AGORA — Papeles de Filosofía — (2003), 22/2: 223-241 ISSN 0211-6642

PRISIONEROS DEL DILEMA

Blanca Rodríguez López
(UCM)

Resumen

Las situaciones de interacción ejemplificadas mediante el Dilema del prisionero plantean, ante todo, un problema práctico. Parfit clasifica las distintas soluciones a este problema en políticas y psicológicas. Dentro de estas últimas, las soluciones morales son las más importantes. En este artículo analizaremos las distintas soluciones morales, así como los distintos cambios morales que pueden dar lugar a las mismas, con especial atención a las ventajas e inconvenientes de cada una.
Palabras clave: dilema del prisionero, Parfit, racionalidad estratégica, teoría de juegos, soluciones morales.

Abstract

Interaction situations of the kind illustrated by the Prisoner's Dilemma pose first and foremost a practical problem. Parfit classifies the different solutions to this problem into political and psychological ones. Among the psychological solutions, moral solutions are particularly important. In this paper we will analyse the moral solutions as well as the changes in the moral attitude that can produce them and will pay special attention to the advantages and disadvantages of each of these attitudes.
Keywords: Prisoner's Dilemma, Parfit, strategic rationality, Game Theory, moral solutions.

"Supongamos que cada uno esta dispuesto a hacer lo que será mejor para sí mismo, o para su familia, o para aquellos a los que ama. Hay entonces un problema práctico: a menos que algo cambie, el resultado real será peor para todos. Este problema es una de las principales razones por la que necesitamos algo más que una economía tipo laissez-faire, por la que necesitamos tanto la política como la moralidad"[1].

Este texto de Parfit resume de modo preciso el problema fundamental al que se enfrenta una de las concepciones más habituales, y por otra parte

[1] Parfit (86), p.62

bien defendida, de un agente racional, según la cual este utilizará siempre la estrategia que maximize su utilidad individual: cuando varios agentes interactúan en un tipo de situaciones comúnmente conocidas como "dilema del prisionero", la situación siempre se resolverá de un modo no cooperativo. Y este resultado es subóptimo. Todos ganaríamos más si se respetara el acuerdo y el juego se desarrollara cooperativamente. El objeto de este artículo es el análisis de las distintas soluciones que pueden ofrecerse en este tipo de situaciones.

El dilema del prisionero

Según la narración histórica que W. Poundstone hace en su obra El dilema del Prisionero, fue en 1950 cuando dos investigadores propusieron "un "juego" simple y desconcertante que puso a prueba parte de la fundamentación teórica de la teoría de juegos (...) se trata de un rompecabezas intelectual que aún hoy nos deja perplejos"[2]. La historia que ilustra el juego es más o menos la siguiente. Dos miembros de una banda criminal son detenidos y encarcelados en celdas separadas. La policía sabe que han cometido un delito grave, pongamos un atraco, pero carece de pruebas para demostrarlo ante un tribunal. Sin embargo, tienen pruebas más que suficientes para acusarles de un delito de menor importancia, digamos un timo de poca monta. En está situación, al jefe de policía se le ocurre ofrecerles un trato que Poundstone califica de "pacto digno de Fausto"[3]: si uno de ellos testifica contra su compañero, de modo que pueda acusarse a este del delito mayor, saldrá en libertad, mientras que su compañero deberá enfrentarse a una pena de 5 años de prisión. La trampa está en que el trato es propuesto a ambos delincuentes, y si ambos lo aceptan y testifican el uno contra el otro, ambos serán condenados a 3 años de prisión. Los dos conocen esta circunstancia, así como que la policía ofrece el trato precisamente por carecer de pruebas, de modo que si ambos permanecen callados y ninguno confiesa, sólo sufrirán una condena de 1 año cada uno.

Conocer todas las circunstancias relevantes no les ayuda a resolver su problema. Más bien este conocimiento es el que les precipita en el dilema. Cada uno de ellos medita por separado en la soledad de su celda: si testifico, es posible que quede en libertad. Claro que si mi compañero también

[2] Poundstone (92), p.21.
[3] Poundstone (92), p.175

testifica me caerán 2 años. Si no testifico, corro el riesgo de que él sí lo haga, en cuyo caso pasaré 3 años interminables en la cárcel. Y, en todo caso, de un año no me libra nadie. Supongamos que mi compañero no testifica. Entonces, lo mejor para mí es testificar y salir de aquí. Pero espera. Supongamos que él sí testifica. Demonios. Pero entonces lo mejor para mí también es testificar. Tendría que ir a la cárcel de todas formas, pero me ahorraría dos años. La cosa está clara: haga lo que haga mi compañero, lo mejor para mí es testificar.

La trampa está servida, y funciona a las mil maravillas. Y aparece el problema: como cada uno está dispuesto a hacer lo que es mejor para sí mismo, es decir, testificar, el resultado será peor para los dos que si no lo hicieran: pasarán 3 años en la cárcel cada uno en vez de uno sólo.

Este tipo de dilemas aparecen en situaciones que Elster califica de *estratégicas*[4]: aquellas en las que el resultado no depende únicamente de la acción de un agente sino de varios. Estas situaciones son estudiadas por la Teoría de Juegos, desarrollada en el segundo cuarto del siglo XX por von Neumann y el resultado de estos estudios fue publicado en 1944 en lo que es considerado el gran manual de referencia de esta teoría[5]. En tanto que teoría normativa, la Teoría de Juegos indica qué debe hacer un agente que se encuentra en tales situaciones. Se trata de una teoría enormemente compleja, pero para nuestro objetivo presente basta con señalar que, en general, lo que debe hacer un agente en estas situaciones es tener en cuenta precisamente su carácter estratégico, es decir, tener en cuenta a la hora de decidir su acción cuales son la acciones que puede esperar del resto del los agentes involucrados. Se trata por tanto de situaciones en las que el agente deberá desarrollar el tipo de acción que Weber llamaba acción social[6]. Esto es precisamente lo que hacen los agentes del Dilema del Prisionero, y es precisamente lo que hace que el dilema sea un dilema: cada uno tiene en cuenta cuales son las posibles acciones del otro pero resulta que, haga lo que haga el otro, lo mejor para él es testificar. Sin embargo, que los dos testifiquen conducirá a un resultado subóptimo, es decir, hay otro resultado de otro par de acciones que es mejor para los dos: si ninguno testifica, pasarán en la cárcel dos años menos cada uno que si testifican.

