Reciprocidad y castigo en la explicación evolucionista de la cooperación

Fernando Melo Acosta

Universidad Nacional de Colombia

Facultad De Ciencias Humanas

Departamento De Filosofía

Bogotá D.C., Colombia

2010

Reciprocidad y castigo en la explicación evolucionista de la cooperación

Fernando Melo Acosta

Tesis presentada como requisito parcial
para optar al título de Magister en Filosofía

Director: M.A. Iván Darío González Cabrera

Codirector: Ph. D. William Augusto Duica Cuervo

Universidad Nacional de Colombia
Facultad De Ciencias Humanas
Departamento De Filosofía
Bogotá D.C., Colombia
2010

A la memoria de mi padre

Agradecimientos

Debo expresar mi agradecimiento al Departamento de Filosofía de la Universidad Nacional de Colombia y especialmente al profesor Iván Darío González Cabrera por el valioso aporte durante la preparación de este trabajo y a los profesores Alejandro Rosas López y William Augusto Duica Cuervo por todo el apoyo proporcionado.

Resumen

En la teoría de la evolución se han desarrollado dos aproximaciones principales frente al problema de la cooperación, el modelo basado en la reciprocidad y el modelo del castigo. En este trabajo se desarrolla una reconstrucción del problema de la cooperación desde la perspectiva evolucionista y se presentan los elementos conceptuales de los mecanismos de reciprocidad y castigo. Se plantea que la principal diferencia entre estos mecanismos consiste en que el castigo, a diferencia de la reciprocidad, incorpora costos adicionales en la función de pagos que modifican el resultado en términos de aptitud. Se argumenta que, en contextos de interacciones multipersonales, no se requiere que la conducta cooperativa sea condicional a la conducta cooperativa de todos los demás individuos como tradicionalmente se ha entendido, sino que sea condicional solamente respecto a un grado proporcional de cooperación por parte de los demás.

Palabras clave: Reciprocidad - Castigo – Cooperación – Evolución de la Cooperación – Selección Natural

Abstract

Evolutionary theory has developed two main approaches to solve the problem of cooperation, the reciprocity-based model and the model of punishment. This work redefines the problem of cooperation from the evolutionary standpoint and provides the conceptual elements of reciprocity and punishment mechanisms. I argue that the main difference between these two mechanisms is that punishment, unlike reciprocity, includes additional costs in the payoff matrix that modify the results in terms of fitness. It is argued that, in multi-person interactions, it is not required that cooperative behavior to be conditional upon the cooperative behavior from all other individuals as traditionally understood, but it is conditional only to a proportional degree of cooperation from others.

Keywords: Reciprocity – Punishment – Cooperation – Evolution of Cooperation – Natural Selection

CONTENIDO

INTRODUCCIÓN

La cooperación entre los organismos se observa con frecuencia en la naturaleza y se ha desarrollado en gran escala en el caso particular de la especie humana. La explicación acerca del origen y estabilidad de la cooperación constituye entonces uno de los principales interrogantes para la teoría de la evolución y en esa materia se han desarrollado varias aproximaciones así como una amplia actividad de investigación experimental.

En esa tarea, el problema general que debe resolver la teoría de la evolución consiste en explicar la razón por la cual la selección natural favorece conductas cooperativas, en situaciones en las cuales, frente a otras alternativas de conducta, la predicción de la teoría consiste en que la cooperación será el resultado menos probable.

Frente a este dilema, en la teoría evolucionista se han desarrollado dos aproximaciones principales, por un lado, el modelo basado en la reciprocidad y por otro, el modelo basado en el castigo, los cuales se han propuesto como mecanismos que explican el surgimiento de la cooperación y su mantenimiento constante. Aunque es abundante el material escrito sobre el tema, con frecuencia los investigadores enfocan su atención hacia la descripción y análisis de las pruebas experimentales, sin que aborden con detalle los elementos conceptuales básicos de los mecanismos de reciprocidad y castigo.

El propósito de este trabajo consiste en realizar una exposición crítica de la discusión e investigación acerca del problema de la cooperación, enfocada hacia dos objetivos, por un lado, desarrollar una reconstrucción de este problema desde la perspectiva evolucionista, y por otro lado, proporcionar claridad conceptual en torno

a los mecanismos de reciprocidad y castigo que se han propuesto como solución a este problema. En el desarrollo de esta tarea, se hará énfasis en el caso particular de la cooperación en la especie humana.

La tarea inicial consistirá en clarificar el problema que enfrenta la teoría evolucionista en la explicación de la cooperación. Para ese propósito será necesaria la exposición con detalle de los modelos evolutivos que se han desarrollado a partir de los mecanismos de la reciprocidad y el castigo.

Una vez se hayan clarificado los elementos conceptuales del problema de la cooperación, será necesaria una breve referencia acerca de aquellos planteamientos que se han formulado acerca de la conceptualización de los mecanismos evolutivos de reciprocidad y castigo. Esta referencia permitirá mostrar la necesidad de esclarecer la distinción entre estos mecanismos, cuestión que tiene incidencia en la discusión acerca de la eficacia de estos mecanismos en la solución del problema de la cooperación. Esa tarea de esclarecimiento será objeto de las secciones subsiguientes, en las cuales se aborda el estudio de los elementos conceptuales básicos que permiten caracterizar cada uno de estos mecanismos, teniendo en cuenta como presupuesto fundamental la identificación del problema de la cooperación realizada en las primeras secciones del trabajo.

En desarrollo de esta reflexión se analiza en primer lugar la caracterización de los rasgos básicos de la estrategia de reciprocidad, comenzando por el modelo simple de la estructura diádica de interacción para luego pasar a revisar el mismo tema bajo el marco de las interacciones multipersonales. Posteriormente se examinan los principales aspectos característicos del castigo, siguiendo una ruta similar a la empleada para el caso de la reciprocidad, de modo que pueda ofrecerse un contraste entre ambos mecanismos evolutivos.

1. EL PROBLEMA DE LA COOPERACIÓN

1.1. Cooperación y conducta cooperativa

En la naturaleza se pueden observar muchos casos de conductas cooperativas en especies animales no humanas y, aunque no es la regla general, se pueden mencionar algunos ejemplos de cooperación que tienen cierta semejanza con las conductas sociales propias de la especie humana, como es el caso de los llamados de alerta en algunas aves, la simbiosis entre individuos de distintas especies, la donación de sangre entre vampiros y el caso de los insectos sociales con castas de individuos estériles cuya exclusiva función consiste en servir a la reina de la colonia. Estos casos de cooperación en el mundo animal se caracterizan o bien por involucrar pocos individuos en la empresa cooperativa o bien por ser ejemplo de cooperación a gran escala entre individuos con una fuerte relación de parentesco, como en el caso de los insectos sociales. En la especie humana, en contraste, la cooperación se ha desarrollado en gran escala entre individuos si ningún grado especial de parentesco, y el origen de este patrón de conducta plantea un acertijo no solamente para la teoría de la evolución sino también para las ciencias económicas y la psicología.

Desde la perspectiva de la biología evolutiva, la cooperación puede definirse en términos de su resultado y de ese modo se entiende como la acción conjunta que produce un beneficio neto para todos los individuos involucrados (Clements & Stephens 1995, Bergmüller et al. 2007a 2007b). Se dice que es una acción conjunta pues requiere el despliegue coordinado de conducta de varios individuos, esto es, por lo menos dos individuos, por lo que no habrá cooperación en las acciones que involucran solamente un individuo. La acción debe ser coordinada en cuanto que las acciones de cada individuo deben estar dirigidas hacia el mismo resultado. Esta

acción conjunta debe estar encaminada al beneficio mutuo, esto es, en términos biológicos, debe propender hacia el aumento de la aptitud biológica de los individuos involucrados. Este beneficio debe ser neto, esto es, se requiere que la diferencia entre el beneficio total y el costo de la acción cooperativa sea positiva, pues de lo contrario, esto es, si el beneficio es menor que el costo, entonces se produce un detrimento en la aptitud biológica. Adicionalmente, es importante que el beneficio sea para todos los individuos involucrados pues esta característica permite distinguir la cooperación respecto de otras interacciones como el parasitismo y la competencia. En el caso del parasitismo se genera un beneficio neto para un individuo y un costo neto para el otro, mientras que en el caso de la competencia se produce un costo neto para ambos individuos (Bergmüller et.al. 2007a).

En biología evolutiva es pertinente distinguir entre cooperación y conducta cooperativa. La acción de cooperar consiste en alcanzar la cooperación o comportarse de manera cooperativa, esto es, de una manera tal que se dirige a producir la cooperación (Mesterton-Gibbons & Dugatkin 1992; Mesterton-Gibbons & Dugatkin 1997). Esto significa que es lógicamente posible que haya conducta cooperativa sin cooperación, puesto que, por ejemplo, un único individuo podría desplegar una conducta cooperativa, sin que esta suponga el despliegue coordinado de la conducta cooperativa de varios individuos, que es requisito para la existencia de cooperación.

Asimismo, la noción de conducta cooperativa supone un espectro amplio de comportamientos que incluye toda conducta que proporciona un beneficio a otro organismo, pero que bien puede beneficiar o por el contrario puede ser costosa para el individuo que la ejecuta. En el primer caso, se dice que la conducta genera un beneficio adaptativo directo y, en el segundo, el beneficio adaptativo es indirecto, esto es, derivado del parentesco según se explicará más adelante (West et al. 2007, Okasha 2003).

Aunque cooperación y conducta cooperativa son nociones estrechamente relacionadas, la distinción es muy útil para el entendimiento que busca proporcionar la teoría evolucionista. En ese sentido debe notarse que la cooperación como tal es el resultado del despliegue de la conducta cooperativa por los individuos y, por tanto, la explicación evolucionista debe dirigir su examen en primer lugar hacia encontrar las razones por las cuales la selección natural puede favorecer esta clase de conductas. La cooperación será entonces explicada de manera derivada, una vez se esclarece la posibilidad biológica de la evolución de la conducta cooperativa.

Con respecto a la noción de conducta cooperativa que se ha explicado es necesario, sin embargo, hacer una precisión que es señalada por West et al. (2007), se requiere que la conducta haya sido seleccionada por su efecto benéfico para el receptor, de modo que pueda diferenciarse de aquellas conductas que reportan beneficios al receptor como subproducto de una conducta diseñada para otro propósito. Para entender mejor este requisito, un ejemplo ilustrativo es presentado también por West y sus colegas. Cuando un elefante produce estiércol, es una conducta que beneficia al elefante pero también puede beneficiar al escarabajo estercolero que utiliza el excremento. El escarabajo estercolero, también conocido como escarabajo pelotero, toma una porción del estiércol con la que hace una bola y mediante rodamiento la lleva a otro lugar donde la entierra para alimentarse y para construir un nido en el que deposita sus huevos. Pero la conducta del elefante no puede considerarse como cooperativa, pues es claro que la conducta del elefante no ha sido seleccionada por el efecto benéfico que tiene para el escarabajo estercolero el cual se beneficia solamente como subproducto de una conducta que ha evolucionado con otra finalidad. Para que una conducta sea cooperativa, ésta debió haber evolucionado como resultado de los beneficios selectivos derivados de la cooperación, lo cual significa a su vez que, así como es lógicamente posible que haya conducta cooperativa sin cooperación, es también lógicamente posible que exista cooperación sin conducta cooperativa, pues la cooperación puede surgir como el resultado de la conducta de varios individuos como un subproducto. Luego, así definidas, las

nociones de cooperación y conducta cooperativa son lógicamente independientes y, por tanto, la relación entre ambas puede ser establecida sólo empíricamente.

1.2. El problema de la cooperación en la teoría evolucionista

En la teoría de la evolución se han planteado múltiples teorías y se ha desarrollado una amplia actividad de investigación experimental, con el objeto de proporcionar una explicación acerca del origen y condiciones de estabilidad de la cooperación, especialmente para el caso de los seres humanos.

Con el fin de proporcionar claridad en torno a esa tarea explicativa, el primer paso consiste en identificar de manera muy general los problemas que enfrenta la teoría de la evolución. En esta materia tradicionalmente se identifican dos problemas principales, el problema del altruismo y el problema de la acción colectiva.

El primer problema que enfrenta la teoría de la evolución consiste en explicar las conductas altruistas, esto es, aquellas en las cuales un individuo incurre en un aparente sacrificio para proporcionar un beneficio a otro organismo. Aunque es frecuente la utilización del término "altruismo" como sinónimo de cooperación, tiene un alcance más estrecho pues se trata de una conducta que siempre es costosa para el donante, mientras que en el caso de la conducta cooperativa, como ya se dijo, también comprende aquellas conductas que representan beneficio directo para el actor (West et al. 2007 Okasha, 2003).

El altruismo es un tipo de conducta que no solamente representa un desafío para la teoría de Darwin sino que además parece estar en contradicción con ella. La selección natural favorece aquellas conductas que aumentan la aptitud biológica del individuo mientras tiende a eliminar aquellas que disminuyen dicha aptitud y por tanto la predicción de la teoría consiste en que las conductas altruistas no pueden surgir ni mantenerse en el mundo biológico. En la competencia por la supervivencia

y la reproducción, los individuos egoístas tienen ventaja sobre los altruistas quienes entonces tienden a desaparecer.

Un segundo problema consiste en explicar la acción colectiva dirigida hacia la provisión de bienes públicos, instancia de cooperación en la cual, a diferencia del altruismo, en principio no se realiza un sacrificio a favor de otro individuo, sino que la conducta cooperativa le produce un beneficio a quien la ejecuta. Conforme se explica en detalle más adelante, cuando la cooperación entre los seres humanos produce un bien público, dada su naturaleza, estará disponible para todos los individuos y, en consecuencia, una mejor alternativa de conducta consiste en beneficiarse de ese bien sin incurrir en el costo que debe asumir cada individuo. La selección natural en principio favorece aquellas conductas que proporcionan un mayor beneficio al individuo y, por ende, nuevamente aquí la cooperación parece contradecir la predicción de la teoría.

En las secciones siguientes se exponen las principales soluciones que desde la perspectiva evolucionista se han dado a estos dos problemas así como sus limitaciones y dificultades, centrando la atención en los mecanismos de reciprocidad y castigo. A partir de los elementos que proporciona esta exposición, posteriormente se revisa nuevamente la cuestión y se replantean los elementos conceptuales que identifican el problema general de la evolución de la cooperación.

1.3. El problema del altruismo

La atención inicial de la teoría evolucionista se centró en el problema del altruismo básicamente por dos razones. En primer lugar, por cuanto el altruismo, al implicar un sacrificio del individuo, parece ser totalmente incompatible con la teoría de la evolución, mientras que los demás casos problemáticos de cooperación, tienen cierto grado de compatibilidad con la teoría, en la medida en que no involucran un sacrificio y, por el contrario, se produce un beneficio para el individuo. En segundo

lugar, es previsible que la explicación evolucionista del altruismo pueda extenderse a otros ámbitos, al menos en algunos de sus elementos, de manera que una vez esclarecido el problema del altruismo, a partir de allí se puede derivar una explicación análoga para solucionar el segundo problema asociado con la evolución de la cooperación.

1.3.1. Selección por parentesco

El primer intento de explicación del problema del altruismo bajo el marco del esquema clásico de selección individual fue realizado por Hamilton (1963, 1964a, 1964b) quien planteó que una conducta altruista puede ser seleccionada cuando beneficia a aquellos individuos que portan los mismos genes del individuo altruista, tesis conocida como la teoría de la selección por parentesco según denominación dada por Maynard Smith (1964).

Según Hamilton, para que gen pueda ser objeto de selección, es suficiente con que incremente la aptitud biológica de aquellos individuos portadores de réplicas de ese mismo gen. El mantenimiento de un gen no depende exclusivamente de la supervivencia del individuo portador sino que también depende de su replicación exitosa. Es por ello que en la naturaleza es común observar que los organismos proporcionan cuidado y alimento a sus descendientes a veces incluso mayor que el que proporcionan a sí mismos.

