fines próximos que todos desean perseguir, no es posible ninguna base racional para la acción común". Si existe al menos este acuerdo limitado, se acepta un ulterior supuesto, tomado de «los economistas del bienestar, herederos de la tradición utilitarista» según el cual «los fines deseados por diferentes individuos –sus 'utilidades'– pueden compararse entre sí en términos de unidades comunes (...) que pueden transferirse de una persona a otra"; si bien este supuesto

[1] BRAITHWAITE, R.B.: *Theory of Games as a Tool for the Moral Philosopher.* Cambridge University Press. 1955; p. 5.

[2] *O.c.*, p. 4.

[3] «Las reglas de la cooperación«, en James GRIFFIN y o.: *Ética y política en la decisión púbilca.* Caracas. Ediciones Angria. 1993, p. 50.

se basa en la dudosa posibilidad de efectuar comparaciones interpersonales de utilidad[4].

En principio, pues, podría parecer que el problema no se plantea para el agente que actúa en situaciones paramétricas: sería razonable para Robinson, mientras estuviese a solas en su isla, proponerse como objetivo «maximizar su propia satisfacción"; sólo a partir de la llegada de Viernes, que hace posible la cooperación, ha de tener en cuenta los objetivos de éste[5]. Pero incluso en este caso habría que especificar lo que se entiende por «su propia satisfacción", porque la expresión carece de un referente unívoco. Un caso tan simple como la elección individual entre ahorro y consumo –paramétrica *ceteris paribus*– pone de manifiesto el conflicto latente entre los objetivos –o, si se prefiere, las preferencias– de los diversos segmentos temporales que integran la identidad del agente y que permite hablar incluso de dilemas del prisionero *intra*personales e *inter*temporales[6].

Dejando de lado esta complicación, Braithwaite considera que sus recomendaciones para distribuir equitativamente los procedimientos de colaboración «serán amorales en el sentido de no basarse en ningún principio moral de primer orden, pero constituirán lo que podría llamarse principios morales de segundo orden que ofrecen criterios de *buen sentido, prudencia* y *equidad* (...) en cierta forma análogos al 'principio suplementario' para la 'justa distribución de la felicidad' que Henry Sidgwick, (su) 'bisabuelo' en la Cátedra Knightbridge, creyó necesario para endulzar la leche pura del evangelio utilitarista»[7].

## 2

Sin renunciar a ser un evangelio, el utilitarismo siempre ha aspirado a constituirse como una *teoría,* al menos en el sentido mínimo de presentarse como un conjunto sistemático de proposiciones que se derivan lógicamente de principios básicos. En una teoría *práctica* estas proposiciones proporcionan, además, razones

4 *Ibid.*
5 *Ibid.*
6 Por ejemplo, a Derek PARFIT: *Prudencia, Moralidad y el Dilema del Prisonero,* Facultad de Filosofía de la Universidad Complutense, Madrid, 1991, § IV; *Reasons and Persons.* Oxford University Press, 1984 § 34.
7 *O.c.*

para actuar, implican recomendaciones que responden a la pregunta por lo que conviene o resulta racional hacer. Una teoría específicamente *ética* define razones admisibles para actuar al imponerles la restricción de ajustarse al criterio de lo que es bueno o justo. La acción es moralmente correcta si es sistemáticamente implicada por el principio que define lo que es bueno o justo, y la obligación objetiva del agente es ejecutarla.

El examen en términos meramente estadísticos de la literatura reciente sobre el utilitarismo revela que una parte muy considerable de las discusiones se centran en los aspectos normativo-prácticos *–evangélicos,* si se quiere– de la teoría, en los efectos de su extensión o aplicación a cuestiones suscitadas, por ejemplo, en el ámbito de la política, el derecho, la ecología o la bioética. Se presta comparativamente menor atención a los presupuestos implícitos o explícitos de su más básica condición de teoría «simplemente» práctica. Más en concreto, al modelo de racionalidad práctica que supone o propone. Y sin embargo, en interpretaciones muy autorizadas de la teoría utilitarista, como la de John Harsanyi, se ha venido proponiendo desde hace largo tiempo «considerar también a la ética como una rama de la teoría general de la conducta racional, ya que la teoría ética puede fundarse en axiomas que representan especializaciones de algunos de los axiomas utilizados en la teoría de la decisión»[8].