El problema no surge de los resultados concretos planteados (el número de años de cárcel a los que serán condenados en cada estado de hechos

[4] Elster (79), p.18
[5] von Neumann y Morgenstern (44)
[6] Weber (21)

alternativo) sino de las características formales y por tanto generales de la situación. El tipo de situaciones estratégica ejemplificadas en el Dilema del Prisionero dependen de que los distintos resultados posibles se atengan a una determinada clasificación. En palabras de Poundstone[7] "existe un resultado de "recompensa" o "premio"...en el caso de que haya cooperación entre ambos; este resultado es más conveniente que el resultado de "castigo" otorgado a ambos si se trata de la situación de no colaboración. Sin embargo, ambos codician el resultado de la "tentación" que tiene lugar en el caso de que haya una única deserción; este resultado de más ganancias que el resultado "recompensa". Ambos jugadores temen ser el que no deserte, y que por tanto le corresponda el resultado de hacer el primo o perder". En este contexto por "cooperar" se entiende realizar la acción que, en caso de ser realizada por ambos, conduciría a un resultado óptimo de Pareto, es decir, aquel que no puede ser mejor para ninguno salvo a costa de ser peor para el otro.

A pesar de las apariencias, el dilema tampoco depende de que cada prisionero este aislado en una celda y tenga que tomar su decisión sin consultar con su compañero y de manera independiente, es decir, sin poder ponerse de acuerdo con el respecto a la estrategia a seguir, sin poder articular un plan conjunto. Podría pensarse que, si se les dejara comunicarse, acordarían una estrategia conjunta de cooperación. Sin duda llegarían a un acuerdo semejante, pero el acuerdo no llegaría a cumplirse: cada uno volvería a plantearse si cumplir o no con lo pactado y llegaría a una conclusión paralela a la primitiva: si el otro cumple el acuerdo, lo mejor para mi es no cumplirlo y, suponiendo que el otro no lo cumpla, con razón de más. Esto sucede porque cada uno sabe que el otro esta tan fuertemente tentado como lo está él para intentar conseguir el resultado de "tentación" e igualmente preocupado por la posibilidad de que todo termine con el resultado de "hacer el primo". En tales situaciones, de nada valen los pactos sin espada.

Utilizando los términos de Parfit, para que aparezca una situación con la estructura del Dilema del prisionero basta con que se cumplan dos condiciones:

Condición *positiva*, (1) cada uno podría, con algún sacrificio por su parte, proporcionar a otros un beneficio mayor y (2) si cada uno proporciona ese beneficio a los otros, cada uno resultará más beneficiado

[7] Poundstone (92), p.179.

Condición negativa, no hay otros efectos indirectos de las acciones de cada uno que supriman esos efectos directos.

Cómo salir del Dilema

Las posibles soluciones, así como su caracterización general, aparecen en un cuadro elaborado por Parfit[8] y que reproduciremos a continuación, ya que, debido a su claridad y exhaustividad, será de gran utilidad en toda la discusión siguiente. Conviene por ello tenerlo siempre presente.

Llamemos "auto-benéfica" (B) a la estrategia no cooperativa y altruista (A) a la cooperativa. El motivo de utilizar estos nombres es claro. En un juego en el que los acuerdos no son imponibles, un agente racional que busca maximizar su utilidad empleará la estrategia no cooperativa. Por este motivo nos referiremos a estos juegos como juegos no-cooperativos. Utilizar la estrategia cooperativa en esas condiciones sería tanto como favorecer al contrario a costa de uno mismo, por lo que podríamos calificar de altruista esta elección. Si todos utilizan B, el resultado será subóptimo. Si, por el contrario, todos utilizan A, el problema quedará resuelto. El dilema se plantea porque, a menos que algo cambie, los agentes tienen todos los motivos para hacer B y ninguna para hacer A. Parece entonces que todas las soluciones posibles han de tener un punto común: proporcionar a los agentes motivos para hacer A. El cuadro que se presenta a continuación refleja los distintos motivos por los que la solución puede ser alcanzada:

Cada uno puede hacer A

a) porque B resulte imposible (1)

b) porque adquiera una disposición a realizar A. Un agente puede adquirir esta disposición:

b1) porque A pase a ser mejor para él:

- por un cambio en su situación (2)
- por un cambio en él (3)

b2) tanto si A es mejor para el como si no. Puede suceder:

- que, debido al cambio operado en él, A ya no sea peor para el (4)
- que, a pesar del cambio, A siga siendo peor para él (5)

En este cuadro aparecen numeradas del 1 al 5 los distintos modos en que puede solucionarse el dilema. Dicho de otro modo, estos son los distintos

[8] Parfit (86), p.63 y (79), p.13

modos que, haciendo que los acuerdos adquieran estabilidad, transforman un juego no cooperativo en uno cooperativo.

Estas distintas soluciones pueden agruparse en dos categorías[9]. Por un lado están las **soluciones políticas**, que resuelven el dilema a través de la introducción de un cambio en la situación en la que se encuentra el individuo. A este grupo pertenecen las soluciones 1 y 2. Por otro lado esta el grupo de las **soluciones psicológicas**, al que pertenecen el resto de las alternativas, y que toman este nombre debido a que dan estabilidad a los acuerdos debido a un cambio producido en el interior del individuo[10].