La idea fundamental de Hamilton consiste en que un gen puede ser exitoso no solamente mediante su propia reproducción sino también mediante la reproducción de aquellos individuos portadores de una réplica de ese gen, lo cual comprende no solamente sus descendientes sino también otros individuos emparentados. Entre más cercano sea el individuo emparentado existen mayores probabilidades de compartir el mismo gen y por tanto, bajo ciertas condiciones, puede evolucionar una conducta altruista hacia individuos emparentados. La denominada Regla de Hamilton en su formulación más conocida expresa que un gen altruista puede ser

objeto de selección cuando $rb>c$, donde r es el coeficiente de parentesco entre los individuos, b es el beneficio para el receptor y c es el costo para el donante. En otras palabras, el beneficio debe ser mayor que el costo del acto altruista pero ese beneficio debe medirse en función del grado de parentesco pues a medida en que exista menor grado de parentesco, en esa misma medida existe menor probabilidad de compartir una réplica del mismo gen.

Para ilustrar la Regla de Hamilton se puede utilizar un ejemplo sencillo que puede encontrarse en el estudio realizado por Alan H. Krakauer (2005) en el cual se llevaron a cabo observaciones de la conducta del pavo salvaje (Meleagris gallopavo) en la reserva natural Hastings en California entre 1999 y 2004. El pavo salvaje es una especie en la cual se forman coaliciones compuestas de dos a cuatro machos, para cortejar a las hembras y defenderlas de otros grupos o de machos solitarios. En estas coaliciones solamente uno de los machos (el macho dominante) se aparea con la hembra y el estudio buscaba explicar el comportamiento altruista de los restantes individuos de la coalición, los machos subordinados. Se encontró que las coaliciones estaban compuestas por parientes y se realizó un cálculo empleando la regla de Hamilton para determinar si se cumplía el requerimiento para que pueda darse la conducta altruista de los machos subordinados. El coeficiente de parentesco r se determinó en 0.42 teniendo en cuenta la media aritmética del coeficiente de parentesco entre los machos subordinados y el macho dominante. El beneficio para el receptor b se determinó en 6.1, calculado por la diferencia entre el número de crías promedio del macho dominante y el número de crías promedio de un macho solitario. El costo para los machos subordinados c se determinó en 0.9, calculado según el número de crías promedio de un macho solitario. Si se aplica la regla de Hamilton se tiene que $rb>c$ pues $0.42X6.1=2.562$, cifra superior a c que es 0.9. Por tanto, Krakauer concluye que la selección por parentesco permite explicar la conducta de cortejo cooperativo en el pavo salvaje. Los machos subordinados, si bien no se reproducen directamente, incrementan su aptitud biológica de forma indirecta, calculada según la regla de Hamilton, compensado el costo de la ayuda en términos de propagación de sus propios genes mediante individuos emparentados.

Aunque la tesis de Hamilton permite explicar la mayor parte de las conductas aparentemente altruistas que se observan en especies no humanas, no representó una solución definitiva al problema pues no explica aquellos casos de altruismo aparente hacia individuos sin parentesco cercano como ocurre de manera constante en la especie humana.

1.3.2. Altruismo recíproco

Ante la limitación del modelo de Hamilton la noción de reciprocidad surge como primera alternativa de solución. En su trabajo pionero, Trivers (1971) plantea la tesis del altruismo recíproco según la cual, bajo ciertas condiciones, las conductas altruistas hacia organismos no emparentados pueden evolucionar. Trivers planteó que la selección natural favorece estas conductas pues proporcionan beneficio a largo plazo para el organismo. Si el beneficio para el receptor es mayor que el costo para el donante, y posteriormente el receptor retorna la ayuda al donante, ello se traduce en un beneficio neto para el donante. Sin embargo, el receptor de la ayuda puede optar por hacer trampa y no devolver la ayuda. Según Trivers la selección desfavorece la trampa cuando sus efectos adversos superan los beneficios de la reciprocidad, lo cual puede ocurrir si el altruista responde a la trampa mediante un recorte de las acciones altruistas hacia el individuo tramposo. Esto da sentido al carácter recíproco del altruismo, esto es, frente al sacrificio del altruista, la acción recíproca consiste en retornar la ayuda, mientras que frente a la respuesta no cooperativa, la conducta recíproca en el siguiente encuentro consiste en abstenerse de proporcionar ayuda.

Trivers sostiene que las posibilidades de que se seleccione una conducta altruista son más grandes cuando hay muchas situaciones de altruismo durante la vida del individuo, cuando un altruista interactúa repetidamente con el mismo grupo de individuos y cuando pares de individuos están expuestos a situaciones altruistas de manera simétrica, esto es, cuando en la interacción los individuos se proporcionan

mutuamente beneficios aproximadamente equivalentes y los costos en que incurre cada uno son también aproximadamente equivalentes.

Según Trivers la situación de dos individuos expuestos repetidamente a situaciones de reciprocidad simétrica es conocida en teoría de juegos como el dilema del prisionero. La descripción del dilema del prisionero es muy conocida y es brillantemente ilustrada por el clásico ejemplo de Luce y Raiffa (1957). Dos sospechosos cómplices de un delito se encuentran en custodia y son separados. El fiscal no tiene suficiente evidencia y entonces plantea a cada uno dos alternativas: confesar (delatar al otro) o no confesar (cooperar con el otro o altruismo). Si ambos no confiesan, solamente pueden ser procesados por un delito menor y ambos reciben una pena reducida (1 año). Si ambos confiesan, serán procesados por el delito pero se les reduce un poco la condena (8 años). Si uno confiesa y el otro no, quien confiesa recibe una pena muy reducida (3 meses) y el otro recibe una pena drástica (10 años).

El dilema del prisionero simple se puede ilustrar con la siguiente matriz de pagos:

		Jugador B	
		Cooperar	Desertar
Jugador A	Cooperar	A=3 (R) B=3 (R)	A=0 (S) B=5 (T)
	Desertar	A=5 (T) B=0 (S)	A=1 (P) B=1 (P)

Si ambos jugadores cooperan, reciben un pago de 3. Si ambos defeccionan, cada uno recibe un pago de 1. Si A escoge cooperar y B desertar, A no recibe pago y B recibe 5, resultado que será inverso si es B quien coopera y A defecciona. Si B escoge cooperar la mejor alternativa para A será desertar y si B escoge desertar, ocurre lo

mismo, la mejor alternativa para A será desertar. Por tanto cualquier sea la decisión del otro jugador, la mejor opción consiste en desertar.

Trivers señala que cuando un altruista se encuentra con otro altruista en repetidas ocasiones (n veces), por ejemplo 5, cada uno recibe nR, esto es, 5X3=15. Si un altruista se encuentra con un individuo no altruista, al verse estafado cambia su estrategia a partir del segundo encuentro, esto es, deja de comportarse de manera altruista y, en ese caso, los pagos serían así: el altruista inicial recibe S+(n-1)P y el no altruista T+(n-1)P, esto es, 0+(5-1)1=4 para el primero, y 5+(5-1)1=9 para el segundo. Trivers señala que a menos que haya una diferencia muy grande entre T y R, incluso con pocas interacciones, nR tiende a ser superior a T+(n-1)P como sucede en el ejemplo. Por tanto si los no altruistas son raros, no podrán diseminarse. Trivers señala que existe una barrera para el altruismo cuando los altruistas al comienzo son pocos pues nP>S+(n-1)P, y este es uno de los problemas que enfrenta esta tesis, la explicación del surgimiento del altruismo en un ambiente inicial no cooperativo. Al respecto Trivers considera que esta barrera es débil en un contexto de muchas interacciones pues a medida que aumenta el número de interacciones los valores tienden a la igualdad. Además, la barrera puede superarse si las ganancias obtenidas por los intercambios entre altruistas superan las pérdidas iniciales que se tuvieron con los individuos no altruistas.

Siguiendo en esta misma línea de la tesis de la reciprocidad, posteriormente Axelrod y Hamilton (1981) complementaron el trabajo de Trivers. Axelrod y Hamilton consideraron que persistían algunas dificultades en la explicación de la evolución de la cooperación, entre ellas, la relativa al inicio de la cooperación a partir de un ambiente no cooperativo que ya se mencionó y también la dificultad referente a su mantenimiento estable una vez surgida, dada la posibilidad de éxito de estrategias no cooperativas que terminen llevando al colapso de la cooperación.

Según Axelrod y Hamilton, en muchas situaciones biológicas un par de individuos se pueden encontrar más de una vez y si puede reconocerse y recordar los

resultados de previas interacciones, la situación corresponde al Dilema del Prisionero iterativo, el cual puede dar lugar a múltiples estrategias a seguir por los jugadores. En la búsqueda de una estrategia evolutivamente estable, es decir, aquella que de ser adoptada por una población, no puede ser invadida por otra, Axelrod y Hamilton adelantaron un torneo computarizado en el cual resultó ganadora la estrategia del toma y daca (Tit-For-Tat, en adelante TFT) consistente en cooperar en la primera jugada y en las restantes jugadas simplemente imitar la jugada previa del otro jugador. En la simulación, esta estrategia mostró mayor capacidad para prosperar en un ambiente diversificado (robustez) y condiciones para resistir la invasión de estrategias mutantes (estabilidad).

Sin embargo, TFT no es la única estrategia estable, la estrategia Siempre Desertar (SD) también lo es, entonces se plantea el problema del surgimiento de la tendencia evolutiva hacia la cooperación en un estado inicial de ausencia de cooperación. Dado que SD es también estable, esto significa que en un ambiente de no cooperación dominado por SD, TFT no puede evolucionar. Lo contrario es cierto también cuando TFT es la estrategia predominante en la población. Luego, el problema de explicar la evolución de TFT se reduciría a explicar su propagación en ambientes dominados por la estrategia estable rival SD. Según Axelrod y Hamilton dos mecanismos permiten el nacimiento de la cooperación.

En primer lugar, las estrategias mutantes pueden estar emparentadas. En situaciones como la del Dilema del Prisionero simple, si los jugadores están lo suficientemente emparentados, el altruismo puede evolucionar por cuanto las condiciones de costo, beneficio y parentesco generan ganancias netas para los genes que residen en los individuos emparentados. Por tanto, puede presumirse que serán los grupos conformados por individuos estrechamente emparentados quienes comienzan a cosechar los beneficios derivados de la cooperación. Una vez existan los genes cooperativos, la selección promoverá estrategias que basan la conducta cooperativa en indicaciones del entorno que permitan el reconocimiento del parentesco en jugadores potenciales. Una indicación de parentesco será el hecho de

la reciprocidad, y a la inversa, una respuesta egoísta indicará que el parentesco es bajo o dudoso. Se adquiere entonces la capacidad de condicionar la propia conducta al comportamiento del otro y la cooperación puede expandirse a grados de parentesco cada vez menor. En la fase final del proceso, cuando la probabilidad de que dos individuos vuelvan a encontrarse sea lo suficientemente grande, es de esperarse que prospere la cooperación basada en la reciprocidad y que se torne evolutivamente estable en poblaciones sin parentesco.

En segundo lugar, otro mecanismo que permite iniciar la cooperación en un ambiente no cooperativo es el agrupamiento. Suponiendo que en una población existe un pequeño grupo de cooperadores que emplean la estrategia TFT y el resto son individuos que emplean la estrategia SD. Si la proporción de individuos agrupados es insignificante frente al total de interacciones de SD, un grupo de individuos TFT puede ser viable si se cumplen las condiciones que se explican en Axelrod (1984).

La primera condición consiste en que el peso relativo de la jugada siguiente en comparación con la jugada actual, esto es, la posibilidad de encontrarse nuevamente con el mismo jugador, debe ser suficientemente grande. Aquí se refiere al grado de incertidumbre sobre el futuro, por ejemplo, si es muy probable que el otro jugador abandone el juego al dar muestras de debilidad, el valor de este parámetro disminuye ostensiblemente y resulta rentable la estrategia SD.

La segunda condición consiste en que la proporción de las interacciones de los TFT con otros TFT debe ser lo suficientemente grande. Una comunidad de SD puede resistir la invasión de cualquier otra estrategia si los nuevos individuos llegan de uno en uno. Ello por cuanto el recién llegado que llega solo, no tiene a nadie que le devuelva su posible cooperación. Sin embargo, según los cálculos de Axelrod, si los que llegan lo hacen agrupados es posible que la cooperación comience. Si los TFT tienen suficientes interacciones con sus semejantes TFT, pueden obtener puntajes promedios superiores a los que obtienen lo SD.

Según Axelrod y Hamilton este análisis tiene algunas implicaciones a nivel biológico. La estrategia TFT exige que el individuo no pueda desertar sin que los otros no puedan ejercer una represalia. La retaliación a su turno exige que el desertor no se pierda en la multitud y en los organismos superiores este problema se soluciona mediante su capacidad para reconocer otros individuos. Cuando un organismo no tiene la capacidad de reconocer al individuo con el que ha tenido interacción, como ocurre en organismos inferiores, un mecanismo eficaz consiste en asegurarse que todas sus interacciones sean con el mismo individuo, lo cual se logra al mantener un continuo contacto. Así sucede en la mayoría de casos de mutualismo que se observan en la naturaleza. Otro mecanismo que tiene el mismo efecto consiste en realizar los encuentros en un lugar fijo como ocurre en los mutualismos acuáticos de limpieza. La cooperación recíproca puede ser estable con una gama amplia de individuos si se cuenta con una capacidad alta de discriminación. Tal es el caso de la especie humana que tiene esta capacidad bien desarrollada y que se basa principalmente en el reconocimiento de rostros. En este sentido, por ejemplo, las personas con prosopagnosia tienen lesiones en regiones identificables del cerebro lo que muestra que se trata de una tarea importante a la que se le dedican recursos cerebrales significativos.

Axelrod y Hamilton señalan que además de la capacidad de reconocimiento, la cooperación estable exige la habilidad de supervisar señales que proporcionen información sobre la continuidad de la interacción pues si la posibilidad de encontrarse nuevamente con el mismo individuo cae por debajo de cierto umbral, deja de ser rentable la cooperación. Un posible ejemplo de esta capacidad ocurre a nivel microbiano pues algunas bacterias que son inofensivas en circunstancias normales, se tornan invasivas y peligrosas cuando el huésped está herido o es anciano.

Ahora bien, debe notarse que, en estudios posteriores, se han planteado estrategias distintas pero similares a TFT que pueden ser evolutivamente estables. En ese

sentido los estudios efectuados en la década de los noventa por Nowak y sus colegas (Nowak y Sigmund 1993, Nowak et al. 1995) señalan que la estrategia conocida como Pavlov ha tenido mayor éxito en simulaciones más completas que las realizadas por Axelrod. Esta estrategia consiste en defeccionar la primera jugada y cambiar a cooperar cuando el otro jugador también defeccione. Más adelante se hará referencia breve a los recientes desarrollos de las investigaciones de Nowak y sus colegas.

1.3.3. Reciprocidad indirecta

A pesar de los desarrollos Axelrod y Hamilton, persistieron dificultades en la explicación de la evolución de la cooperación. Continuaba sin resolver el interrogante acerca de aquellas conductas cooperativas hacia individuos extraños con quienes no se tiene una expectativa de cooperación futura. Ello parece ser un tipo de interacción social rutinaria en la especie humana y sugiere la hipótesis de una predisposición psicológica hacia el aprendizaje de patrones de conducta que promueven la cooperación, hipótesis que como se ha dicho parece contradecir la teoría de la evolución.