Es un lugar común entre los filósofos morales la observación de Alfred Ayer según la cual «el sistema ordinario de ética, tal como es elaborado en las obras de los filósofos éticos está muy lejos de constituir un todo homogéneo; no sólo es susceptible de contener partes de metafísica y análisis de conceptos no-éticos, sino que sus propios contenidos éticos son ellos mismos de tipos muy diferentes[9]. No es ésta una característica peculiar de las teorías éticas: de forma implícita o explícita toda teoría acepta como presupuestos indemostrables proposiciones que sólo son demostrables en una teoría de nivel más básico, de la cual depende a través de una estructura de gran complejidad lógica. Distintos filósofos morales han destacado esta estratificación estructural de las teorías éticas. Basten un par de ejemplos: Abraham Edel aplica en su análisis metodológico de las teorías éticas el concepto de «niveles ins-

[8] HARSANYI, J.C.: Advances in understanding rational behavior. En J.C. HARSANYI (Ed.): *Essays on ethics, social behavior and scientific explanation.* Dordrecht. D. Reidel. 1976, pp. 89-118.

[9] *Languaje, truth and logic.* Harmondsworth: Penguin Books. 1971, pp. 136-7.

trumental-funcionales (...como los de) conducta, pauta moral y teoría ética»[10], mientras que Georg von Wright considera que las teorías éticas emplean tres tipos de conceptos lógicamente distintos: deontológico-normativos, axiológico-valorativos y antropológicos –siendo estos últimos los propios de la filosofía de la acción y de la mente[11].

## 3

Para describir, comprender y explicar por medio de proposiciones indicativas la acción considerada como algo dado, la filosofía de la mente se sitúa en la perspectiva externa propia del espectador. Pero esa misma acción es considerada por el agente desde una perspectiva interna en el proceso de deliberar, elegir, decidir y actuar; proceso que se traduce en proposiciones valorativas o normativas que ofrecen razones para actuar. Esta perspectiva interna es la que adoptan las teorías normativas o prácticas, y en especial la ética.

La cuestión que se plantea es si una teoría normativa –en este caso el utilitarismo– puede cumplir esta función sin presuponer a su vez un modelo de agente racional, una teoría acerca de la naturaleza «real» de los agentes que deliberan, deciden y actúan, es decir, una teoría de la racionalidad práctica. Teniendo en cuenta que tal teoría no puede a su vez ser normativa sino que ha de presentarse como esencialmente explicativa o, en definitiva, científica. Y lo que caracteriza a una teoría de este tipo es que puede ser verdadera o falsa, plausible o implausible, en la medida exacta en que dé cuenta de los hechos que se propone explicar. En recientes estudios sobre la evolución y el estado actual de la teoría de la elección racional, Martin Hollis y Robert Sugden atribuyen ciertos dilemas y paradojas que aquejan a la Teoría de la Elección Racional precisamente a su aceptación incondicionada de supuestos procedentes de una determinada filosofía de la mente[12].

10 *El método en la teoría ética.* Madrid. Tecnos. 1968, p. 192.

11 *The logic of preference.* Edinburgh University Press. 1970, § 1.

12 «Utility theory is not and cannot be innocent of all philosophy of mind»: SUGDFN, Robert; HOLLIS, Martin: «Rationality in action». *Mind,* 102, 1993, p. 1, 32; SUODEN, Robert: «Rational choice. a survey of contributions from economics and philosophy». *The Economic Journal,* 101, 1991, pp. 751-785.