Las soluciones políticas presentan sobre las psicológicas la ventaja de ser más fáciles de introducir. Pensemos en el caso del dilema n-personal presentado por la contribución al bien público mediante los impuestos: cada uno se beneficia más si no paga impuestos, pero si nadie los paga todos resultan perjudicados y, bajo cualquier supuesto acerca de la conducta de los demás, lo mejor para cada uno es no pagarlos. La solución 1 supondría disponer de medios que hicieran imposible la evasión de impuestos, por ejemplo mediante algún tipo de control de ingresos por parte del estado. La solución 2 pasaría por el establecimiento de un sistema de castigos y recompensas que hiciera beneficiosa individualmente la contribución. Estos sistemas son fáciles de introducir en el sentido de que pueden introducirse de una vez por todas mediante la promulgación de una ley. Esto traería en consecuencia la ventaja adicional de que sería fácil que todos conocieran de inmediato la existencia de estas medidas, lo cual haría que cada uno a) conociera que, en la nueva situación, ya no le es posible o no le es beneficiosa la no cooperación y b) supiera que todos los demás están igualmente

[9] Conviene recordar que estas soluciones lo son de cara al problema práctico mencionado por Parfit en el texto citado al principio. El problema teórico tiene peor solución si es que tiene alguna. Por eso es un dilema.

[10] Esta división en dos grupos de las razones por las que alguien puede adoptar una estrategia altruista no es nueva. Por el contrario, responde a lo que clásicamente se consideraban sanciones externas e internas. Por ejemplo, Bentham , en un intento de explicar porque alguien puede actuar de un modo que contribuya a la felicidad ajena, ofrece una lista de cinco tipos de sanciones, de entre las cuales las 4 primeras (física, político, popular y religiosa) serían sanciones externas, mientras que la última, la sanción social o simpática, sería externa (Bentham (89) Cap.3) De modo similar, Sidgwick habla de sanciones externas e internas para tratar de explicar la coincidencia deber/interés (Sidgwick (81)p.164). La misma división puede encontrarse en Mill (74), capítulo III. El esquema clásico presentado por estos autores coincide con el de Parfit no sólo en esta división en dos grupos, sino también en algunas de las alternativas ofrecidas. Otros, como Hobbes, sólo admiten las soluciones políticas.

informados del cambio de la situación y por tanto, también actuarán de forma cooperativa. Supondría una espada capaz de hacer cumplir el pacto.

Sin embargo, las soluciones políticas cuentan también con no pocos inconvenientes. En el caso de la solución 1, el principal problema es que es, en muchos casos, sencillamente inaplicable. Piénsese por ejemplo en el caso de los soldados presentado por Parfit. Cada uno se enfrenta con la elección de desertar o mantenerse en su puesto. Si los demás desertan, lo mejor que puede hacer es desertar también. Si los demás se mantienen en su puesto, entonces también es mejor desertar. Puede haber distintos modos de convertir esta situación en cooperativa, pero parece que la solución 1 es difícilmente aplicable. Siempre se puede atar a los soldados a sus puestos, o romperles un tobillo. Entonces no podrán huir. Pero tampoco podrán avanzar.

La solución 2 tampoco es aplicable en todos los casos. Pensemos en la práctica de ayudar al prójimo. Si nadie ayuda a los demás, lo mejor será que yo tampoco lo haga y si los demás lo hacen también es mejor para mí si no lo hago. Pero de esta forma, todos salimos perdiendo. A veces es posible utilizar 2 como solución, pero no siempre. Porque para eso debería disponerse de un medio para saber en todos los casos cuando alguien ha ayudado a otro y cuando no. Naturalmente, no bastaría con la palabra de ninguno de los implicados. Y, a menos que dispongamos de una especie de vigilante personal para cada uno, esto parece en muchos casos difícil de averiguar.

La única desventaja de las soluciones políticas no es que no sean universalmente aplicables. Incluso en los casos en los que pueden emplearse (que, dicho sea de paso, son bastantes) no constituyen una buena solución. El motivo es que introducir estas soluciones resulta tremendamente costoso. Si se emplearan en todos los casos, el costo de mantenimiento del sistema coercitivo y del sistema de recompensas sería tan grande que probablemente saldríamos perdiendo aun más que con la no cooperación. En el mejor de los casos, la situación alcanzada mediante estos mecanismos no será óptima. Si podemos conseguir algún otro medio que nos ahorre el costo supuesto por estas soluciones, todos saldremos ganando.

Además, estas soluciones tienen otras desventajas notables. Por ejemplo, tal y como apunta Rawls[11], la creación de este sistema cuenta con el inconveniente de ser un peligro para la libertad individual, existiendo siem-

[11] Rawls (77) p.240

pre el riesgo de una interferencia indebida en los asuntos privados. Esto sin contar con que la aplicación universal de estas medidas sin duda sería contemplada como un "no dejar vivir" por parte del estado, y crearía una sensación tan angustiosa de falta del libertad que sus posibles beneficios se verían considerablemente mermados.

Por todo ello, parece que estas soluciones deben ser consideradas como un último recurso que debe aplicarse en ausencia de un método mejor. De todas formas, y en tanto que su existencia da estabilidad a la cooperación, en ausencia de otra solución, será racional adoptarlas siempre y cuando sus desventajas no hagan que la situación resultante sea aun peor que la no cooperación.