En este contexto surgió una extensión de la tesis del altruismo recíproco, la propuesta de Alexander (1987), quien planteó el mecanismo de la reciprocidad indirecta, entendida como aquella en la cual el individuo cooperador no recibe un retorno directo por parte del receptor de la cooperación, sino indirectamente por parte de otro individuo del grupo social. Si bien este mecanismo complementa la tesis del altruismo recíproco, debe aclararse que una idea similar ya aparece en el planteamiento de Trivers (1971). En ese aspecto Trivers señala que en casos como el llamado de alerta en ciertas especies de aves se puede establecer una cadena causal que produce beneficio indirecto al, por ejemplo, impedir la obtención por parte del predador de información útil sobre el área en el cual las aves viven, evitando así que el predador se especialice en su caza. En el caso de los seres humanos, Trivers sostiene que, según el modelo del altruismo recíproco, la selección natural puede

favorecer el aprendizaje a partir de las interacciones cooperativas exitosas y fallidas que permitan la formación de expectativas sobre el comportamiento futuro de los demás, a la manera en que opera la reputación, como por ejemplo ayudar a otros a ejercer coerción sobre los tramposos, crear sistemas de altruismo multipersonal entre individuos cooperadores y formular reglas que regulen las interacciones dentro de tales sistemas:

> In the close-knit social groups that humans usually live in, selection should favor more complex interactions than the two-party interactions so far discussed. Specifically, selection may favor learning from the altruistic and cheating experiences of others, helping others coerce cheaters, forming multiparty exchange systems, and formulating rules for regulated exchanges in such multiparty systems (Trivers 1971 p. 52).

De acuerdo a Trivers esto significa que la selección también favorece el aprendizaje acerca de las tendencias altruistas o tramposas de otros individuos y ello tendrá como consecuencia que el individuo prestará atención a las actitudes de los individuos, lo que los demás dicen de ellos y no sólo directamente al comportamiento cooperativo bruto.

Debe notarse que muy pronto la tesis de la reciprocidad indirecta en la versión más refinada de Alexander recibió críticas. En ese sentido Boyd y Richerson (1989) desarrollaron un modelo matemático simple para someter a prueba la posibilidad de cooperación mediante redes de reciprocidad indirecta. En este modelo el individuo que recibe ayuda, en lugar de retornarla al donante, en la siguiente interacción proporciona ayuda a un tercero quien a su vez replica la conducta hacia otro individuo, de manera tal que se forma una cadena de interacciones que en una instancia posterior se traduce en un acto de ayuda hacia el donante inicial. Boyd y Richerson encontraron que es improbable que la cooperación prospere bajo este escenario, a menos que los grupos sean muy pequeños.

La primera modelación computacional exitosa del mecanismo de reciprocidad indirecta fue realizada por Nowak y Sigmund (1998) quienes realizaron simulaciones de la estrategia de puntaje de imagen en la cual cada jugador tiene un puntaje que se incrementa en un punto cuando proporciona ayuda y disminuye también en un punto cuando se niega la asistencia. Los autores encontraron que luego de un número considerable de generaciones todos los jugadores adoptan la misma estrategia. También concluyeron que la cooperación por medio de la reciprocidad indirecta sólo se puede establecer si la posibilidad de saber la imagen de otro jugador excede la relación costo-beneficio del acto altruista.

Nowak y Sigmund consideran que el resultado obtenido por Boyd y Richerson obedeció al esquema que se sometió a prueba, en el cual se asume que el donante recibe un retorno por su ayuda en una ronda posterior de la misma cadena de interacciones sin que se hayan utilizado puntajes de imagen, lo cual con seguridad habría modificado los resultados.

Posteriormente los resultados de Nowak y Sigmund fueron cuestionados por Leimar y Hammerstein (2001). Según los autores, la estrategia de puntaje de imagen solamente prospera bajo condiciones muy exigentes, bien bajo fuertes efectos de deriva genética o bien cuando el costo de la asistencia es muy bajo. Leimar y Hammerstein realizan simulaciones bajo condiciones diferentes y encontraron que una estrategia similar de reciprocidad indirecta, la planteada por Sugden (1986) y que utiliza el concepto de buena reputación (*good standing*), tiene mejor desempeño que la estrategia de puntaje de imagen. En la estrategia de buena reputación un individuo pierde prestigio cuando no ayuda a un receptor de buena reputación pero, a diferencia de las estrategias de puntaje de imagen, no pierde buena reputación cuando se rehúsa a ayudar a un individuo que no tiene buena reputación.

Milinski y sus colegas (2001) condujeron experimentos con la participación de estudiantes universitarios, con el fin de comparar ambas estrategias. Los resultados

favorecieron la estrategia de puntaje de imagen y se encontró que la estrategia de buena reputación puede ser muy exigente en términos de capacidad de memoria.

Sin embargo, los estudios de Panchanathan & Boyd (2003) llevan a una conclusión diferente. Consideran que el modelo de Nowak y Sigmund se basa en el supuesto de que los individuos nunca cometen errores. Es por ello que toman el modelo y le incorporan errores con un parámetro que denota la probabilidad de que por error no se produzca una donación. Bajo este escenario se encontró la estrategia de puntaje de imagen fracasa y, por el contrario, la estrategia de buena reputación resulta evolutivamente estable.

1.4. El problema de la acción colectiva

Conforme se indicó al comienzo, en la explicación acerca del origen y condiciones de estabilidad de la cooperación, la teoría de la evolución enfrenta no solamente el problema del altruismo, sino que también ha encontrado un segundo problema, el problema de la acción colectiva dirigida hacia la provisión de bienes públicos, en la cual la conducta cooperativa del individuo se traduce en beneficio personal pero enfrenta una alternativa de acción que puede representar un mejor resultado.

Se entiende por bien público aquel que está disponible para todos y del cual su uso por una persona no disminuye el uso por otras personas (Hess & Ostrom 2007 p. 351). Bajo esta noción se entiende que los bienes públicos tienen dos características: no-rivalidad y no-exclusión (Cornes & Sandler 1996). No-rivalidad, o indivisibilidad, significa que una unidad del bien puede ser consumida por una persona sin disminuir las oportunidades de consumo de ese bien que estará aun disponible para los demás. La no-exclusión significa que una vez producido, los beneficios del bien están disponibles para todos, sin excluir a ninguna persona.

En 1965 Olson plantea el problema de la acción colectiva al manifestar que a menos de que se trate de un grupo reducido de personas o que exista algún tipo de coerción, los individuos racionales, que persiguen su interés propio según la tradición económica ortodoxa, no van a actuar para conseguir el interés común o del grupo (Olson 1965 p. 2).

En contraste con el problema del altruismo analizado en la sección anterior, aquí el escenario es distinto, el problema se sitúa en un contexto mucho más amplio, ya no se trata solamente de interacciones de dos individuos, sino que se busca explicar la cooperación en contextos multipersonales, incluyendo grandes grupos. Aquí se observa que muchos individuos están enfrentados a la posibilidad de una acción colectiva, y la mejor alternativa para cada individuo, aquella que maximiza su utilidad, consiste en gozar del beneficio del bien público sin asumir el costo, de manera que si todos actúan de la misma manera, el resultado consistirá en que el bien público no podrá producirse. El individuo no cooperador es denominado "gorrón" (en adelante *free rider*), quien al estar expuesto a incentivos para negarse a contribuir, opta por usar el bien sin pagar el costo (en adelante *free riding*).

Si se examina el problema desde el punto de vista de la teoría de la evolución, puede decirse que en principio, la selección natural tiende a favorecer aquellos genes que generan patrones de conducta de *free riding* (Price 2006) pues los individuos portadores de dichos genes tienen mayores ganancias que los individuos que contribuyen a la generación de bienes públicos. Por tanto, la predicción de la teoría consiste en que no podrán surgir mecanismos cognitivos orientados hacia la acción colectiva.

En aparente contradicción con la predicción de la teoría evolucionista, la propensión de los seres humanos hacia la acción colectiva, parece haber evolucionado en lo seres humanos. En materia de bienes públicos se ha realizado una extensa investigación experimental y sus resultados muestran evidencia acerca de la tendencia de los seres humanos hacia la cooperación en ese ámbito. El diseño

experimental básico consiste en reunir un grupo de personas que no se conocen entre si y asignar a cada individuo una suma de dinero inicial; luego se les ofrece la oportunidad de aportar una parte, o toda su asignación, a un fondo común, bajo el compromiso de que los recursos totales que se recauden en el fondo común serán duplicados y la suma final será repartida en partes iguales entre los jugadores. Ledyard (1995) realizó un recuento de esta clase de experimentos y entre las conclusiones obtenidas está que en experimentos de una ronda, así como en las primeras rondas de experimentos de múltiples rondas, se observan contribuciones entre el 30 y el 70% y en ningún caso la cooperación decrece a niveles inferiores al 10%.

Adicionalmente, debe notarse que la cooperación entre los individuos encaminada hacia la provisión de bienes públicos se ha desarrollado ampliamente en la especie humana, no solamente en sociedades industriales sino también en comunidades de cazadores-recolectores, en las cuales pueden verse muchos ejemplos como compartir comida, cooperar en la cacería, así como la defensa y el ataque colectivos (Price et al. 2002). Un extenso estudio liderado por Joseph Henrich que abarcó 15 sociedades pequeñas de los cinco continentes muestra que la contribución media en los experimentos de bienes públicos oscila entre 40 y 60% (Heinrich et al. 2004).

Planteado en los anteriores términos el problema de la acción colectiva desde la perspectiva evolucionista y siguiendo la ruta trazada, la tarea que sigue consiste en examinar las soluciones al mismo basadas en los mecanismos de reciprocidad y castigo, tema que será objeto de las secciones siguientes.

1.4.1. El modelo basado en la reciprocidad y su extensión a interacciones multipersonales

Frente al problema de la acción colectiva el paso natural en este análisis consiste en examinar el resultado de extender a este problema la solución al problema del altruismo basada en el mecanismo de la reciprocidad. En esta tarea una

consideración importante a tener en cuenta consiste en que, como antes si dijo, ahora el problema se ubica en un escenario de muchos individuos, razón por la cual debe examinarse la aplicación del modelo de reciprocidad al contexto de interacciones multipersonales, especialmente en el caso de los grandes grupos.

En esta materia es interesante observar que si bien Trivers proporciona una explicación más o menos detallada del mecanismo de reciprocidad en situaciones diádicas de dilema del prisionero iterado, sostiene también que la reciprocidad puede extenderse a situaciones de *n*-personas (Trivers 1971: 52).

Posteriormente Boyd y Richerson (1988) extendieron el modelo de Axelrod y Hamilton a grupos de interacción repetida en una situación de dilema del prisionero con n personas. El detalle del modelo es complejo pero puede examinarse la idea básica. Los individuos altruistas solamente tienen ganancias cuando se encuentran con otros altruistas y, en los demás casos, las estrategias no altruistas tienen mejor resultado. Es por ello que la única estrategia cooperativa estable es aquella en la cual solamente se coopera cuando todos los demás cooperan, pues de lo contrario, unos pocos desertores podrán obtener mejores resultados y tendrán mayor éxito reproductivo que los cooperadores. Esa clase de estrategias, cuando son raras, solamente pueden prosperar cuando exista la posibilidad de formar grupos de cooperadores y dicha posibilidad disminuye geométricamente a medida en que se incrementa el tamaño del grupo, al menos cuando los grupos se forman de manera aleatoria. En caso de que los grupos no se formen de manera aleatoria como lo sugiere Axelrod (agrupamiento), las condiciones necesarias para el éxito de la estrategia cooperativa son extremadamente exigentes y se requiere un grado demasiado alto de agrupamiento. Boyd y Richerson concluyen que según el análisis efectuado, la reciprocidad es el resultado evolutivamente menos probable a medida que los grupos se tornan más grandes.

No obstante, atrás se explicó que la tesis de la reciprocidad posee una variante importante: el modelo de reciprocidad indirecta. Por consiguiente, resulta también pertinente referirse a sus posibilidades de éxito en escenarios de muchos individuos.

En esa materia los estudios experimentales de Milinski y sus colegas (2002) respaldan la eficacia del mecanismo de la reciprocidad indirecta para explicar la cooperación en contextos de bienes públicos. En los experimentos participaron 116 individuos divididos en 19 grupos de 6 personas. Se alternaron rondas de juegos de bienes públicos con rondas de juegos de reciprocidad indirecta, en dos modalidades. En la primera, 9 grupos jugaban 8 rondas de bienes públicos y luego 8 rondas de reciprocidad indirecta, y en la segunda modalidad, 10 grupos jugaban una ronda de reciprocidad indirecta seguida de una ronda de bienes públicos hasta completar 16 rondas. Bajo la primera modalidad la cooperación declinó en las rondas de bienes públicos y se restableció en las rondas de reciprocidad indirecta, mientras que en la segunda modalidad, la cooperación se mantuvo siempre en niveles altos. Los autores sostienen entonces que estos resultados demuestran que la necesidad de mantener la reputación en la reciprocidad indirecta permite mantener niveles altos de contribución para los bienes públicos.

Sin embargo, estudios recientes conducidos por Suzuki y Akiyama (2005) han arrojado dudas al respecto. Suzuki y Akiyama realizaron un análisis de interacción en situación de dilema del prisionero con n personas incluyendo el efecto de reputación e investigaron la evolución de la reciprocidad indirecta en escenarios de grupos grandes. Los resultados mostraron que la evolución de la reciprocidad indirecta se torna difícil a medida en que se incrementa el tamaño del grupo. El estudio solamente analiza la evolución de la reciprocidad indirecta bajo el sistema de puntaje de imagen y sin incluir el efecto de errores de implementación. Por esa razón los mismos investigadores realizaron un estudio posteriormente (Suzuki y Akiyama 2007) en el cual incluyeron errores de implementación y se analizó el modelo también bajo el concepto de reputación. El nuevo estudio arrojó un resultado similar al anterior: a medida en que se incrementa el tamaño del grupo, la

condición para que la reciprocidad indirecta sea estable se torna más restrictiva en cuanto a la frecuencia inicial de discriminadores requerida para la estabilidad de la sociedad de reciprocidad indirecta.

En sección posterior se hará referencia a recientes investigaciones que proporcionan nuevos elementos en torno a la discusión sobre la viabilidad de la estrategia de reciprocidad en contextos de grupos grandes.

1.4.2. El modelo basado en el castigo

Una línea de investigación ha planteado que la aplicación del castigo a los *free riders* constituye una mejor alternativa de solución al problema de explicar el mantenimiento de condiciones estables de cooperación en contextos multipersonales.

En esta línea de investigación se destaca el trabajo de Boyd y Richerson (1992) quienes utilizaron un modelo teórico de interacción repetida tipo dilema del prisionero con n personas en el cual, luego de cada interacción, los individuos tienen la posibilidad de imponer un castigo a otro miembro del grupo asumiendo un costo por ello. En el ejercicio se asume que el costo por ser castigado es mayor que el costo por cooperar y que ocasionalmente los jugadores cometen errores. Se consideraron inicialmente dos estrategias: los castigadores cooperadores quienes son los que cooperan y castigan a los desertores, y los cooperadores renuentes que son los que defeccionan hasta que son castigados y entonces cooperan pero nunca castigan. Si al comienzo los cooperadores renuentes son comunes y los beneficios de largo plazo por la cooperación son mayores que el costo de castigar, los cooperadores castigadores pueden prosperar, pero ello también beneficia a los cooperadores renuentes quienes comienzan a tener mayores ganancias y eventualmente se llega a un equilibrio en el que coexisten muchos cooperadores renuentes con unos pocos cooperadores castigadores. Al contrario si al comienzo quienes son comunes son los castigadores cooperadores, se favorece la cooperación pues los cooperadores renuentes evitan el

castigo. Teniendo en cuenta que castigar es costoso, este último escenario debe considerar el efecto de la defección de segundo orden, esto es, la renuencia a cooperar en la imposición de castigos. Para ello Boyd y Richerson introducen una nueva estrategia, la del cooperador tolerante que es quien siempre coopera y nunca castiga. En ese caso se llega a varios posibles equilibrios en los cuales coexisten las tres estrategias empleadas, siempre y cuando los costos de ser castigado sean muy superiores a los costos de castigar. De ese modo Boyd y Richerson proponen que el castigo constituye una mejor opción que la reciprocidad para explicar la cooperación en contextos grupos grandes.