En tanto que constructo artificial todo modelo posee una dimensión *sintáctica* que se agota en su coherencia formal e interna. Pero desde la perspectiva de su función científica –explicativa, predictiva e incluso hermenéutica– el modelo no puede ser *autocontenido* ni estar vacío de contenido. Ni sus conceptos *primitivos* ni las *premisas* que los contienen pueden ser simplemente *estipulados.* Si el valor epistémico de un modelo depende de su adecuación –su isomorfismo– con la realidad, ha de incorporar necesariamente una hipótesis que permita de manera implícita o explícita su interpretación sustantiva y *material*

Es un lugar común en las ciencias sociales y en disciplinas filosóficas como la filosofía de la mente la necesidad y la dificultad –ajena a las ciencias naturales– de tener en cuenta los conceptos que los propios agentes emplean en sus deliberaciones prácticas para entender la lógica de la situación desde la perspectiva de estos. Aunque la ciencia adopta respecto de la acción la perspectiva del *espectador,* para hacer inteligible la conducta del *agente* necesita al menos *mencionar* unos conceptos que el propio *agente,* en cambio, se ve ineludiblemente forzado a *usar* en su deliberación práctica.

La simple tipología de von Wright antes aludida permite clasificar estos conceptos en dos categorías bien diferenciadas:

– *valorativos,* que permiten discriminar entre las alternativas en función de *preferencias* y *utilidades.*

– *normativos,* que permiten discriminar entre las alternativas en función de su coherencia con *normas* o *principios* que enuncian lo que debe hacerse.

Ambos *tipos* de conceptos corresponden a dos *tipos* muy diferentes de situaciones de elección y obedecen a una lógica asimismo diferente en la deliberación práctica. Los *valorativos* caracterizan la deliberación en aquellas situaciones en las que *de hecho* los agentes se consideran libres de decidir en función de sus puras preferencias personales –incluyendo desde gustos a preferencias meditadas– sin más restricciones que las que le impongan las condiciones materiales, los costes alternativos de sus decisiones o las limitaciones de capacidad computacional, de organización y utilización de la memoria, etc. Los conceptos *normativos* se aplican, asimismo típicamente, en aquellas otras situaciones en las que *de hecho* el agente reconoce la existencia de restricciones de naturaleza distinta a las anteriores pero que coinciden con ellas en impedirle dar libre curso a sus *preferencias* –las *obligaciones.*

## 4

Cuándo, cómo y por qué los agentes definen o perciben ciertas situaciones de elección como pertenecientes a uno u otro tipo es una cuestión de trascendental importancia filosófica, pero que podemos dejar de lado ahora. Más trascendente es la cuestión del *status* propio de las obligaciones. No es lo mismo considerarlas como *pronósticos* –simples *rules of thumb* que ayudan al cálculo de probabilidades y utilidades esperables de las acciones que se ajustan a la pauta restrictiva– o, por el contrario, como auténticas *normas* cuya obligatoriedad no deriva directamente de su utilidad. En el primer caso tendrían la condición de razones *prudenciales* –término que traduce literalmente la expresión inglesa *forward-looking:* el *prudens* es el *prae-videns,* que mira hacia adelante y ve de antemano. En el segundo serían *backward-looking,* bien en el sentido estrictamente temporal, por ejemplo, de atenimiento a un pacto o promesa anterior, o incluso en el sentido de ser independientes de consideraciones prudenciales, por lo que bien podría llamárselas *inward-looking.* Esta distinción es la que recoge el contraste tradicional entre *interés* y *deber* que David Gauthier formula acertadamente cuando cuestiona la identificación que hace Hume de ambos: »Si el deber no fuese más que el interés, la moral sería superflua. ¿Por qué apelar a lo correcto o incorrecto, al bien o el mal, a la obligación o al deber, si fuese posible apelar en cambio al deseo o la aversión, al beneficio o el coste, al interés o el provecho? Apelar a la moral tiene sentido precisamente a raíz de la insuficiencia de estas consideraciones como guía de lo que debemos hacer»[13]

Incluso a solas en su isla Robinson tiene que rendirse a la necesidad racional de imponerse *obligaciones interesadas,* es decir, restricciones a sus posibles elecciones encaminadas a promover su propio beneficio e interés a medio y largo plazo. La necesidad de restringirse se debe a las múltiples causas ya aludidas: la escasez intrínseca de todo recurso finito, el conflicto interior entre su «razón» y sus «pasiones», su voluntad débil, su racionalidad imperfecta que le impide ordenar adecuadamente sus propias preferencias, la contingencia que afecta de incertidumbre sus previsiones de futuro, etc.