Las alternativas numeradas del 3 al 5 son soluciones psicológicas Todas ellas solucionan el dilema, e.d., hacen que cada agente seleccione la estrategia A, debido a un cambio producido en el propio agente. Estas soluciones son soluciones morales cuando el cambio operado en el agente no es un cambio de carácter específico sino de tipo general. Al hablar de "cambio de carácter específico" nos referimos a cambios que sólo representa una solución para un dilema concreto. Un caso de cambio específico sería el que se produciría en el ejemplo de los soldados mencionado más arriba si estos desarrollaran una tendencia compulsiva a obedecer las órdenes de los generales, o si adquirieran un odio hacia el enemigo que les hiciera desear su muerte aun a costa de arriesgar seriamente su propia vida. Estos cambios producidos en los soldados harían que en el caso concreto de la elección entre desertar y permanecer en sus puestos todos eligieran esto último. Pero si estos mismos soldados se encontrarán involucrados en un situación distinta, por ejemplo si se encontrarán en el dilema de pagar o no impuestos, la disposición adquirida no serviría de nada y ellos volverían a elegir la no cooperación.

Las soluciones morales son siempre de carácter general, y consisten en unos cambios en el agente que son operativos en una amplia gama de situaciones. Estos cambios de carácter moral producidos en el agente pueden dar lugar a las soluciones numeradas del 3 al 5. Si consideramos estas soluciones veremos que se dividen en dos grupos. Por un lado están las soluciones 3 y 4. En ellas, la estrategia A deja de ser peor para el agente. La diferencia entre ellas esta en que en la solución 3 el motivo por el que el agente esta dispuesto a seleccionar la estrategia A es por que A es ahora mejor para el. Sin embargo, en la solución 4, el agente actúa motivado por el cambio moral operado en él y el hecho de que A, debido a este cambio, ya no sea peor para él, es un simple efecto colateral. La solución 5, por otro

lado, tiene lugar cuando el agente elige A a pesar de que esta elección sigue siendo peor para él.

Aparte de su característica común de resolver los dilemas operando en el agente un cambio de actitud general, estos cambios, y por consiguiente, las soluciones morales, son diversos. El estudio de estos cambios y de sus ventajas relativas tiene por sí mismo un considerable interés, y dedicaremos el resto del artículo al mismo.

Qué moralidad escoger para salir del Dilema

Parfit señala cuatro de estos cambios de carácter moral que, de darse en todos los agentes involucrados en un dilema, darían solución a este[12]. En primer lugar, el agente puede convertirse en una persona *digna de confianza*. Puesto que sabemos que en determinadas situaciones es racional llegar a un acuerdo, los agentes implicados lo harán. Una vez alcanzado este, todos prometerán llevarlo a cabo. Y si son personas dignas de confianza todos cumplirán lo prometido, e.d., todas realizarán la estrategia A.

En segundo lugar, el agente puede hacerse reacios a ser aprovechados. Un agente con esta característica preferirá cumplir el acuerdo si los demás lo hacen a pesar de que ganaría más no haciéndolo, debido a la repugnancia que le produce obtener beneficios a costa de los demás. Por eso, si sabe que los demás harán A, el también hará A.

En tercer lugar, el agente puede hacerse *kantiano*. Ante cualquier alternativa, el agente se preguntará si puede querer racionalmente que todos actúen según esa estrategia y sólo la realizará él mismo si la respuesta es positiva. Puesto que la situación en la que todos hacen B es una situación en la que todos salen perdiendo, el agente así dispuesto no realizará B, sino A.

Por último, el agente puede convertirse en una persona *altruista*. En una situación en la que puede elegir entre una estrategia B que le beneficia a él y una estrategia A que beneficia al otro, siempre elegirá A.

Cada uno de estos cambios tiene sus inconvenientes a la hora de solucionar el dilema. Según Parfit, los dos primeros no solucionarían todas las

[12] Es posible que alguien pueda pensar en algún cambio de actitud moral distinto de estos cuatro. Yo no encuentro ninguno, pero, naturalmente esto no quiere decir que no pueda haberlos. Esto sin embargo no me parece de mucha importancia. Mi propósito es analizar el funcionamiento de las soluciones morales en general y, más concretamente, que implicaciones teóricas tiene la existencia de la solución 5. Y basta con que esta solución pueda plantearse con cualquier tipo de cambio moral, cosa que parece indiscutible.

situaciones de Dilema, sino sólo las versiones más simples. Sin embargo algunos de los inconvenientes señalados no son importantes. Por ejemplo, el segundo de estos cambios presenta, en opinión de Parfit, el inconveniente de que, si hay un número suficiente de personas con esta característica, entonces no sería necesaria la existencia de un acuerdo condicional[13]. Bastaría con la seguridad de que un número suficiente va a realizar A para que los que tienen esta disposición hicieran A a su vez. Pero los agentes así dispuestos no podría crear esta seguridad. Este inconveniente puede tener su importancia, pero puede ser superado si suponemos la existencia de un acuerdo condicional y suponemos también el conocimiento por parte de cada uno de los agentes de lo que los otros están haciendo.

Otro problema más importante relacionado con este segundo cambio moral y al que sin embargo Parfit no presta tanta atención es quién va a establecer la práctica de la cooperación. Si el único motivo que mueve a los agentes es su repugnancia a ser aprovechados, una vez que existe una práctica en la que todos o un número suficiente cooperan, estos cooperarán también. Pero si esta práctica esta aun por establecerse, un agente impulsado sólo por esta motivación no podría dar el primer paso. Si no hay nadie cooperando, no se puede ser un aprovechado por no cooperar. Pongámonos en el puesto de un agente con esta característica que tienen que elegir entre B y A. ¿Cuales son los motivos para realizar B?. Según hemos dicho, el motivo es que si los demás hacen B, lo mejor es hacer B también, y si los demás hace A también lo es. Esto último es precisamente lo que cambia en este caso. Si los demás o un número suficiente hacen A, entonces lo mejor para mi es hacer A también, porque de otro modo estaría actuando como un aprovechado, cosa que por hipótesis quiero evitar. Por tanto, sólo en el supuesto de que los demás o un número suficiente de los demás van a hacer B tendré yo motivos para hacer B. Que yo sepa que todos tienen la misma disposición que yo y que todos saben esto de todos los demás no soluciona nada.