En esta materia debe mencionarse el trabajo de Fehr & Gächter (2002) quienes sostienen que los modelos basados en la reciprocidad no son suficientes por cuanto no explican la razón por la cual la cooperación es frecuente entre personas que no están genéticamente relacionadas, en interacciones no repetidas y cuando las ganancias por reputación son pequeñas o ausentes. Para Fehr & Gächter el castigo proporciona una solución a este problema y plantean que si los *free riders* son castigados, la cooperación puede ser rentable. Para examinar si los seres humanos se involucran en castigo altruista, Fehr y Gächter condujeron experimentos en situaciones controladas de bienes públicos en los cuales se buscaba establecer si los individuos tienen la inclinación a castigar a los *free riders* incluso si ello es costoso y no reporta un beneficio material. En los experimentos se utilizaron unidades monetarias, bajo dos clases de condiciones: castigo y no castigo. En los pruebas se observó que 84.3% de los individuos castiga al menos una vez, 34.3% más de 5 veces y 9.3% más de 10. El patrón seguido fue claro: el 74.2% de los actos de castigo se impuso sobre individuos no cooperadores y fue ejecutado por cooperadores. También se encontró que el castigo sobre los no cooperadores incrementó sustancialmente la cantidad invertida en bienes públicos.

Los resultados de los experimentos de Fehr y Gächter son consistentes con la evidencia que han arrojado otros estudios experimentales. Joseph Henrich (2006) lideró un extenso estudio que abarcó 15 sociedades pequeñas de los cinco

continentes, en el cual se encontró que en experimentos de una sola ronda se observa la aplicación del castigo costoso en todos los grupos estudiados. Así mismo, Gürerk y sus colegas (Gürerk Ö., Irlenbusch B. & Rockenbach B. 2006) adelantaron un estudio con el propósito de comparar la ventaja competitiva de la institución sancionatoria frente a las instituciones sin sanción. En este experimento participaron 84 individuos que interactuaron de manera anónima en situaciones de dilema del prisionero con 30 repeticiones. El estudio concluye que la institución sancionatoria es claramente ganadora frente a las instituciones sin sanción.

Múltiples experimentos en los que se emplean juegos de bienes públicos han mostrado que los individuos cooperadores tienden a castigar a los *free riders* y además están dispuestos a asumir con sus propios recursos el costo de castigar (Fehr and Gachter 2000, Masclet et al. 2003, Ostrom et al. 1992).

En otros experimentos se ha analizado más a fondo esta tendencia en los seres humanos y se ha planteado que la selección natural ha favorecido el desarrollo de una adaptación motivacional que genera un sentimiento punitivo hacia los *free riders*. Según Price et al. (2002) esta hipótesis plantea que la función del sentimiento punitivo consiste en motivar acciones que reversan el diferencial adaptativo que tiene un *free rider*. Para someter a prueba esta hipótesis los autores realizaron una encuesta a estudiantes universitarios en la cual planteaban preguntas sobre situaciones hipotéticas de movilización con ocasión de un enfrentamiento militar con otro país. Según Price y sus colegas los resultados de la encuesta apoyan las predicciones de la hipótesis y respaldan la tesis de que existe un conjunto de adaptaciones orientadas a la acción colectiva que incluyen un subsistema motivacional que produce sentimientos punitivos dirigidos específicamente a los *free riders* con el fin de eliminar o reducir sus ventajas adaptativas y sin que tenga una finalidad distinta como, por ejemplo, la optimización de la acción colectiva.

Si bien el mecanismo el castigo parece ser una solución adecuada para la explicación de la cooperación en grupos grandes, esta solución no está exenta de dificultades. El acto de imponer un castigo implica un costo para el individuo que lo impone y por ende existe un incentivo para evitar ese costo, entonces puede decirse que se trata de un castigo altruista pues implica un acto de sacrificio para beneficio de otros individuos. De manera análoga, puede entenderse que el castigo es un tipo de bien público y, por ende, cada individuo obtiene un mejor resultado mediante el *free riding*, que será un *free riding* de segundo orden. Desde el punto de vista de la teoría de la evolución se tiene aquí nuevamente el problema del altruismo. La única diferencia consiste en que ahora, dado que el comportamiento altruista de los castigadores resulta en sí mismo un problema de bienes públicos, se sigue que éste no puede ser explicado por los recursos usuales de selección de parentesco y altruismo recíproco. Según se indicó atrás la teoría de selección de parentesco no explica la cooperación hacia individuos sin parentesco cercano que sería el caso más común en contextos de bienes públicos. A su turno, la tesis del altruismo recíproco se enfoca en el análisis de interacciones diádicas y no es claro que sea aplicable para el caso de grupos grandes. Debe notarse, que bajo tales aproximaciones el altruismo se torna un altruismo aparente pues como lo señala Trivers, estos modelos basados en la selección natural, son modelos diseñados para remover el altruismo del altruismo (Trivers 1971: 35). Por tanto, se plantea la posibilidad de que el castigo altruista se trate de un caso de altruismo no aparente, esto es, verdadero. De allí se derivan dos posibilidades de explicación, por un lado, el castigo altruista es en sí mismo mal-adaptativo, y por otro lado, puede existir otra explicación adaptacionista que aún no ha sido considerada.

1.5. Recapitulación acerca del problema de la cooperación

En las secciones anteriores se han presentado los dos principales problemas que enfrenta la teoría de la evolución en la explicación de la cooperación, los intentos de solución basados en la reciprocidad y el castigo, así como las dificultades que se han

encontrado. Teniendo entonces en consideración los resultados de las distintas aproximaciones y siguiendo el plan trazado en las primeras secciones y en la sección introductoria, se puede ahora revisar la cuestión acerca de las relaciones entre estos dos problemas, con el fin de plantear nuevos elementos conceptuales que tendrán incidencia en la discusión acerca de la eficacia de los mecanismos de reciprocidad y castigo como modelos explicativos de la evolución de la cooperación.

En la sección anterior se mencionó una diferencia importante entre los dos problemas planteados en torno a la cooperación: mientras en el problema del altruismo la discusión está situada en el campo de las interacciones diádicas, en el caso de la acción colectiva, la cuestión involucra muchos individuos. Sin embargo, la diferencia de escala puede no ser criterio para diferenciar entre estos problemas. Para mostrar esto es conveniente advertir cómo, a través de un camino diferente al seguido por la teoría de la evolución, la teoría de los bienes públicos llegó a la formulación de un problema de características similares al problema de la cooperación en teoría de juegos.

En efecto, Hardin (1971, 1982) demuestra que el problema de la acción colectiva puede ser representado como un juego con una estructura similar a la del dilema del prisionero, concretamente el modelo de juego multipersonal (n personas). Antes se explicó el dilema del prisionero como una situación en la cual dos jugadores están expuestos a la posibilidad de cooperación y se indicó que la mejor opción para cada individuo consiste en no cooperar. Ello a pesar de que el resultado, la mutua defección, no es el resultado ideal pues no es el que reporta el mayor beneficio conjuntamente para los dos jugadores. En otras palabras, ese resultado no constituye un óptimo de Pareto en la medida en que es posible mejorar el resultado para los dos jugadores sin desmejorar a ninguno de ellos, lo cual solamente se logra si ambos cooperan. Pero en el caso de los bienes públicos ocurre en realidad lo mismo pero en un escenario que involucra muchos individuos y en dicho escenario, conforme antes se explicó, frente a la posibilidad de una acción colectiva, la mejor alternativa para el individuo consiste en gozar del beneficio del bien público sin

asumir el costo. Tal como ocurre en las interacciones diádicas, aquí la mejor opción consiste en no cooperar, de manera que si todos los involucrados actúan de la misma manera, el resultado consistirá en que el bien público no podrá producirse de manera voluntaria, resultado del todo distante con respecto al óptimo de Pareto.

Se trata entonces del mismo problema pero en contextos diferentes. En ambos casos el problema consiste en explicar la razón por la cual la selección natural favorece conductas cooperativas en situaciones en las cuales el individuo cuenta con alternativas que en principio pueden representar mayores beneficios. En ambos casos el problema consiste en que según la teoría, la conducta cooperativa será el resultado evolutivamente menos probable, incluso a pesar de que dicha cooperación podría generar mayor beneficio para cada uno de los individuos involucrados.

Si bien la contribución voluntaria dirigida a la producción del bien público parece no involucrar el sacrificio propio del altruismo, dado que puede generar un beneficio para el individuo, tiene un carácter altruista consistente en que el individuo incurre en un costo en condiciones en las cuales tiene una mejor alternativa de conducta: el *free riding*. Al cooperar, los individuos proporcionan un beneficio al *free rider* a expensas de su propio beneficio, mientras el *free rider* obtiene un beneficio a expensas de los individuos cooperadores. Al igual que en el caso del altruismo, aquí la conducta parece estar en contradicción con la teoría de la evolución. Para el individuo y, en última instancia, para sus propios genes, la opción que favorece su replicación es el *free riding*, mientras que el acto de contribuir a la provisión del bien público lo coloca en desventaja.

Adicionalmente, la generación de beneficio no constituye una regla general para el individuo que contribuye a la provisión del bien público. Puede ocurrir que las contribuciones sumadas no sean suficientes para que pueda generarse el bien público, lo que supone una pérdida para el individuo cooperador. También puede ocurrir que la viabilidad de la acción colectiva dependa de incrementar sustancialmente la contribución de aquellos individuos cooperantes, por lo que

puede darse el caso en el cual el retorno de la inversión sea inferior al aporte realizado.

En cualquiera de las anteriores hipótesis, se aprecia que el individuo cooperador, incurre en un comportamiento altruista respecto a la alternativa que supone el *free riding*, de manera que si persiste en su decisión de aportar incurre en el mismo sacrificio en que incurre el individuo cooperador en interacciones diádicas. En ambos casos se tiene una opción diferente de mayor rentabilidad y sin las pérdidas que se asumen por la participación de *free riders*.

Ahora bien, aunque el problema de la provisión de bienes públicos es estructuralmente un problema de altruismo, es ampliamente aceptado que las teorías que se han examinado en secciones anteriores, de selección de parentesco y altruismo recíproco, no ofrecen una solución adecuada al problema de la acción colectiva. Como antes se dijo, la tesis de la selección por parentesco no explica los casos de cooperación entre individuos no emparentados y la ausencia de parentesco constituye la regla general en las situaciones de cooperación dirigida a la provisión de bienes públicos. Por otra parte, la tesis del altruismo recíproco desarrolla su análisis a partir de una estructura de interacción diádica y ha recibido objeciones importantes en cuanto a su aplicación a contextos de muchos individuos. Bajo esta consideración, surgió el modelo del castigo que pretende superar las limitaciones del modelo basado en la reciprocidad.

2. CONCEPTUALIZACIÓN DE LOS MECANISMOS EVOLUTIVOS DE RECIPROCIDAD Y CASTIGO

2.1. Aproximaciones acerca de la distinción entre reciprocidad y castigo

En la primera parte de este trabajo ha quedado redefinido el problema que enfrenta la teoría de la evolución en el campo de la cooperación. Conforme se explicó, se identifican dos aproximaciones principales para su solución: por una parte el modelo basado en la reciprocidad, y por otra parte, la tesis del castigo. Se debe entonces iniciar ahora la tarea encaminada hacia el segundo objetivo del presente trabajo, esto es, esclarecer los elementos conceptuales de dichos mecanismos evolutivos.

Aunque es abundante el material escrito sobre el problema de la evolución de la cooperación humana, con frecuencia los investigadores se enfocan principalmente en los parámetros, resultados y análisis de las pruebas experimentales, sin abordar con detalle la conceptualización básica de los mecanismos de reciprocidad y castigo, o haciendo mención a este tema únicamente de manera indirecta.

En este sentido, resultan ilustrativos y especialmente importantes para el desarrollo del análisis propuesto las aproximaciones de algunos autores como Sripada (2005) y Rosas (2008). Si bien estos autores dirigen su atención fundamentalmente hacia otros problemas, en su análisis aportan elementos clave para la discusión que será objeto de las secciones posteriores. Se trata de reflexiones que proporcionan una base importante para el objetivo de este trabajo, razón por la cual en las siguientes secciones, se presenta una breve reconstrucción de los planteamientos de Sripada y Rosas.

2.1.1. El planteamiento de Sripada

En su artículo "Punishment and the strategic structure of moral systems" (2005), Sripada realiza un análisis de las diferencias entre reciprocidad y castigo con el fin de plantear que la explicación del acatamiento moral basada en el castigo constituye una mejor alternativa que la explicación basada en la reciprocidad.

Sripada considera que la existencia de las normas morales representa un verdadero acertijo para la teoría de la evolución. El autor señala que en múltiples contextos de la vida social, las normas morales obligan al individuo a realizar acciones contrarias a su interés personal, lo cual significa que en tales casos los individuos pueden tener un interés egoísta en la violación de las reglas morales, al menos en aquellas situaciones en las cuales dichas reglas obligan a la solidaridad y al sacrificio. Es por ello que surge el interrogante acerca de la estabilidad a largo plazo de las normas morales y su habitual cumplimiento por los individuos, cuestión que se conoce como el problema del acatamiento moral.

Según Sripada los teóricos de la evolución en realidad no han examinado directamente este problema sino que se han enfocado en otro problema relacionado, el problema de la cooperación. Para Sripada el problema de la cooperación (acción colectiva) es un problema distinto respecto del problema del acatamiento moral que es el que constituye objeto de su atención. Sostiene que si bien algunas normas morales proporcionan una solución a problemas de acción colectiva, existen muchas normas morales que no regulan la acción colectiva, por ejemplo aquellas relacionadas con la violencia, el adulterio, el comportamiento sexual, las relaciones de autoridad, los hábitos alimenticios entre otras.

Al margen de la discusión en torno a ese planteamiento, las reflexiones de Sripada son pertinentes para esclarecer los conceptos de reciprocidad y castigo en el ámbito de la cooperación, que es el objeto de análisis aquí, pues como bien lo señala el

propio autor, en el caso de las normas que considera que no se relacionan con la acción colectiva, se enfrenta también el problema de la existencia de incentivos egoístas para su incumplimiento.

Sripada considera que la aproximación basada en la reciprocidad, según se formula a partir de los trabajos de Trivers y Axelrod, no es una solución adecuada frente al problema del acatamiento moral. Uno de los problemas que Sripada advierte en este enfoque es el aumento de escala. En ese sentido considera que en un escenario que involucre muchos jugadores, la reciprocidad no sería un mecanismo de disuasión que opere de manera selectiva, por cuanto la defección que sea respuesta a una defección en una ronda previa, afecta no solamente al jugador que no cooperó inicialmente sino a todos los demás jugadores. Los cooperadores deberán ser entonces muy estrictos, esto es, cada jugador continuará cooperando siempre y cuando todos los demás hayan cooperado en rondas previas, y entonces para que haya reciprocidad sería necesario que el grupo sea muy homogéneo, de manera tal que todos los individuos sean cooperadores, lo cual no es factible en el mundo real. Además, incluso si el grupo fuere homogéneo, señala Sripada, el equilibrio resultante sería muy sensible a errores. Es normal que los jugadores cometan algún error ocasionalmente y en esos casos el modelo predice que la defección por un individuo, aun cuando sea por error, genera la defección de todos los demás, con lo cual la cooperación termina colapsando.

En contraste, según Sripada, la explicación del acatamiento moral basada en el castigo no tiene el problema del aumento de escala. La razón fundamental es que, a diferencia de la reciprocidad, el castigo es un mecanismo que impone selectivamente costos a los individuos no cooperadores, lo que ha permitido explicar, por ejemplo, la cooperación en grupos grandes.

Sripada sostiene que existen otras diferencias entre reciprocidad y castigo, además de la que se acaba de indicar:

1) Según Sripada en el castigo, a diferencia de la reciprocidad, la relación entre la conducta y el castigo es arbitraria. En la reciprocidad la conducta opera bajo el principio "like begets like", que aunque no tiene traducción precisa, significa que una conducta o acción engendra una reacción semejante. En ese orden, la cooperación engendra cooperación y la defección engendra defección. Por el contrario, en el caso del castigo no opera ese principio sino que opera bajo un principio diferente, allí la defección engendra daño.

2) Los modelos basados en la reciprocidad se estructuran a partir del dilema del prisionero simple, mientras que en la tesis basada en el castigo se tiene una segunda fase del dilema del prisionero que no está presente en el dilema del prisionero simple en la cual cada jugador decide si castiga o no al otro jugador, dependiendo de la acción previa del primer jugador.