La aparición de Viernes obliga a Robinson a modificar su hipótesis científica acerca del entorno de sus decisiones para integrar

[13] *Morals by agreement.* Oxford. Clarendon Press. 1986, p. 1.

el nuevo dato: sería erróneo atribuir a lo que es en realidad otro sujeto –alter ego– las mismas características que al resto de las cosas y objetos de la naturaleza. Y este error tendría para Robinson la grave consecuencia técnica de impedirle deliberar y actuar adecuadamente en su propio beneficio, ya que le haría incapaz de prever el «comportamiento» de un elemento del entorno que posee la singular capacidad de tomarle a él en cuenta y formarse expectativas respecto de su conducta. En principio Viernes aparece como una fuente de restricciones simplemente interesadas para Robinson: la conducta de aquél 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 otra es la interpretación que cabe dar a la afirmación de Julia Barragán: «para que la convivencia en sociedad sea posible, se hace *necesario* que el individuo acepte ciertas formas de restricción de sus intereses individuales»[14].

## 5

Aunque más arriba se han definido las *preferencias* y las *obligaciones* como dos *tipos* conceptuales cuasi-excluyentes lo cierto es que guardan entre si unas relaciones mucho más complicadas de lo que parece deducirse de su distinción abstracta. Una situación normativamente definida puede dejar la decisión al libre arbitrio del agente: cuando a éste se le reconoce el derecho o se le permite hacer u omitir algo, 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. Más aún: el agente podría tener metapreferencias, esto es, preferencias de segundo orden sobre sus preferencias de primer orden, que podrían ejercer sobre éstas una influencia normativa: alguien puede estar insatisfecho consigo mismo y preferir ser de otra manera.

Pero existe una tercera categoría de conceptos cuya función queda perfectamente confinada al ámbito del discurso filosófico-científico *sobre* la acción y, en consecuencia, a la perspectiva propia del espectador. Son los conceptos *psicológico-filosóficos,* que

[14] «Las reglas de la cooperación», p. 41.

describen la naturaleza y las propiedades *reales* de los agentes y las acciones –motivos, razones, intenciones, etc. Son precisamente estos conceptos los que permiten la interpretación material de los modelos formales de la elección.

La filosofía moral ha venido aplicando modelos elaborados por la Teoría de Juegos –en particular el Dilema del Prisionero– a las elecciones en situaciones de conflicto y cooperación con propósitos tanto analíticos como normativos. En términos de la coherencia formal 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 Juegos procede precisamente de la aparente plausibilidad empírica de que goza el concepto primitivo de *interés* cuyo contenido, en palabras de Barragán, «no difiere fundamentalmente de lo que en el plano intuitivo entendemos por interés y que equivale al de *preferencia*»[15]

Los citados estudios de Hollis y Sugden apuntan como posible causa de determinadas paradojas y dilemas de la racionalidad estratégica el hecho de presuponer una muy concreta interpretación de la naturaleza de las razones que son capaces de motivar a la acción. No es posible en esta comunicación extenderse en los detalles de su argumentación. Baste mencionar que 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 (también de explicar) la cooperación en los términos estrictos de la Teoría de Juegos. 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.

[15] *Ib.*, p. 44.

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. Es dudoso que la propia teoría de la racionalidad prudencial podría hacer frente a los problemas de simple cooperación –no digamos ya equitativa– por medio de las estrategias estrictamente interesadas que analiza la Teoría de Juegos. Bien es verdad que se hace necesario transformar los presupuestos de la racionalidad paramétrica en la medida necesaria para hacer frente a las situaciones estratégicas. Pero estas transformaciones se reducen a cambios de escala en un gradiente de complejidad creciente pero sin solución de continuidad *dentro* de un universo homogéneo de racionalidad; mientras que lo que llamamos moralidad parece producir una mutación *de* este universo. En definitiva, la necesidad de *dar cuenta* de la realidad de la cooperación fuerza, por razones meramente filosóficas y no por motivos evangélicos, el propio modelo de racionalidad.