---

[13] Con "**acuerdo condicional**" Parfit se refiere al acuerdo de realizar determinada estrategia conjunta si los demás, por su parte, también lo hace.

Es importante no confundir este acuerdo condicional, que en algunos casos parece prescindible, con el acuerdo al que se llega acerca de la estrategia conjunta a realizar. Este no es prescindible en ningún caso. En el caso del Dilema del Prisionero esta paso es trivial pues solo hay una solución óptima posible. En los casos más complejos en los que hay varios resultados óptimos es imprescindible un acuerdo previo acerca del resultado concreto a perseguir. Sólo cuando existe este acuerdo es posible plantearse si seguirlo o no, o si seguirlo sólo a condición de que los demás lo hagan también.

El cambio consistente en que los agentes se conviertan en personas dignas de confianza tiene problemas aun mayores. Si cada uno sólo se comprometerá a realizar su parte si todos los demás también se comprometen, esta característica moral de los agentes de ser personas dignas de confianza puede solucionar el problema. Pero otra cosa muy distinta sucede si no se alcanza una unanimidad total. En efecto, puede no ser necesaria ni posible la unanimidad, sino que puede suceder que la cooperación de un número n de personas sea suficiente para lograr el resultado apetecido. Por ejemplo, supongamos que 3 personas quieren ir a jugar al frontón, pero para eso tienen que ponerse de acuerdo en quien irá a por la pista. Como quieren jugar dos horas (menos no vale la pena) tienen que ir dos, ya que las normas del polideportivo establecen que sólo se alquilará una hora por persona. Los tres son personas de confianza y si se comprometen a ir, irán. Ahora bien, cada uno puede hacer el siguiente razonamiento. Si los otros dos se comprometen a ir, yo no necesito comprometerme. Y si ninguno de ellos lo hace, entonces no sirve de nada que lo haga yo. Sólo si uno (y sólo uno) de ellos se compromete haré yo bien en comprometerme también. Por supuesto que si me comprometo a colaborar cumpliré mi palabra, pero sólo en el caso de que yo haga el número 2 será razonable que adquiera ese compromiso.

Esta situación es parecida a la que se plantea entre agentes no dignos de confianza. En ese caso vimos que es racional cooperar si mi cooperación es condición necesaria para que el resultado deseado se realice, y si es necesaria la participación de todos los implicados entonces el acuerdo es estable. Pero en otro caso surgía el dilema. Aquí sucede algo similar. La diferencia está en que si los agentes en cuestión son dignos de confianza yo se que, si se han comprometido a hacer A, harán A. Y si hacen A en un número suficiente, entonces yo no gana nada comprometiéndome a hacer A. Si me comprometo a ello tendré que hacerlo, pero es mejor para mi hacer B. Y si no se han comprometido, también es mejor que haga B.

Sin embargo, y a pesar del parecido, la situación no es la misma. Porque en el primer caso el problema es si yo voy o no voy a cumplir el acuerdo, mientras que en el segundo caso, e.d., en el caso de que se trata de personas dispuestas a mantener el acuerdo, el problema se transforma en qué acuerdo va a lograrse. Puede en efecto efectuarse una negociación para acordar una estrategia conjunta de la forma "X e Y irán y Z no irá". Si nadie puede hacer una presión razonable para ser él el favorecido por el trato, pueden jugárselo a los chinos. Y una vez acordada esta estrategia, todos la seguirán. De hecho, ante una situación como esta, esto es lo que hace la gente

digna de confianza. Naturalmente, en situaciones que involucran a más jugadores la situación es más difícil de resolver. Pero la dificultad nunca está en si uno se comprometerá a hacer algo o no. La dificultad está, más bien, en cómo acordar la estrategia conjunta a seguir. Pero una vez que esta ha sido acordada, cada uno cumplirá su parte. Dicho de otra forma, a lo que cada uno se compromete es a cumplir su parte. Y cual sea esta parte es algo que se decide en el proceso de negociación mediante procedimientos de regateo.

Puede plantearse en este punto una objeción. Esta surge porque cada uno se compromete a realizar su parte después de que se haya acordado qué parte corresponde a cada uno. Ahora bien, los acuerdos para los que no se requiere unanimidad son tales que el resultado se producirá aunque no todos cumplan su parte. Hablando entre gente decente, aunque no todos se comprometan a cumplir lo que les ha tocado. En el ejemplo anterior, si se acuerda que vayan los tres, el resultado deseado (la obtención de la pista) se sigue aunque sólo dos se comprometan (y cumplan, cosa que ahora damos por supuesta). Pero, de nuevo, el problema es ¿quien se compromete? Y el dilema vuelve a plantearse.

Esta objeción, así planteada, puede ser contestada con facilidad. Por definición, el resultado del acuerdo será óptimo. Y un acuerdo que consiste en que vayan tres donde pueden ir dos no lo es. Hay otra situación que es mejor para alguno sin que sea peor para ninguno, a saber, la situación en la que sólo van los necesarios. Por consiguiente, parece que no puede plantearse un acuerdo en el que el costo de uno sea gratuito. Sin embargo, hay otros casos en los que la respuesta no es tan sencilla, a saber, los casos del tipo del dilema del contribuyente. En estos casos la contribución individual se ve como algo insignificante que no hace ninguna diferencia. De modo que si poca gente se compromete a pagar, que yo pague no sirve para alcanzar los bienes públicos deseados, y si pagan los suficientes, seguro que aunque yo no pague no va a pasar nada. De modo que lo mejor es no comprometerme. Por supuesto que si yo no colaboro el resultado no será exactamente igual, pero si infinitamente parecido. Por ello, el que todos seamos dignos de confianza no parece suficiente para solucionar este tipo de casos[14].