3) Sripada sostiene que el castigo es una acción costosa para quien lo aplica, mientras que negar reciprocidad frente a la acción del infractor, no es algo costoso para quien lo ejecuta.

Clarificada la distinción Sripada considera que una acción que se denomine "castigo" puede ser objeto de una descripción en términos conductuales. Por un lado, está el castigo que consiste en una directa disminución del resultado para el infractor, el cual puede denominarse como castigo directo. Tal es el caso de la multa, la destrucción de propiedad, etc. Por otro lado, existen otras clases de castigos, por ejemplo aquellos basados en la exclusión y aquellos basados en la reputación. Si bien parecen evocar el modelo basado en la reciprocidad, tienen rasgos que son propios del modelo basado en el castigo: son selectivos, su imposición tiene costo y son arbitrarios respecto de la conducta del infractor.

Sripada considera que la aproximación basada en el castigo, si bien es mejor que la basada en la reciprocidad para explicar el acatamiento moral, genera un nuevo interrogante que ya se mencionó atrás. Teniendo en cuenta que el castigo

invariablemente tiene un costo para quien lo impone, debe explicarse la manera en que se sustenta el castigo costoso. En este sentido, según Sripada, algunos rasgos intrínsecos del castigo permiten reducir su costo. Por un lado está la asimetría, pues se observa que el costo para quien impone el castigo es solamente una pequeña fracción del daño causado para quien recibe el castigo. Por otro lado, debe tenerse en cuenta que el castigo es una estrategia condicional, esto es, solamente se castiga cuando alguien viola las normas morales.

Sripada considera que adicional a esta reducción de costos, existen mecanismos de estabilización del castigo que hacen que su imposición favorezca el interés personal de quien lo impone. Uno de ellos es el castigo de orden más alto, esto es, aquel que se impone a quien infringe la norma que ordena castigar las violaciones de normas morales. La idea central tras esta noción es que la ejecución del castigo es considerada como un deber moral. Por tanto, tal como ocurre cuando se presenta un incumplimiento de cualquier otra norma moral, quien incumple el deber de castigar es susceptible de recibir castigo por ese motivo. Según Sripada dos factores permiten estabilizar el castigo de orden más alto, por un lado la asimetría ya mencionada y, por otro lado, lo escasas que son las oportunidades que se tienen para obtener un beneficio por abstenerse de imponer esa clase de castigo.

Sripada señala que algunos autores han propuesto otros mecanismos de estabilización del castigo. Entre ellos está la tesis de la señal costosa según la cual quien castiga infractores anuncia al resto del grupo que es un individuo cooperador. Otro mecanismo que se ha propuesto es el basado en la transmisión cultural conformista, esto es, la tendencia de los individuos a adoptar variantes culturales basadas en su uso habitual en la población.

2.1.2. Las aproximación de Rosas

En fuerte contraste con la concepción anterior, Rosas (2008) ha cuestionado la actitud teórica basada en la distinción radical entre reciprocidad y castigo. De

acuerdo con Rosas, tanto la reciprocidad como el castigo han sido formuladas en la teoría evolucionista y en la teoría de juegos evolutiva como estrategias, esto es, categorías de conducta. Rosas plantea que la distinción entre dichas estrategias opera sólo a nivel conductual y no puede aplicarse en el plano de los mecanismos proximales de carácter psicológico que soportan tales estrategias.

En este orden de ideas, Rosas plantea una distinción entre una noción estrecha y una noción amplia de reciprocidad. La diferencia entre ambas radica fundamentalmente en el nivel de exigencia para el organismo en términos de los mecanismos cognitivos intervinientes. En la noción estrecha la reciprocidad es cognitivamente poco exigente, mientras que en la noción amplia se requieren mecanismos psicológicos más sofisticados.

Bajo una noción estrecha, Rosas toma el siguiente significado de reciprocidad: responder a una acción de A con una acción idéntica. En ese caso, no se requiere una alta capacidad cognitiva del organismo para su ejecución, sino sólo la capacidad de desplegar una conducta similar como respuesta a las acciones que otro organismo dirige hacia él, sin que sea necesario que el organismo entienda sus acciones en términos de cooperación o reciprocidad. Se trata de copiar lo que el otro organismo ha hecho previamente y, por tanto, opera de manera similar a como lo haría un organismo que sigue una estrategia del tipo TFT a partir de la segunda ronda.

Rosas señala que según la evidencia empírica, la mayor parte de interacciones del tipo TFT, en especies no humanas, ocurre cuando el lapso de tiempo entre acción y reacción es corto y no exige seguir una regla de reciprocidad en el sentido amplio y psicológico del término. Cuando ese lapso es prolongado, las interacciones cooperativas son poco comunes y ello se debe a que aumentan las exigencias de capacidades cognitivas del individuo en cuanto a memoria y reconocimiento individual. En ese tipo de interacciones se requiere un aparato cognitivo más sofisticado que proporcione la capacidad de seguir reglas de acción prudenciales,

como es el caso de la regla de reciprocidad en los seres humanos. Sin embargo, bajo esta noción estrecha de reciprocidad, el organismo no requiere seguir una regla de reciprocidad y, por tanto, dicha noción no puede jugar un papel importante en la explicación de la cooperación humana.

Rosas explora entonces una noción amplia de reciprocidad. Para ello considera necesario incorporar elementos que no se tomaron en cuenta bajo la noción estrecha, concretamente, los mecanismos proximales involucrados en la producción de las conductas de cooperación y deserción. En contraste con la noción estrecha, la noción amplia de reciprocidad se define en función de tales mecanismos, de manera que la distinción entre reciprocidad y castigo se convierte en una distinción que se traza a nivel psicológico. En este sentido, Rosas sostiene que en el caso de los seres humanos existirá una regla en su estructura psicológica, formulada en términos de esas conductas, una norma que prescribe un determinado comportamiento como bueno o malo.

Rosas inicia su cuestionamiento a la distinción radical entre reciprocidad y castigo señalando que reciprocidad y castigo tienen un efecto similar. Una forma común de castigo en la vida social, por ejemplo, es el ostracismo o la exclusión y su efecto es similar al que se produce en el altruismo recíproco cuando en las interacciones diádicas se defecciona frente a la deserción del otro individuo. La similitud en los efectos sugiere, al menos en seres humanos, la existencia de un mecanismo general que impone costos al desertor y a la vez produce ciertas conductas con ese mismo propósito en diferentes circunstancias.

Rosas, siguiendo el planteamiento de otros investigadores entre ellos Sripada y Stich (2005), considera que es plausible una aproximación acerca de las normas, bajo la cual se revela un mecanismo psicológico subyacente que incluye una motivación intrínseca hacia la conducta prescrita en la norma y una disposición a castigar a quienes la incumplan. Según Rosas esta caracterización corresponde a la noción amplia de la norma reciprocidad que se indicó atrás y emerge así la

reciprocidad como una norma que exige el cumplimiento condicionado de las normas. Esta norma tendría la siguiente formulación: Cumplir en respuesta al cumplimiento; no cumplir en respuesta al no cumplimiento. Se trata entonces de una meta-norma pues prescribe el cumplimiento de las normas. En ese orden de ideas, la meta-norma de reciprocidad se generaliza para todos los niveles de interacción no solamente en el caso de las bipersonales sino también en las de muchos individuos. La regla de reciprocidad demanda la retaliación en todos los niveles e involucra terceros. De ese modo, en el caso de las diadas el desertor debe soportar la respuesta retaliatoria del otro individuo mediante la defección y en el caso de la provisión de bienes públicos el *free rider* deberá enfrentar la retaliación que se ejerce mediante el castigo.

Rosas sostiene que, bajo el mismo mecanismo motivacional de la reciprocidad, tanto la retaliación como el castigo directo, se diferencian solamente por ser métodos diferentes de castigo, y la diferencia se origina en consideraciones de eficiencia. Para demostrarlo desarrolla un modelo de juego repetido indefinidamente y compuesto de dos rondas alternativas: la primera consiste en un juego multipersonal de bienes públicos, y la segunda en un juego de castigo bipersonal. El autor plantea dos variantes para la segunda ronda, en la primera, el método de castigo es el castigo directo (TPP), mientras que en la segunda el método es el castigo indirecto con la estructura de dilema del prisionero (TPR). Así mismo, se establecieron cuatro diferentes escenarios, en los cuales la variación se refería al número de jugadores desertores y a la posibilidad que tienen los desertores de explotar a los cooperadores.

Una vez efectuado el cálculo correspondiente, los resultados muestran que TPR es estable mientras los desertores no tengan la posibilidad de explotar a los cooperadores. TPP fluctúa dependiendo del porcentaje de cooperadores. Aunque TPR es un escenario realista, también es frecuente en la vida real que los desertores logren explotar a los cooperadores. Por tanto, los resultados en términos de pagos para cooperadores y desertores, dependerán de las oportunidades de explotación.

Con base en lo anterior Rosas plantea que los individuos estarán motivados para castigar si se les permite alternar el método de castigo según las circunstancias o por consideraciones de eficiencia. Sostiene que esto es consistente con la evidencia experimental, específicamente, las pruebas efectuadas por Milinski y sus colegas (2002) y mencionadas en sección anterior. En dichas pruebas los resultados mostraron que los desertores en juegos de bienes públicos recibían mala reputación y eran castigados con la defección en interacciones bipersonales. Rosas aclara que Milinski interpreta los resultados de otra manera pues considera que negar cooperación a un desertor no es una forma de castigo. Contrario a ello Rosas considera que los resultados indican que los jugadores perciben las rondas bajo la estructura de dilema del prisionero, como oportunidades de castigo.

2.2. La distinción entre reciprocidad y castigo

El examen de los planteamientos de Sripada y Rosas muestra que la cuestión acerca de la conceptualización de los mecanismos de reciprocidad y el castigo es relevante en la discusión acerca de la evolución de la cooperación. Aunque las posturas analizadas no necesariamente se contraponen en todos sus aspectos, persisten en cada una ciertas dificultades que hacen necesario el análisis propuesto como segundo objetivo de este trabajo.

En el análisis de Sripada se concluye que el mecanismo del castigo ofrece una mejor explicación que la tesis de la reciprocidad, por razones que están ligadas a la manera en que se definen conceptualmente estos mecanismos. Específicamente, al describir la reciprocidad como un mecanismo no selectivo, el autor hace inviable el mecanismo como explicación de la cooperación en interacciones multipersonales. Sin embargo, conforme se verá más adelante, el mecanismo de reciprocidad debe ser definido de manera menos exigente con lo cual adquiere mayor alcance en el ámbito de las interacciones que involucran muchos individuos. Lo que aquí debe

enfatizarse es que la propuesta de Sripada muestra que la conceptualización de los mecanismos de reciprocidad y castigo tiene una incidencia determinante en la solución del problema de la cooperación. La posibilidad de postular el castigo como una mejor explicación que la reciprocidad está ligada de manera indisoluble a la manera en que se definen los mecanismos y las diferencias entre uno y otro.

Con respecto al planteamiento de Rosas, un aspecto positivo consiste en que permite apreciar la ambigüedad existente en la distinción entre reciprocidad y castigo como mecanismos explicativos de la cooperación humana. Como señala Rosas, dicha distinción depende esencialmente de la perspectiva de análisis que se emplee. Una perspectiva puramente conductual conduce a una noción estrecha de reciprocidad donde reciprocar es responder con el mismo comportamiento. Dado que el castigo puede ser un tipo de comportamiento cualitativamente distinto de aquel respecto del cual es respuesta, reciprocidad y castigo en sentido estrecho ya no pueden ser por principio equiparadas. Pero esta distinción es susceptible de desaparecer desde una perspectiva psicológica, puesto que, desde el punto de vista psicológico del propio sujeto, ser golpeado o humillado públicamente, por ejemplo, podrían ser formas legítimas de reciprocidad frente a la negativa a compartir la carne producto de la caza. Es por esta razón que la perspectiva psicológica supone en principio una noción más amplia de reciprocidad.

Desde la perspectiva de Rosas, ambas opciones suponen estrategias explicativas distintas. Por un lado, la noción estrecha se encuentra asociada con una estrategia explicativa enfocada directamente en la evolución de las conductas, más que en los mecanismos cognitivos mediante los cuales se realizan. La clasificación de los mecanismos explicativos de la cooperación está entonces asociada a los efectos de dichas conductas sobre la aptitud biológica. Por otro lado, la noción amplia está ligada con una estrategia explicativa en la cual la teoría evolutiva explica la aparición de ciertos mecanismos psicológicos que son la explicación próxima de la conducta cooperativa. De acuerdo a este último enfoque, la clasificación de los mecanismos explicativos de la cooperación, como la reciprocidad y el castigo, se

traza a nivel de los mecanismos psicológicos donde estos podrían resultar indistinguibles.

Sin embargo, algunos aspectos de la propuesta de Rosas requieren revisión. Por un lado, la noción estrecha no parece capturar adecuadamente el funcionamiento del mecanismo de reciprocidad. Resulta insuficiente una descripción conductual de la reciprocidad bajo la cual se reduce dicho mecanismo a una acción de copia pues ello restringe artificialmente el dominio de comportamientos que pueden considerarse recíprocos. La noción estrecha está basada en la estrategia TFT que como antes se dijo, constituye la modelación típica de la estrategia de reciprocidad en teoría de juegos. Sin embargo, no parece viable construir a partir de dicha estrategia una noción conductual de reciprocidad cuya descripción se agote en una simple acción de copia. Como se propone más adelante, lo que resulta esencial a la reciprocidad es el despliegue condicionado de conductas (no necesariamente idénticas) en situaciones de transferencia potencial de beneficios equivalentes. De hecho, si fuese viable una descripción conductual así limitada, no sería posible distinguir la reciprocidad respecto de otras conductas del organismo que no están relacionadas con la cooperación, de modo que cualquier acción de imitación podría ser considerada como reciprocidad en el sentido estrecho.

Por otro lado, la estrategia explicativa asociada con la noción amplia de reciprocidad supone un cambio importante en la manera en que se explica la evolución de la cooperación. El razonamiento que subyace a dicha estrategia consiste en que se busca explicar esa evolución de manera indirecta, esto es, se explica la evolución de los mecanismos psicológicos que intervienen en el proceso causal de la conducta cooperativa del individuo y, de este modo, se explica el surgimiento y estabilidad de la cooperación. Aunque dicha estrategia de explicación es posible, no puede ser adoptada en este trabajo, por cuanto no trata a la reciprocidad ni al castigo como mecanismos evolutivos en sí mismos. En efecto, conforme se analiza más adelante, existen claras diferencias entre reciprocidad y castigo en cuanto a los efectos en aptitud, aspecto que no puede ser excluido de una

explicación darwinista de la cooperación. En la medida en que la selección natural explica la evolución apelando a las diferencias de aptitud entre los organismos, los mecanismos de reciprocidad y castigo deberían ser clasificados en función de los efectos que dichos mecanismos tienen sobre la aptitud (y no, por ejemplo, en función de que dichos comportamientos sean explicados o no por el mismo mecanismo psicológico).

Dado que el reto explicativo para la teoría evolucionista consiste en determinar el mecanismo que ofrece una mejor solución al problema de la evolución de la cooperación, los mecanismos de reciprocidad y castigo deben ser clasificados y analizados en principio como mecanismos evolutivos. Hacia esta cuestión se dirigen las secciones finales. En este sentido, lo que se propone a continuación es una clasificación darwiniana alternativa de dichos mecanismos, que se aparta tanto del análisis de Sripada como del propuesto por Rosas, y que a su vez implica algunas consecuencias importantes en la evaluación de los méritos relativos de ambos mecanismos como explicaciones evolutivas de la cooperación.

2.2.1. La reciprocidad en interacciones bipersonales

De manera general, 'reciprocidad' puede definirse como el intercambio condicionado de beneficios equivalentes en aptitud. Esta definición general puede ser aplicada tanto en casos de interacciones bipersonales como multipersonales. En esta sección se analizará sólo el primer tipo de escenario, para luego analizar el caso de interacciones multipersonales aplicando la definición propuesta.

Conforme se indicó anteriormente, en interacciones bipersonales el modelo típico de interacción recíproca presenta una estructura en la cual los individuos siguen la estrategia TFT que consiste en cooperar en la primera jugada y a partir de allí simplemente ejecutar el mismo movimiento del otro jugador. La principal característica de la estrategia TFT consiste en que es condicional, o en palabras de Sripada (2005), la cooperación es contingente a la cooperación del otro jugador en la

ronda previa. El carácter condicional de la estrategia TFT también puede verse desde la primera jugada y puede decirse que en la reciprocidad se proporciona un beneficio condicionado a que el receptor proporcione un beneficio en retorno (Cosmides & Tooby 2005 p. 594) y en caso de no recibirse ese retorno en el siguiente encuentro, allí la respuesta no será la cooperación. Es posible reciprocar no solamente en respuesta a una ayuda previa recibida sino también bajo la condición de que se reciba cooperación en la siguiente jugada, pues de lo contrario no se proporciona ayuda. En cualquiera de estos dos casos, el requisito de condicionalidad se cumple, al margen de los mecanismos cognitivos que sirven de explicación próxima.

Según lo anterior, la reciprocidad no es solamente una acción de copia de la conducta del otro individuo, no es una mera imitación, es una replicación de un tipo especial, es condicional respecto de la conducta cooperativa del otro individuo. Esto significa que un individuo reciproca cuando genera un beneficio como respuesta a un beneficio esperado o conferido previamente, independientemente de que el comportamiento sea o no idéntico. En este sentido, contrario a lo señalado por Sripada, la relación entre la conducta desplegada en la primera jugada y la siguiente es tan arbitraria como en el caso del castigo. Lo importante es que ambas conductas sean comportamientos donde se confieran beneficios equivalentes.

Otra característica importante de la reciprocidad está relacionada con los costos y beneficios en aptitud derivados de la estrategia. Para entender esta característica podemos examinar un modelo sencillo de interacción recíproca como TFT. Bajo esta estrategia en caso de deserción del otro jugador, el individuo se abstiene de realizar la inversión y de ese modo minimiza las pérdidas potenciales de interacciones con desertores, entonces el riesgo se reduce a la primera jugada. Si el otro jugador decide cooperar, tendrá la ganancia propia de la cooperación que a su vez estará determinada por la diferencia entre costo y beneficio. El individuo que emplee la estrategia TFT obtiene las ganancias propias de sus interacciones con cooperadores y reduce al mínimo las pérdidas por encuentros con desertores.

Lo importante aquí es que en una estrategia como TFT no incorpora costos adicionales a la inversión que debe ser efectuada ni tampoco modifica los beneficios derivados de la cooperación. El principal beneficio es el que obtienen los individuos como resultado de la cooperación, que como antes se dijo es un beneficio neto, se obtiene un ingreso mayor que el costo de la inversión realizada. El costo estará determinado por la inversión que debe realizarse para cooperar, o en caso de deserción, el costo de la retaliación pues si el individuo defecciona la respuesta es inmediata, el otro individuo reacciona en la siguiente jugada con la defección. Cualquiera de los dos individuos tendrá en algún momento la opción a desertar pero, en ese caso, se asume el costo consistente en la pérdida de la ganancia derivada de la futura instancia de cooperación.

Lo anterior implica una diferencia importante entre reciprocidad y castigo pues en este último se modifican sustancialmente los costos para los jugadores. Si se adiciona el castigo en la modelación se genera una reducción de las ganancias del desertor. Adicionalmente quien impone el castigo incurre en un costo adicional a los costos asociados a la cooperación. Por tanto, a diferencia de la reciprocidad, en el castigo se incorporan en la función de pagos costos adicionales no asociados con los costos propios de la interacción cooperativa. Además, frente al nivel inicial de dotación de recursos con que cuentan los jugadores, en la reciprocidad el desertor puede obtener ganancias sin realizar ninguna inversión, mientras que con la imposición del castigo, el desertor debe deducir a sus ganancias el costo que significa el castigo, y es por ello que puede decirse que se pierde esa ganancia que obtendría con su negativa a cooperar.

Pese a esta limitación, según las simulaciones computacionales y demás trabajos teóricos reseñados en sección anterior, bajo un ambiente no cooperativo la estrategia tipo TFT puede surgir y mantenerse estable en contextos de interacciones bipersonales. Se puede decir entonces que la reciprocidad se caracteriza por tener un esquema de pagos que en principio hace posible la

cooperación en interacciones diádicas. Por tanto es presumible que en el entorno ancestral la selección natural haya favorecido la propensión de los seres humanos hacia patrones de conducta de reciprocidad en interacciones diádicas.

La propensión hacia la reciprocidad en interacciones diádicas está presente en todas las culturas, por lo que es considerada como un universal transcultural (Kurzban y DeScioli 2008). Esta propensión ha sido constatada por la evidencia experimental. Smith y sus colegas (Smith 2004, McCabe, Rigdon y Smith 2003) adelantaron pruebas en las cuales se utilizó el denominado "Juego de Confianza" que es un juego secuencial de dos personas, con el fin de someter a prueba la hipótesis de que los individuos tienden a reciprocar, esto es, a devolver el favor que otros le han hecho. Participaron 54 estudiantes organizados en 27 parejas. Los individuos no se conocían entre sí y a cada uno se le asignaba alguna de las dos posiciones, Jugador 1 o Jugador 2. La persona 1 debe empezar y tiene dos alternativas: la primera consiste en una distribución $20 para sí mismo y $ 20 para el jugador 2 y la segunda consiste en trasladar la decisión al Jugador 2. Si no traslada la decisión el juego termina y cada uno se lleva $20. Si traslada la decisión, el Jugador 2 tiene dos alternativas: $25 para cada uno o $15 para el Jugador 1 y $30 para el Jugador 2. Hasta aquí la primera modalidad del ejercicio, la modalidad voluntaria. Luego se realizó un nuevo ejercicio, la modalidad involuntaria en la cual se repite el juego con una variante, el Jugador 1 ya no tiene dos alternativas sino solamente una, debe trasladar la decisión al jugador 2. Una vez trasladada la decisión, el Jugador 2 tiene dos alternativas: $25 para cada uno o $15 para el Jugador 1 y $30 para el Jugador 2. En la modalidad voluntaria, 63% de los jugadores 1 trasladaron la decisión al Jugador 2 y de éstos 65% eligió la repartir $25, $25 y 35% eligió $15, y $30. En la modalidad involuntaria, solamente el 33% eligió repartir en forma igualitaria. Según lo anota Smith, en la modalidad voluntaria el Jugador 2 ve que el Jugador 1 ha realizado una acción que lo beneficia y en su turno puede devolver el favor. En el caso de la modalidad involuntaria no es viable esta interpretación.

2.2.2. La reciprocidad en interacciones multipersonales

En la sección anterior se presentaron los rasgos básicos de la reciprocidad en interacciones diádicas. A continuación se examina la cuestión en el caso de las interacciones multipersonales de bienes públicos, esto es, de tres o más individuos, donde la reciprocidad adquiere una complejidad mayor. En interacciones multipersonales la deserción individual en respuesta a la deserción de cualquier otro individuo, lleva indefectiblemente al colapso de la cooperación. Este es el problema que, según Sripada (2005), tiene la reciprocidad como explicación de la cooperación en los seres humanos y que denomina el problema del aumento de escala. Conforme se indicó anteriormente, Sripada considera que en un contexto más amplio de acciones colectivas que involucre a muchos jugadores, la reciprocidad no sería un mecanismo de disuasión que opere de manera selectiva, por cuanto la defección que sea respuesta a una defección en una ronda previa, afecta no solamente al jugador que no cooperó inicialmente sino también a todos los demás jugadores. Los cooperadores deberán entonces ser muy estrictos, esto es, cada jugador continuará cooperando siempre y cuando todos los demás hayan cooperado en rondas previas, lo cual hace de la cooperación recíproca un escenario poco probable.

Este análisis de Sripada encuentra aparente respaldo en la evidencia experimental. Según las conclusiones de los experimentos reseñados por Dawes y Thaler (1988), en escenarios de juegos múltiples se observa cooperación en las rondas iniciales y luego de unas cuantas repeticiones, la cooperación declina en forma significativa. Sin embargo, el planteamiento de Sripada se apoya en una particular conceptualización del funcionamiento del mecanismo de reciprocidad que sólo puede ser aplicable al caso de interacciones bipersonales. La reciprocidad no puede ser modelada en interacciones multipersonales de la misma manera que en el caso de las interacciones diádicas sino que debe tomar en consideración las características del entorno multipersonal. En ese ámbito varios individuos deben efectuar un aporte y, si todos participan, se obtiene el beneficio máximo, pero si muy pocos

participan, la cooperación naufraga. Así mismo, es posible obtener un resultado cooperativo incluso en presencia de *free riders*, por ejemplo, cuando solamente una proporción mínima de los individuos opta por no contribuir. Puede decirse entonces que en interacciones multipersonales puede haber cooperación sin que haya una conducta cooperativa por parte de la totalidad de miembros del grupo. Esto permite plantear, contrario a lo que sostiene Sripada, la viabilidad de un mecanismo de reciprocidad que no exija la cooperación por parte de todos y cada uno de los individuos, dado que ello constituye una restricción innecesaria de cara a la consecución de un bien público.

Ciertamente, la reciprocidad es una estrategia condicional, pero no se reduce a una estrategia condicional como TFT, pues ésta, aunque constituye el caso típico de reciprocidad, es una estrategia en interacciones diádicas. La reciprocidad en interacciones multipersonales debe entenderse como una estrategia en la cual el individuo contribuye de manera condicionada a la conducta cooperativa por parte de los demás individuos. Allí existen dos opciones, la primera, que el aporte esté condicionado a que todos contribuyan, caso en el cual la cooperación colapsa solamente con la deserción de un individuo (esta, por ejemplo, es la opción adoptada por Sripada). Una segunda posibilidad consiste en que se coopere bajo la condición de que un número suficiente de individuos aporte en función de la consecución del bien público. Esta segunda opción es consistente con el requisito de condicionalidad descrito anteriormente, pues lo que se requiere es simplemente producir un beneficio bajo la condición de esperar o haber recibido un beneficio equivalente. Ciertamente, en este caso puede producirse el resultado cooperativo, esto es, con beneficio neto para todos los aportantes, aun cuando el beneficio neto para el *free rider* es mayor que para los demás partícipes. Pero este tipo de situaciones puede darse en contextos bipersonales, por ejemplo, cabe la posibilidad de recibir retorno por la cooperación que no sea exactamente equivalente a la ayuda proporcionada, puede ser inferior o superior en una proporción que no sea significativa y que no lleve al fracaso de la cooperación.

La caracterización de la reciprocidad en interacciones multipersonales descrita en estos últimos términos es consistente con la evidencia experimental en la materia. Conforme lo señalan Johnson y sus colegas (Johnson et al. 2008), en los experimentos de juegos de bienes públicos se ha encontrado que la mayoría de los individuos son cooperadores condicionales, esto es, cooperan más si perciben que los demás jugadores cooperan y cooperan menos si perciben *free riding* en los demás jugadores (Fischbacher et.al. 2001, Kurzban & Houser 2005, Kurzban et.al. 2001). Según Johnson y sus colegas la evidencia experimental sugiere que entre la mitad y dos tercios de los individuos en juegos de bienes públicos adoptan una estrategia de reciprocidad cuando sus contribuciones tienen una correlación positiva con el promedio observado o esperado de las contribuciones de los demás jugadores. Johnson y sus colegas agregan que según el estudio de Croson (1999) las contribuciones de los jugadores se podrían predecir mejor por sus expectativas acerca de la contribución media de los demás jugadores.

Según lo anterior, en contextos multipersonales la condicionalidad de la reciprocidad debe ser entendida en el sentido de que la contribución del individuo está condicionada a que la cooperación por parte de los demás individuos del grupo sea proporcional a la contribución efectuada. Aunque el mecanismo de reciprocidad entendido de esa manera tenga entonces mayor aplicación en interacciones multipersonales, la investigación realizada en torno a la reciprocidad como explicación de la cooperación en interacciones diádicas, no puede ser extendida de manera automática a contextos mutipersonales. La modelación teórica de Boyd y Richerson (1988) mostró que a medida que aumenta el tamaño del grupo social, se disminuye la posibilidad de que la selección favorezca las estrategias de reciprocidad. Sin embargo, debe notarse que en la modelación se utilizó una noción de reciprocidad donde se coopera bajo la condición de que todos cooperen, supuesto que ha sido alterado en una modelación reciente que recoge la noción de reciprocidad condicionada que se ha explicado en esta sección. Takesawa y Price (2010) adelantaron una revisión del ejercicio Boyd y Richerson y sustituyeron el supuesto de Boyd y Richerson al que denominan estrategia de reciprocidad discreta,

por la que denominan estrategia de reciprocidad continua que consiste en mantener la cooperación en niveles que se acercaran a la media de las contribuciones de los demás individuos. Como resultado de la modelación se encontró que la estrategia de reciprocidad continua puede evolucionar en grandes grupos pero solamente bajo condiciones de alta eficiencia productiva. De todas maneras resultó ser una estrategia con superiores resultados a los arrojados por la estrategia discreta.

A pesar de que la modelación teórica no proporciona una evidencia sólida acerca de la eficacia del mecanismo de reciprocidad continua, lo que debe enfatizarse aquí es que resulta factible la conceptualización alternativa del mecanismo de reciprocidad que se plantea en este trabajo. Los resultados de la modelación teórica y el ejercicio como tal, permiten plantear que la reciprocidad, con su rasgo básico consistente en el carácter condicionado puede ser entendida en la forma amplia que se ha descrito y ello tiene como consecuencia que se fortalece su relevancia en la explicación acerca del origen y mantenimiento de la cooperación en entornos de grupos grandes. El reexamen que se pretende aquí en torno a la noción de reciprocidad se encamina en la misma dirección de las investigaciones acerca de la estrategia de reciprocidad continua, que puede proporcionar un nuevo camino a seguir en la investigación acerca de la evolución de la cooperación. Es importante observar que se trata de una línea de investigación muy reciente que puede tener desarrollos en el futuro con variantes que puedan arrojar elementos nuevos a su favor.

Adicionalmente, la aproximación propuesta es consistente con el modelo conceptual que subyace a dos planteamientos que de manera independiente apoyan la viabilidad de las estrategias de reciprocidad como mecanismo explicativo de la cooperación en grupos grandes. En primer lugar, algunos como por ejemplo Tooby y sus colegas (Tooby et al. 2006, Price en prensa), plantean la hipótesis según la cual la reciprocidad directa evolucionó en la especie humana en contextos de grupos pequeños que eran la regla general en entornos ancestrales, pero continúa siendo empleada en las interacciones que ocurren en los grupos más grandes de las sociedades modernas. Bajo esta perspectiva Tooby y sus colegas plantean como

posibilidad que la reciprocidad haya evolucionado de manera simultánea y en sinergia con otros mecanismos como la reciprocidad indirecta, el agrupamiento y el castigo. En tiempos ancestrales, señalan los autores, muy probablemente las oportunidades para la cooperación más frecuentes se daban en diadas, eran menos frecuentes en triadas o tétradas y eran realmente raras en grupos de veinte personas. Tooby y sus colegas sostienen que incluso en tiempos actuales, en las sociedades grandes la gran mayoría de las interacciones cooperativas involucran un número limitado de individuos.

La evidencia de la vida cotidiana es clara en señalar que la mayor parte de interacciones cooperativas se desarrollan con pocas personas y de manera repetida durante períodos prolongados, con la mediación de un previo conocimiento acerca del historial de interacciones previas de los individuos involucrados. Conforme lo anotan Johnson y sus colegas (2008), cuando las interacciones son aleatorias puede desaparecer la reciprocidad según el modelo de Boyd y Richerson (1988), pero si los individuos pueden escoger las personas con quienes interactúan, entonces la reciprocidad puede sobrevivir en grupos grandes. Esto último aporta un elemento clave en la reflexión acerca de la modelación desarrollada por Takesawa y Price (2010). Es posible afirmar que el carácter aleatorio de las interacciones también afecta los resultados de dicha modelación, de manera que sin esa restricción del modelo, es factible obtener un resultado muy favorable al mecanismo de reciprocidad continua.