[14] Sin embargo, esta solución es buena si se trata de acuerdos condicionales, donde el compromiso de todos los demás es condición indispensable para que se comprometa cada uno. Podría aun surgir, como en el caso anterior, el problema de quien se comprometa primero, pero esto puede solucionarse de diversos modos, por ejemplo mediante compromisos simultáneos.

El tercer cambio, consistente en que los agentes se conviertan en kantianos[15], y el último, según el cual los agente se harían altruistas, parecen ser los menos problemáticos. Sobre este último cambio conviene no obstante hacer una matización. Una vez realizado el acuerdo, si todos somos altruistas el problema se resuelve bien. Lo mejor para ti es que yo colabore, de modo que haré A. Supuesto que el otro también es altruista, hará A también, de modo que la cooperación está garantizada. Pero que los agentes sean altruistas al realizar el acuerdo, en contra de lo que pueda parecer a primera vista, no cambia el carácter de este proceso. Esto puede ejemplificarse bien en el caso de los prisioneros. Supongamos que ambos son altruistas y prefieren siempre el bien del otro al beneficio propio. Esto invierte sus preferencias acerca de los resultados respecto a las que tendrían dos jugadores no altruistas. Si utilizamos los términos empleados por Poundstone en el texto citado más arriba, el orden en que los jugadores no altruistas prefieren los distintos estados cosas posibles es: tentación, recompensa, castigo y hacer el primo. Por otro lado, si los agentes involucrados fueran altruistas, su ordenación sería: hacer el primo, recompensa, castigo y tentación. De todos los puntos óptimos posibles, yo prefiero los que te favorecen a ti y tu los que me favorecen a mi, de modo que, al igual que sucede entre jugadores auto-interesados[16].El dilema sería exactamente el mismo, salvo que, en el caso de los jugadores altruistas, cada uno estaría jugando, por así decirlo, por el otro.

Este dilema surgiría entre dos altruistas puros cuando se encuentran en situación de elegir entre a) darse a si mismo un gran beneficio o b) dar al otro un beneficio más pequeño. Esta *condición Positival* sustituiría a la *Condición Positiva* originaria. Un altruista puro elegiría b, con lo cual el resultado sería peor para todos. Por ejemplo, consideremos la siguiente matriz:

[15] Tal y como señala Parfit, el test kantiano, tomado literalmente, tiene sus propios problemas. El ejemplo habitual de estos es el de la práctica de una profesión, digamos la medicina. Yo no puedo racionalmente desear que todo el mundo practique la medicina, de modo que yo tampoco debo hacerlo. Y si todos somos kantianos, es de temer que las consecuencias fueran desastrosas. No obstante, el test puede refinarse lo suficiente como para impedir que surjan estos problemas. Por ejemplo, puede plantearse que la aplicación del test se limita a los casos en los que todos deseamos hacer algo, es decir, a los casos en los que la pregunta ¿que pasaría si todos hiciéramos lo mismo? surge de manera natural.

[16] Hay que notar que en el Dilema del prisionero todos los puntos son óptimos excepto, precisamente, el que llamamos castigo.

	B_1	B_2
A_1	(1,1)	(0,9)
A_2	(9,0)	(8,8)

Cada uno de los jugadores puede a) darle 1 al otro o b) darse 8 a si mismo. Llamaremos A a la estrategia que da 1 al otro (e.d., las estrategias A1 y B1) y B a la que da 8 a uno mismo (A2 y B2). Si ambos eligen A, el resultado será (1,1). Pero hay otro resultado que los dos preferirían a este, a saber (8,8). Supongamos que los dos se ponen de acuerdo en no ayudarse y elegir ambos la estrategia B, con el fin de obtener ese resultado. Pero, de nuevo, este acuerdo no es estable. El jugador uno razonaría así. Si el jugador 2 mantiene el acuerdo y hace B, entonces es mejor para él que yo haga A. Y si no lo mantiene y hace A, entonces también es mejor para él que yo haga A. De modo que haré A. El jugador 2 haría un razonamiento similar y llegaría a la misma conclusión. Por tanto, sería (1,1) el resultado.

Parece por tanto que la transformación de los agentes en altruistas puros no solucionaría los dilemas. Parfit matiza esta cuarta vía diciendo que los participantes no deben ser altruistas puros, es decir, personas que no tienen ningún interés por sí mismas, ya que de este modo se plantearía un dilema parecido al que se intenta solucionar mediante los cambios morales. Más bien, la cuarta solución moral deberá consistir en que los agentes sean lo suficientemente altruistas, no sobrepasando en ningún caso el límite de la benevolencia imparcial, es decir, una consideración imparcial de los intereses de todos, incluidos los propios.

Podemos concluir esta parte del análisis afirmando que los cuatro cambios morales solucionarían el problema práctico presentado por el Dilema, si bien los dos primeros resultan más problemáticos. Queda aun por analizar la relación de tales cambios con las distintas soluciones apuntadas.

Qué solución aportan los cambios morales

Según quedó dicho más arriba, las soluciones morales pueden resolver el dilema con las soluciones 3, 4 y 5. Vimos también que estas podían clasificarse en dos grupos: en las soluciones 3 y 4, la estrategia A deja de ser peor para el agente. En la solución 3 el motivo por el que el agente esta dispuesto a seleccionar la estrategia A es por que A es ahora mejor para el y en la solución 4, el agente actúa motivado por el cambio moral operado en

él y el hecho de que A ya no sea peor para él es un simple efecto colateral. En la solución 5, por otro lado, el agente elige A a pesar de que esta elección sigue siendo peor para él

Quiero insistir en esta última parte del artículo en que cualquiera de los cambios morales apuntados pueden dar lugar a cualquiera de las soluciones morales. Dado que sería demasiado largo dar un ejemplo de cada motivación en relación con cada una de las soluciones, me limitaré ver algunos de ellos que pueden resultar más problemáticos.