En segundo lugar se han desarrollado nuevos modelos evolutivos que tienen variaciones respecto del modelo desarrollado por Boyd y Richerson (1988) el cual es un modelo determinista con población infinita. Recientemente Nowak y sus colegas (2004) desarrollaron un modelo estocástico de dos jugadores para especificar las condiciones requeridas para que la selección natural favorezca la cooperación en poblaciones finitas. El estudio indicó que una estrategia TFT puede fijarse en una población de SD si la aptitud de TFT es mayor que la de SD cuando TFT tiene una frecuencia de 1/3 que fue denominada la “regla del tercio”. En reciente estudio de

Imhof y Nowak (2010) se introdujeron algunas variaciones al modelo y se encontró que las estrategias más exitosas fueron SD y la denominada TFT generosa (GTFT) que es aquella que coopera cuando la otra persona coopera pero ocasionalmente coopera cuando la otra persona ha defeccionado. Según ello consideran que la evolución de la cooperación tiene un carácter oscilatorio en el cual se presentan ciclos interminables entre los diferentes tipos de estrategias.

Kurokawa e Ihara (2009) extendieron el modelo de Nowak a un juego con n jugadores e investigaron el efecto del tamaño del grupo. Concluyeron que la selección natural puede favorecer que TFT reemplace a SD y contrario a lo esperado, el aumento del tamaño del grupo facilita la evolución de la cooperación bajo ciertas condiciones. De acuerdo a lo anterior, es presumible que si se utiliza el modelo estocástico de Nowak junto con la estrategia de reciprocidad continua de Takesawa y Price (2010), el resultado sería favorable al surgimiento y mantenimiento de la cooperación en grupos grandes en la medida en que dicha estrategia es menos exigente que la estrategia de reciprocidad discreta.

Ahora bien, al igual que en el caso de las interacciones diádicas, en contextos multipersonales el carácter condicional respecto de la cooperación de los demás no agota la caracterización de la estructura estratégica de la reciprocidad. Según se explicó atrás, la reciprocidad tiene una característica importante consistente en que los pagos no tienen variaciones que modifiquen los resultados asociados con los beneficios de la interacción cooperativa. Pues bien, no ofrece discusión afirmar que este rasgo se mantiene en el caso de las interacciones multipersonales y en dicho escenario es viable el mismo contraste entre reciprocidad y castigo que se indicó para el caso de las interacciones diádicas. Allí sucede lo mismo, la imposición de un castigo genera una reducción de las ganancias del desertor, lo cual no ocurre en el caso de la reciprocidad. La diferencia en el análisis sería sólo en cuanto a la magnitud de los costos asociados a la imposición del castigo. Se dijo anteriormente que en las interacciones diádicas quien impone el castigo incurre en un costo adicional que no se genera en el caso de la reciprocidad; este costo adicional, si bien

se presenta en el caso de las interacciones multipersonales, su magnitud se reduce sustancialmente por cuanto no debe ser asumido en su totalidad por un individuo sino por todos los cooperadores. Se distribuye entonces el costo y el valor que asume cada individuo se torna relativamente bajo, lo cual favorece el margen de ganancia.

Por otra parte, el incremento del margen de ganancia podría tener un impacto importante en la evolución de la cooperación por reciprocidad en contextos de bienes públicos. Conforme se indicó con anterioridad, en esa materia se ha realizado una extensa investigación experimental y se han utilizado pruebas en las que los individuos tienen la oportunidad de aportar a un fondo común cuyos recaudos son posteriormente duplicados y repartidos entre los individuos. Ledyard (1995) realizó un estudio comparativo de los experimentos realizados hasta ese momento para determinar los aspectos uniformes que se observan en los resultados obtenidos. Si bien las contribuciones nunca llegan a cifras inferiores al 10% y en general no se cumplen las predicciones de la teoría de juegos, se aprecia en los experimentos que las contribuciones son sensibles a los rendimientos marginales del individuo y cerca del 50% de los individuos responden a incentivos egoístas una vez entienden el juego. También se encontró que un nivel bajo de rendimientos marginales siempre causa una reducción en la tasa de las contribuciones y paralelamente un nivel alto de rendimientos marginales genera un incremento en la tasa de las contribuciones.

Los experimentos de bienes públicos con grupos grandes son escasos debido a que resultan bastante costosos al requerir sumas importantes para motivar a los participantes a tomar el experimento con seriedad (Cornes & Sandler 1996 p 511). Sin embargo, en esta materia un ejemplo son los estudios de Isaac y Walker (1988, 1994) en los cuales se encontró que el monto del retorno marginal individual tiene incidencia sobre los niveles de eficiencia en la provisión de bienes públicos y en ese sentido los resultados son consistentes con lo que se ha encontrado en tratándose de grupos pequeños. En cuanto a la incidencia del tamaño del grupo se encontró que en el rango de retorno marginal entre 0.30 y 0.75 se logra mayor eficiencia en grupos

grandes, lo cual sugiere que en ese contexto la conducta es influenciada por una sutil interacción entre el tamaño del grupo y el rendimiento marginal individual.

2.2.3. El castigo como instrumento para la cooperación

De manera similar a lo que ocurre en el caso de la reciprocidad, establecer una definición de castigo tiene bastante complejidad. En biología evolutiva suele entenderse el castigo como la respuesta a acciones que reducen la aptitud biológica de un individuo mediante conductas que reducen la aptitud biológica del instigador y que lo desestimula o le impide repetir la acción inicial (Clutton-Brock & Parker 1995). Se puede especificar un poco más el concepto para adecuarlo al contexto de la cooperación que es el objeto de interés aquí. Puede decirse entonces que el castigo constituye una respuesta a una conducta de un individuo que se abstiene de cooperar en interacciones sociales, mediante la imposición de un costo que reduce sus ganancias.

A partir de esta descripción general se pueden entonces especificar los rasgos básicos de este mecanismo que lo contraponen a la reciprocidad. En primer lugar, el castigo se caracteriza por ser una reacción frente a una conducta previa que lo desencadena. Aparentemente se encuentra de nuevo aquí el carácter condicionado característico de la reciprocidad que atrás se explicó. Podría decirse que la imposición del castigo está condicionada a una conducta previa de otro individuo y, por tanto, ello no permitiría una distinción frente a la reciprocidad. Sin embargo, en el caso del castigo este carácter condicionado opera de manera diferente a como opera en la reciprocidad. En la reciprocidad la cooperación está condicionada a la cooperación del otro u otros individuos. En el caso del castigo, su imposición está condicionada a la defección previa del individuo.

Pese a lo anterior, es aún posible imaginar estrategias mixtas de reciprocidad y castigo. Tal sería el caso de una estrategia que consista en cooperar bajo la condición de un retorno y, en caso de no recibirse ese retorno, el individuo que

incumple es castigado. Allí opera el castigo como instrumento o si se quiere, complemento de la reciprocidad y es por ello que para algunos la secuencia de acciones adquiere la denominación de reciprocidad negativa, según sugerencia de Bergmüller et al. (2007), por oposición a la reciprocidad positiva que correspondería más a la noción de reciprocidad que se analizó con anterioridad. Según Bergmüller y sus colegas se denomina reciprocidad negativa aquella encaminada a la inversión por un actor que con ello evita una respuesta negativa (el castigo) a cualquier falta de inversión (hacer trampa).

Al margen de las diferencias terminológicas, lo que se enfatiza aquí es que si bien reciprocidad y castigo son mecanismos diferentes en cuanto al tipo de condicionalidad, de allí no se sigue que sean incompatibles como mecanismos explicativos de la evolución de la cooperación. Por el contrario, cabe la posibilidad de que, en la historia evolutiva de la cooperación, estos mecanismos hayan coexistido y hayan operado en sinergia en la evolución de los mecanismos psicológicos asociados con la conducta cooperativa.

Ahora bien, al igual que en el caso de la reciprocidad, el carácter condicional del castigo no agota la caracterización de su estructura estratégica. El castigo puede tener una amplia variedad de formas, por ejemplo, la amonestación, la multa, la exclusión, el ostracismo, etc. En todos los casos el individuo castigado enfrenta un resultado que reduce su aptitud biológica. En otras palabras, se modifican los beneficios derivados de la interacción cooperativa mediante una deducción a los pagos a favor del desertor. Según se indicó atrás, aquí se encuentra una diferencia clave frente a la reciprocidad y consiste en que los resultados en términos de pagos tienen una variación sustancial. En términos biológicos se puede decir que el castigo elimina la ventaja adaptativa que tiene el tramposo o el *free rider*. Por tanto, puede decirse que el costo que se impone al individuo es un aspecto clave en la estructura estratégica de las interacciones mediadas por el castigo.

Paralelamente a los costos impuestos al desertor, otro efecto del castigo consiste en que normalmente implica costos para quien lo impone. Conforme se indicó atrás, Sripada señala que el castigo es una acción costosa para quien lo aplica, mientras que negar reciprocidad frente a la acción del infractor, no es algo costoso para quien lo ejecuta. En ese aspecto es importante una precisión. Por regla general el castigo tiene un costo, pero esto no es necesario pues formas de castigo no costosas son siempre posibles. Un castigo puede incluso beneficiar directamente al castigador. Esto ocurre, por ejemplo, cuando el castigo constituye la confiscación de un bien para usufructo del castigador, lo cual resultaría equivalente a una modificación experimental en la que la sanción monetaria en contexto de juegos de bienes públicos entrara directamente a engrosar los dividendos de los castigadores.

En las líneas precedentes se ha visto el contraste entre reciprocidad y castigo en sus rasgos básicos. Sin embargo, es pertinente una breve referencia a otros aspectos que según Sripada (2005) constituyen también una diferencia entre estos mecanismos. Según se mencionó en sección anterior, para Sripada en el castigo el daño aplicado al infractor es selectivo, mientras que en la reciprocidad no se da este carácter selectivo pues la retaliación no se dirige exclusivamente en contra del desertor. Siguiendo la caracterización de la reciprocidad en entornos multipersonales que antes se explicó, puede decirse que si un individuo decide no seguir cooperando como respuesta a un nivel promedio muy bajo de las contribuciones de los demás, es claro que su acción afecta no solamente a los *free riders* sino también a aquellos que hayan realizado aportes. Esa respuesta del individuo puede llevar a generar pérdidas para aquellos que continúan contribuyendo. El castigo por el contrario no tiene este efecto, pues al estar dirigido exclusivamente sobre el *free rider*, simplemente le impone un costo adicional que compensa la ventaja obtenida sin afectar a los individuos que cooperan.

Aunque lo anterior parece ser coherente con el criterio de Sripada, si se examina el punto en detalle puede verse que el carácter selectivo del castigo no proporciona un elemento adicional a la distinción. Por un lado, esta característica solamente puede

aplicarse en contextos de interacciones multipersonales y no puede afirmarse en casos de interacciones diádicas, en las cuales, por definición, la reciprocidad opera de manera selectiva, en la medida en que la respuesta solamente se dirige al infractor sin que afecte a otros individuos. Por otro lado, se trata de un rasgo inherente a lo que se ha explicado anteriormente, en el sentido de que el castigo impone costos adicionales al desertor. Es precisamente en ese sentido en el que puede decirse que es selectivo, pero más allá de ello, pierde relevancia como aspecto diferenciador.

Ahora bien, otro aspecto que Sripada señala como una importante diferencia entre los modelos de reciprocidad y castigo, consiste en que, en el caso del castigo la relación entre la conducta y el castigo puede ser bastante arbitraria, lo cual no ocurre en el caso de la reciprocidad. Según se indicó atrás, Sripada considera que en el castigo, a diferencia de la reciprocidad, la defección no genera otra defección sino que produce un daño. Pero aunque esta afirmación sea correcta, tal planteamiento es esencialmente distinto a afirmar que la relación conducta y castigo es siempre arbitraria. Por un lado, y según se mencionó atrás, la reciprocidad no se reduce solamente una acción de copia de la conducta del otro individuo, se trata de una conducta en la cual lo clave es la trasferencia de beneficios y por tanto la relación entre las conductas de los individuos que interactúan puede ser tan arbitraria como en el caso del castigo. Lo importante es que en ambas conductas se confieran mutuamente beneficios que son equivalentes por sus efectos en la aptitud de los individuos. Por otro lado, el carácter arbitrario entre conducta y castigo, tampoco parece ser la regla general pues existen muchas excepciones. Un ejemplo familiar puede verse en la denominada ley del talión que tuvo mayor auge en el derecho antiguo, como puede verse en algunas prescripciones de la Biblia (Éxodo 21, 23; Levítico 24:19–21), el Código de Hammurabi (¶197, ¶200, ¶229-230) y el Corán (5: 45, 2:178-179). Debe aclararse que si bien los sistemas legales actuales no contemplan esta clase de castigos, aun persiste su aplicación en algunos países.

CONCLUSIONES

La teoría de la evolución ha enfrentado un importante reto explicativo en la tarea de dar cuenta del origen y mantenimiento estable de la cooperación, labor que adquiere mayor complejidad en el caso de los seres humanos en quienes la cooperación se ha extendido de manera significativa. En esa materia tradicionalmente se han identificado dos problemas. Por un lado, se deben explicar las conductas altruistas, esto es, aquellas en las cuales un individuo incurre en un aparente sacrificio para proporcionar un beneficio a otro organismo. Por otro lado, está el problema de la acción colectiva dirigida hacia la provisión de bienes públicos, pues en ese contexto la alternativa de conducta más favorable para la aptitud del individuo, consiste en beneficiarse del bien sin incurrir en el costo que debe asumir para la producción del mismo.

En procura de una solución al problema del altruismo, se propuso el modelo basado en la reciprocidad propuesto inicialmente por Trivers en 1971 según la cual, si el beneficio para el receptor es mayor que el costo para el donante, la selección natural puede favorecer conductas altruistas pues se producen beneficios para el donante derivados del retorno de la ayuda por parte del receptor. El trabajo de Trivers fue complementado posteriormente por Axelrod y Hamilton (1981) quienes con el fin de explicar el inicio y persistencia de la cooperación a partir de un ambiente no cooperativo, realizaron un experimento computacional que mostró cómo una estrategia de cooperación recíproca como TFT podía evolucionar de manera estable. Con posterioridad a los estudios de Axelrod y Hamilton, se desarrolló una ramificación importante de la tesis de la reciprocidad, el mecanismo de reciprocidad indirecta, entendida como aquella en la cual el individuo cooperador no recibe un retorno directo por parte del receptor de la cooperación, sino indirectamente por parte de otro individuo del grupo social.

Frente al problema de la acción colectiva, situado en contextos de muchos individuos, la tesis del altruismo recíproco no resultó exitosa. Las simulaciones de Boyd y Richerson (1988) mostraron que la reciprocidad es el resultado evolutivamente menos probable a medida que los grupos se tornan más grandes y, en cuanto al mecanismo de la reciprocidad indirecta, los resultados no han sido concluyentes en cuanto al efecto del aumento del tamaño del grupo.

Otra línea de investigación que ha buscado resolver el problema de la acción colectiva es el modelo basado en el castigo, según el cual, la aplicación de sanciones a los *free riders* es un mecanismo que explica el mantenimiento de la cooperación en contextos multipersonales. En esta materia se obtuvieron resultados a favor de la eficacia de este mecanismo en el trabajo teórico de Boyd y Richerson (1992) y en múltiples experimentos en contextos de juegos de bienes públicos (Fehr and Gachter 2000, Masclet et al. 2003, Ostrom et al. 1992).

Con base en el análisis de las soluciones dadas al problema del altruismo y al problema de la acción colectiva, se logró establecer que no son dos problemas distintos sino que en realidad se trata de un mismo problema, el problema del altruismo. Aunque en escenarios distintos, en ambos casos el dilema es idéntico, la mejor opción para cada individuo consiste en abstenerse de cooperar y, si todos actúan de la misma manera, la cooperación fracasa. En ambos contextos el problema para la teoría de la evolución consiste en explicar la razón por la cual la selección natural favorece conductas cooperativas en situaciones en las cuales el individuo cuenta con alternativas que en principio pueden representar mayores beneficios y por tanto la opción de cooperar podría reducir su aptitud biológica.

Esclarecido el problema que enfrenta la teoría de la evolución en la explicación de la cooperación, se asumió el segundo objetivo del trabajo consistente en esclarecer los elementos conceptuales de los principales mecanismos que se han planteado como solución a este problema, a saber, la reciprocidad y el castigo. Para ello se adelantó un examen de las aproximaciones de Sripada (2005) y Rosas (2008), dada su

pertinencia para el tema bajo estudio. Por un lado, Sripada considera que la aproximación basada en la reciprocidad no es una solución adecuada al problema del acatamiento moral, principalmente por cuanto en un escenario que involucre muchos jugadores, ese mecanismo no opera de manera selectiva, es decir, la defección que sea respuesta a una defección en una ronda previa, afecta no solamente al jugador que no cooperó inicialmente sino a todos los demás jugadores, lo cual conduce a que la cooperación frecuentemente fracase. Por el contrario, la explicación basada en el castigo constituye una mejor explicación pues no tiene esta limitación, esto es, el mecanismo opera de manera selectiva imponiendo selectivamente costos a los individuos no cooperadores. Por otro lado, y en fuerte contraste con Sripada, Rosas cuestiona la distinción radical entre reciprocidad y castigo, arguyendo que tal distinción opera sólo a nivel conductual, bajo una noción estrecha de reciprocidad donde reciprocar es fundamentalmente responder con el mismo comportamiento. En contraste, una estrategia explicativa que tome en cuenta los mecanismos proximales de carácter psicológico, conduce a la noción amplia de reciprocidad, bajo la cual se desvanece la distinción entre reciprocidad y castigo pues la similitud entre la retaliación del altruismo recíproco y ciertos castigos, como la exclusión o el ostracismo, sugiere la existencia de un mecanismo psicológico general que produce ciertas conductas con el mismo propósito en diferentes circunstancias.

La propuesta de Rosas sugirió la distinción entre dos tipos distintos de enfoques al problema de la distinción entre reciprocidad y castigo. El primero distingue reciprocidad y castigo en función de los efectos que ambos tipos de conductas tienen sobre la aptitud. El segundo lo hace en función de los mecanismos psicológicos que controlan dichas conductas. Se arguyó que sólo la primera resuelve la cuestión de la distinción entre reciprocidad y castigo, entendidos estos como mecanismos darwinianos de la evolución, pero se adujo que aún así entendidos los planteamientos de Sripada y Rosas presentaban algunas dificultades que justificaban proponer una dirección alternativa. En el caso de Sripada, se consideró que el mecanismo de reciprocidad debía ser conceptualizado de manera distinta,

menos restrictiva, de una manera tal que tenga en cuenta la naturaleza interacciones entre muchos individuos. Con respecto a la propuesta de Rosas, la noción estrecha resultó insuficiente pues al describir la reciprocidad como una simple acción de copia, ésta restringía en forma excesiva la esfera de comportamientos que podían considerarse recíprocos.

Con estas consideraciones en mente, se analizaron los elementos conceptuales de los mecanismos de reciprocidad y castigo. Se propuso como definición general de reciprocidad el intercambio condicionado de beneficios equivalentes en aptitud y se pasó a examinar los rasgos particulares de éste mecanismo en contextos bipersonales y multipersonales. A partir del modelo típico de reciprocidad que es la estrategia TFT, se identifican dos rasgos básicos del mecanismo de reciprocidad para el caso de las interacciones bipersonales. El primero de ellos consiste en que se trata de una estrategia condicional, dado que es contingente respecto de la conducta cooperativa del otro jugador. El segundo rasgo consiste en que no incorpora costos adicionales a la inversión que debe ser efectuada, ni tampoco modifica los beneficios derivados de la cooperación. Ello implica una diferencia clara respecto del mecanismo del castigo pues este último incorpora en la función de pagos, costos adicionales, generando una reducción en las ganancias del desertor.

En el caso de las interacciones multipersonales, se constatan los elementos básicos de la reciprocidad que se han mencionado. Por un lado, con respecto al segundo rasgo consistente en que los pagos no tienen variaciones que modifiquen los resultados de la interacción cooperativa, no ofrece discusión afirmar que éste se mantiene en el caso de las interacciones multipersonales. Por otro lado, en relación con el primer rasgo, esto es, el carácter condicional de la reciprocidad, su alcance en escenarios multipersonales debe considerar la naturaleza de las interacciones en ese entorno. Allí la cooperación debe estar condicionada al actuar cooperativo de los demás individuos, pero no necesariamente de todos los individuos. Esto puede ser inferido de la definición general propuesta. Dada esta definición, puede afirmarse que en contextos multipersonales la condicionalidad de la reciprocidad debe ser

entendida en el sentido de que la contribución del individuo está condicionada a que la cooperación por parte de los demás individuos del grupo genere un retorno proporcional al beneficio conferido a los demás. Pero esta condición no necesariamente se cumple con la contribución de todos los individuos.

Descritos los principales rasgos de la reciprocidad, el siguiente paso consistió en esclarecer los elementos básicos del mecanismo del castigo y su contraste frente a la reciprocidad. De manera muy general, el castigo puede describirse como una respuesta a una conducta de un individuo que se abstiene de cooperar en interacciones sociales, mediante la imposición de un costo que reduce sus ganancias. Por tratarse de una respuesta, el castigo tiene un carácter condicionado, su imposición está condicionada a la defección previa de un individuo, en contraste con la reciprocidad en la cual la cooperación está condicionada a la conducta cooperativa de otro u otros individuos. Similarmente, el castigo tiene un segundo rasgo característico ya mencionado, pues se modifican los costos de la interacción cooperativa mediante una deducción a los pagos a favor del desertor, eliminando la ventaja adaptativa que tiene el *free rider,* aspecto que traza una diferencia clave con respecto a la reciprocidad.

BIBLIOGRAFIA

Alexander R. D. (1987). The Biology of Moral Systems. Aldine Transaction.

Axelrod R., Hamilton W.D. (1981). The Evolution of Cooperation. Science 211 (4489): 1390-1396.

Axelrod R. (1984). The Evolution of Cooperation. Basic Books.

Bergmüller R. et al. (2007a). Integrating cooperative breeding into theoretical concepts of cooperation. Behavioural Processes, 76 (2): 61-72.

Bergmüller R. et al. (2007b). On the further integration of cooperative breeding and cooperation theory. Behavioural Processes, 76 (2): 170-181.

Boyd R., Richerson P. J. (1988). The evolution of reciprocity in sizable groups. J. Theor. Biol. 132, 337–356.

Boyd R., Richerson P.J. (1989). The evolution of indirect reciprocity. Social Networks, 11(3): 213-236.

Boyd R., Richerson P.J. (1992). Punishment allows the evolution of cooperation (or anything else) in sizable groups. Ethol. Sociobiol. 13, 171–195,

Clements K.C., Stephens D.W. (1995). Testing models of non-kin cooperation: mutualism and the prisoner's dilemma. Anim. Behav. 50: 527–535.

Clutton-Brock T. H., Parker G. A. (1995). Punishment in animal societies. Nature, 373 (6511): 209-216.

Cornes R., Sandler T. (1996). The Theory of Externalities, Public Goods and Club Goods, Cambridge University Press.

Cosmides, L. & Tooby, J. (2005). Neurocognitive adaptations designed for social exchange. In D. M. Buss (Ed.), The Handbook of Evolutionary Psychology (pp. 584-627). Hoboken, NJ: Wiley.

Croson R.T.A. (1999). Contributions to public goods: altruism or Reciprocity? The Wharton School. University of Pennsylvania.

Dawes R. M., Thaler R. H. (1988). Anomalies: Cooperation. The Journal of Economic Perspectives, 2 (3): 187-197.

Fehr E. & Gächter S. (2000). Cooperation and Punishment in Public Goods Experiments, American Economic Review 90, (2000): 980-994.

Fehr E., Gächter S. (2002). Altruistic Punishment in Humans, Nature 415 (10): 137-140.

Fischbacher U., Gächter S., Fehr E. (2001). Are people conditionally cooperative? Evidence from a public goods experiment. Economics Letters, 71 (3), June 2001: 397-404.

Gürerk Ö., Irlenbusch B. & Rockenbach B. (2006). The Competitive Advantage of Sanctioning Institutions. Science, 312: 108-111

Hamilton, W. D. (1963). The evolution of altruistic behavior. Am. Nat., 97: 354-356.

Hamilton W. D. (1964a). The genetical evolution of social behavior I. Journal of Theoretical Biology, 7 (1): 1-16.

Hamilton W. D. (1964b). The genetical evolution of social behavior II. Journal of Theoretical Biology, 7 (1): 17-52.

Hardin R (1971). Collective action as an agreeable n-prisoners' dilemma. Behavioral Science. 16 (5): 472–481.

Hardin R (1982). Collective Action. Baltimore, MD: Johns Hopkins University Press.

Henrich, J. et al. (2004). Overview and Synthesis. In Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies, edited by Henrich, J., et al. Oxford University Press.

Henrich J. et al. (2006). Costly Punishment Across Human Societies, Science 312: 1767-1770.

Hess C, Ostrom E. (2007). Glossary, in Hess C, Ostrom E. (eds.) Understanding Knowledge as a Commons: From Theory to Practice, Cambridge MA: MIT Press.

Imhof, L. A. & Nowak, M. A. (2010). Stochastic evolutionary dynamics of direct reciprocity. Proc. R. Soc. B 277: 463–468

Isaac, R. M. & Walker, J. M. (1988). Group size effects in public goods provision: The voluntary contribution mechanism. Quarterly Journal of Economics, 53, 1988: 179-200.

Isaac R. M., Walker J. M. & Williams A.W. (1994). Group size and the voluntary provision of public goods: Experimental evidence utilizing large groups. Journal of Public Economics, 54(1): 1-36.

Johnson D. D. P., Price M. E., Takezawa M. (2008). Renaissance of the individual: Reciprocity, positive assortment, and the puzzle of human cooperation. In Foundations of Evolutionary Psychology, C. Crawford & D. Krebs (eds.), pp. 331-352. New York: Lawrence Erlbaum.

Krakauer A. H. (2005). Kin selection and cooperative courtship in wild turkeys. Nature 434: 69-72.

Kurokawa, S., Ihara Y. (2009). Emergence of cooperation in public goods games. Proceedings of The Royal Society B Biological Sciences 276 (1660): 1379-1384.

Kurzban R. et al. (2001). Incremental commitment and reciprocity in a real time public goods game. Personality and Social Psychology Bulletin, 27: 1662-1673.

Kurzban R., Houser D. (2005). Experiments investigating cooperative types in humans: A complement to evolutionary theory and simulations. PNAS, 102 (5): 1803–1807.

Kurzban R., DeScioli P. (2008). Reciprocity in groups: Information-seeking in a public goods game. European Journal of Social Psychology Eur. J. Soc. Psychol. 38: 139–158.

Ledyard, J. (1995). Public goods: a survey of experimental research. In J. H. Kagel, & A. E. Roth (eds.), The handbook of experimental economics (pp. 111–194). Princeton, NJ: Princeton Univ. Press.

Leimar O., Hammerstein P. (2001). Evolution of cooperation through indirect reciprocity. Proc. R. Soc. Lond. B 268: 745-753.

Luce R.D., Raiffa H. (1957). Games and decisions: introduction and critical survey. Dover Publications.

Masclet D. et al. (2003). Monetary and nonmonetary punishment in the voluntary contributions mechanism. Am. Econ. Rev. 93: 366–380.

Maynard Smith J. (1964). Group selection and kin selection. Nature 201: 1145–1147.

McCabe, K. A., Rigdon, M. L., & Smith, V. L. (2003). Positive reciprocity and intentions in trust games. Journal of Economic Behavior & Organization, 52, 267-275.

Mesterton-Gibbons M., Dugatkin L. A. (1992). Cooperation among unrelated individuals: evolutionary factors. Q. Rev. Biol., 67: 267–281.

Mesterton-Gibbons M., Dugatkin L. A. (1997). Cooperation and the Prisoner's Dilemma: towards testable models of mutualism versus reciprocity. Anim Behav. 54 (3): 551-7.

Milinski, M. et al. (2001). Cooperation through indirect reciprocity: image scoring or standing strategy? Proc. R. Soc. Lond. B 268: 2495-2501.

Milinski M., Semmann D., Krambeck H.-J. (2002). Reputation helps solve the `tragedy of the commons'. Nature, 415: 24.

Nowak M., Sigmund K. (1993). A strategy of win-stay, lose-shift that outperforms tit-for-tat in Prisoner's Dilemma, Nature, 364: 56-58.

Nowak MA, May RM, Sigmund K (1995). The arithmetics of mutual help. Sci Am 272: 76-81.

Nowak M. A., Sigmund K.(1998). "Evolution of indirect reciprocity by image scoring." Nature 394 (6685): 573-577.

Nowak MA et al. (2004). Emergence of cooperation and evolutionary stability in finite populations. Nature 428: 646-650.

Okasha S. (2003). Biological Altruism. Stanford Encyclopedia of Philosophy.

Olson M. (1965). The Logic of Collective Action: Public Goods and the Theory of Groups. Cambridge, MA: Harvard University Press.

Ostrom E., Walker J., and Gardner R. (1992). Covenants with and without a sword: Self-governance is possible. Am. Pol. Sci. Rev. 86: 404–417.

Panchanathan K., Boyd R. (2003). A tale of two defectors: the importance of standing for evolution of indirect reciprocity. Journal of Theoretical Biology 224 (2003): 115–126.

Price M., Cosmides L., Tooby J. (2002). Punitive sentiment as an anti-free rider psychological device. Evolution and Human Behavior, 23(3): 203-231.

Price M. E. (2006). Monitoring, reputation and "greenbeard" reciprocity in a Shuar work team. Journal of Organizational Behavior 27: 201-219.

Price, M. E. (En prensa). Cooperation as a classic problem in behavioural biology: Past progress, current challenges. In Evolutionary Psychology: A Critical Introduction, V. Ed. Swami, Oxford, Wiley-Blackwell. (Draft - Not the final version)

Rosas A. (2008). The return of reciprocity: a psychological approach to the evolution of cooperation. Biology and Philosophy, 23 (4): 555-566.

Smith V. L. (2004). Human Nature: An Economic Perspective. Daedalus, Vol. 133 (4): 67-76

Sripada C. S. (2005). Punishment and the strategic structure of moral systems. Biology and Philosophy, 20: 767–789.

Sripada C. & Stich S. (2006). A Framework for the Psychology of Norms, in P. Carruthers, S. Laurence & S. Stich (eds.), The Innate Mind: Culture and Cognition. Oxford University Press: 280-301

Sugden R. (1986). The economics of rights, co-operation and welfare. Oxford, UK: Basil Blackwell.

Suzuki S., Akiyama E., (2005). Reputation and the evolution of cooperation in sizable groups. Proc. R. Soc. London B 272: 1373–1377.

Suzuki S., Akiyama E., (2007). Evolution of indirect reciprocity in groups of various sizes and comparison with direct reciprocity. Journal of Theoretical Biology 245 (2007): 539–552

Takezawa M., Price M. E. (2010). Revisiting "The evolution of reciprocity in sizable groups": Continuous reciprocity in the repeated N-Person prisoner's dilemma. Journal of Theoretical Biology.

Tooby J., Cosmides L., Price, M.E. (2006). Cognitive Adaptations for n-Person Exchange: The Evolutionary Roots of Organizational Behavior. Managerial and Decision Economics, 27 (2/3): 103-129

Trivers R.L., (1971). The evolution of reciprocal altruism. Quarterly review of Biology, 46(1): 35-37.

West S.A., Griffin A.S., Gardner A. (2007). Social semantics: altruism, cooperation, mutualism, strong reciprocity and group selection. Journal of evolutionary biology 20(2): 415-32.