Consideremos el primer cambio moral, consistente en que el agente se convierta en una persona de confianza. Debido al carácter de este cambio, puede pensarse que una persona así dispuesta elegirá siempre A con independencia de que A sea mejor o peor para el. Es decir, puede pensarse que un agente con esta disposición elegirá siempre A movido, no por la consideración de los beneficios que A le suponga, sino, simplemente, porque se ha comprometido a realizar A. Esto es en parte verdad. Un agente así dispuesto respetará el acuerdo porque se ha comprometido a hacerlo. Pero este "porque" no tiene necesariamente una lectura causal. Su lectura fundamental es condicional, es decir, un agente así dispuesto cumplirá el trato si se ha comprometido a hacerlo, puesto que es una persona de confianza. Pero puede suceder que un agente con esta disposición elija A porque para él ser un hombre de palabra tenga tanta importancia que, de no cumplir lo acordado, los beneficios conseguidos no compensen el haber perdido su fama de hombre honrado, ante los demás o ante si mismo[17]. No voy a discutir si las personas de confianza dan lugar por lo general a las soluciones 4 y 5 y sólo excepcionalmente a la 3 o viceversa. La cuestión importante es que este cambio moral puede dar lugar a todas ellas.

Respecto a los cambios morales mediante los cuales los agentes pasan a ser reacios a ser aprovechados o se convierten en kantianos puede pensarse también que dan lugar a las soluciones 4 y 5. Sin embargo esto no es así, por el mismo motivo que no lo era en el caso anterior.

[17] Tomada de modo literal, la presentación de las soluciones morales que hace Parfit parece obligar a esta posibilidad, ya que la solución 3 surge cuando el agente tiene una disposición a hacer A y esta disposición ha sido adquirida porque A es mejor para él. Sin embargo, creo que esta interpretación es errónea. No se trata de por qué un agente haya adquirido una determinada disposición, sino de por qué este dispuesto a hacer A en un momento determinado. Por tanto, la solución 3 se produce cuando un agente tiene una disposición a realizar A debido a que A, a raíz de un cambio interno del agente, es mejor para él.

Acerca de el último cambio moral, por el contrario, puede pensarse que sólo puede dar lugar a la solución 3. Si alguien adquiere un interés suficiente por los asuntos de los demás, y más aun si llega al punto de conceder la misma importancia a estos que a los suyos propios, seguramente la estrategia A será mejor para el y, por esto, estará dispuesto a realizar A. Pero tampoco en esta ocasión sucede esto de manera necesaria. Desde luego, es indudable que nuestros intereses y preferencias, y, por tanto, nuestra función de utilidad, se alteran cuando empezamos a tomar en cuenta los intereses de los demás. Pero también es cierto que siempre hay un sentido en que nuestros propios intereses y los intereses de los demás son distintos, aunque nosotros estemos dispuestos a tomar en cuenta estos últimos.

Puede suceder que algunas personas tengan lo que podríamos llamar un "interés personal" por el bienestar de todo el mundo, es decir, que sientan hacia el bienestar de todos los demás el mismo tipo de interés que otro puede sentir por el de sus hijos, pareja, padres, amigos íntimos o por el de uno mismo. Mostrarían, en términos de Hume, una benevolencia universal e ilimitada. Para tales personas, su función de utilidad esta determinada por estos intereses, del mismo modo que la mía lo esta por mi interés hacia ciertas personas. También en estos casos es posible distinguir entre sus intereses y los de los demás. En parte esta distinción se hace debido a que, de hecho, la mayoría de estas personas tienen una marcada tendencia a opinar sobre cuales son los intereses de los demás, separándose en este punto de la opinión de los propios interesados. Aparte de este hecho, la distinción se establece acudiendo a lo que serían los intereses de estas personas si estas no se preocuparan tanto de los demás, cosa siempre posible. Esta distinción es la que hace posible decir que alguien está obrando en contra de su propio interés.

Por lo general, en estos casos estamos más dispuestos a hacer esta distinción de lo que lo estamos, por ejemplo, en el caso de una madre con su hijo. Cuando una madre se sacrifica por su hijo, incluso cuando llega a poner en peligro su propia vida, en cierto modo podemos decir que actúa contra sus propios intereses, pero realmente no consideramos que los intereses de su hijo no sean los suyos también. El motivo es que consideramos la preocupación de la madre como algo natural, o, si se prefiere, como algo común que nos ocurre a la gran mayoría de los seres humanos[18]. Aunque

[18] También en este caso hay matices. Habitualmente, no nos causa mucho problema afirmar que los intereses del hijo son los de la madre cuando el primero tiene 10 años. Cuando tiene 40 nos sentimos más inclinados a distinguir entre los intereses de ambos.

en menor grado, lo mismo ocurre con el interés de un hombre por sus amigos. No obstante, aquí no vamos a ocuparnos de si un interés personal por el bienestar de todo el mundo es natural o no. Lo cierto es que algunas personas (pocas, por cierto) lo sienten. Y cuando esto sucede, y estas personas escogen la estrategia A, se produce la solución 3.

Sin embargo, esto no es lo que sucede en todos los casos en los que el agente es lo suficientemente altruista para elegir A. Yo puedo estar interesada en el bienestar de los demás, por ejemplo en el bienestar de los argentinos, sin que esto se confunda en ningún momento con el interés personal que siento por un amigo. La diferencia es tremendamente difícil de señalar, pero existe.

Si yo tuviera que elegir entre salvar la vida de uno de aquellos o hacerle un regalo a mi amigo, elegiría lo primero. Pero si tuviera que elegir entre hacer un regalo a uno de los dos de tal modo que mi regalo les beneficiara a ambos igual, o incluso si beneficiara a mi amigo un poco menos, le haría el regalo a mi amigo. Si viera que alguien esta maltratando a mi amigo, saldría corriendo sin pensarlo dos veces, a cualquier hora y a cualquier lugar, para ayudarle. Cuando veo un reportaje en el que alguien está siendo maltratado, me indigno e incluso, en ocasiones, pienso si debería correr a prestar mi ayuda, pero no lo hago. El reportaje acaba y empieza la película. Si la cosa era muy tremenda, me siento mal durante un rato, y es posible que me apunte a algún organismo internacional de ayuda. Esto indica una diferencia de grado.

Por otra parte, si tuviera que entregar algo valioso para sacar a un amigo de un lío, lo haría gustosa, mientras que si tuviera que hacerlo por un desconocido, aunque lo haría (al menos, lo haría si yo fuera una persona suficientemente altruista) no lo haría tan gustosamente. Me costaría más. Ante el agradecimiento de mi amigo, diría "de nada" de todo corazón. Ante el del extraño, estas palabras no pasarían de ser una frase hecha. Esto indica algo más que una diferencia de grado. Indica hasta qué punto mi interés por el bienestar de mi amigo es personal, mientras que mi interés por el extraño no lo es.

Mi interés personal por mi amigo es algo "natural", un interés que yo, simplemente, tengo, como lo tengo, en mayor o menor grado, por mi misma. Mi interés no personal por los extraños no es natural en este sentido. O, mejor dicho, si lo es un cierto interés. A mi no me da igual lo que les suceda a los demás. Prefiero que les vaya bien. Esto es un interés que yo tengo, igual que tengo interés en que determinado partido gane las elecciones o en que mi coche sea de un determinado color. Pero un interés que

pueda ser calificado de "suficientemente altruista", es decir, que me haga estar dispuesto a elegir la estrategia A en vez de B, ya es otro asunto. Este tipo de interés sólo lo tengo si me obligo a considerar la situación desde un punto de vista imparcial. Este punto de vista siempre puede ser adoptado. Pero hacerlo o no es algo que depende de mi voluntad. No es el punto de vista que yo tengo. Es el punto de vista que yo puedo adoptar si me determino a hacerlo. Y puede ser que yo piense que debo adoptar ese punto de vista. Pero esto no tiene nada que ver con mis intereses personales. Es algo que debo hacer. Si lo hago, y desde ese punto de vista realizo mi elección, elegiré A. Pero lo elegiré tanto si A es mejor para mi como si no lo es. Puede ser que, al adoptar ese punto de vista, A ya no sea peor para mi. Pero también puede suceder que si lo sea. Y, a pesar de eso, elegiré A. De modo que podemos concluir que el cambio moral consistente en que el agente pase a tener suficiente altruismo puede dar lugar tanto a la solución 3 como a la 4 y la 5.

Creo que estas consideraciones son suficientes para establecer que ninguno de los cuatro cambios morales analizados está necesariamente ligado a una determinada solución moral de los dilemas. Si esto es cierto, parece que la elección entre los distintos cambios morales analizados, especialmente entre los dos últimos y menos problemáticos de convertirse en kantiano o en una persona suficientemente o moderadamente altruista no puede dirimirse por su comportamiento en los casos de dilema. Este hecho es especialmente digno de atención si consideramos que las distintas soluciones afectan de forma también distinta al dilema. Lo que tienen en común las soluciones pertenecientes al primer grupo, esto es las soluciones 3 y 4, es que representan situaciones en las cuales la función de utilidad de los agentes se transforma de tal modo que el dilema deja de ser tal. Por tanto, no solo presentan una solución al problema práctico, sino que disuelven el problema teórico: en tanto que anulan la Condición Positiva, el dilema, simplemente, desaparece. Otra cosa muy distinta es lo que sucede con la solución 5. Al igual que las otras, resuelve el problema práctico, pero no el problema teórico. Este tampoco se disuelve, sino que, al contrario, se presenta en toda su magnitud: mediante esta solución práctica, los agentes hacen lo que consideran peor para sí mismos y, sin embargo y precisamente por esto, consiguen lo que es mejor para todos. Todos lo cambios morales resulelven el problema práctico. Ninguno de ellos, en los casos en los que conduce a la solución 5, resuelve el problema teórico. Pero puesto que nuestro objetivo presente era analizar las soluciones al problema práctico planteado por el dilema, y este se alcanza con cualquiera de las soluciones

mencionadas, podemos decir del problema teórico que eso ya es otra historia.

REFERENCIAS BIBLIOGRÁFICAS

Bentham (89), *Introduction to the Principles of Morals and Legislation*, Londres 1982.
Elster (79), *Ulysses and the Sirens*, Cambridge U.P., 1979.
Mill (74), *El Utilitarismo*, Madrid, Alianza Editorial, 1984.
von Neumann-Morgenstern (44) *Theory of Games and Economic Behaviour*, Princeton U.P. 1944.
Parfit (79), *Prudencia, moralidad y el Dilema del Prisionero*, Madrid, U. Complutense 1991.
—— (86), *Reasons and Persons*, Oxford U.P., 1986.
Poundstone (92), *El Dilema del Prisionero*, Madrid, Alianza Editorial, 1995.
Rawls (77), *A Theory of Justice*, Cambridge, Harvard U.P., 1977.
Sidgwick (74), *The Methods of Ethics*, Cambridge, Hackelt P.C., 1981.
Weber (21), *Economia y sociedad*, Mejico, F.C.E., 1944.