LAS EMOCIONES MORALES COMO ADAPTACIONES PARA LA COOPERACIÓN EN DILEMAS SOCIALES

ALEJANDRO ROSAS

ABSTRACT. Economic experiments have consistently shown that human subjects cooperate in social dilemmas, in spite of the contrary predictions of game theory. Some social scientists have argued that moral emotions can explain this anomaly. I briefly present the experimental evidence for such anomalous behavior and for the thesis that emotions explain it. I argue that moral emotions promote cooperation when integrated into a normative mechanism. The normative mechanism requires the ability to discriminate between cooperators and defectors and facilitates control of the latter. Following Frank (1988), I suggest that the normative mechanism and its moral emotions evolved as a solution to cooperation in social dilemmas, benefiting both the individual and the group.

KEY WORDS: Altruism, anomalies, cooperation, emotions, evolution, experimental economics, norms, punishment, selfishness, social dilemmas, temporal discounting.

1. INTRODUCCIÓN

Desde los desarrollos de la filosofía moral británica en la modernidad y en especial a partir del *Tratado de la naturaleza humana* de Hume, diversas disciplinas han reflexionado sobre el papel que corresponde a las emociones, tanto en la producción de los juicios morales como en la conducta moral. Recientemente, propuestas provenientes de las neurociencias (Damasio 1994, 1996; Rilling 2002) así como de la economía (Frank 1988; Fehr & Gachter 2002; Bowles & Gintis 2002), han sugerido que las emociones juegan un papel positivo en las decisiones racionales y morales. Las propuestas están motivadas por observaciones o experimentos y tienen consecuencias relevantes para la concepción teórica de agente racional. Ellas sugieren, por ejemplo, la necesidad de modificar el modelo de agente que subyace a la teoría económica. Es interesante que el comportamiento cooperativo, que parece anómalo desde ese modelo, también sea consi-

Departamento de Filosofía, Universidad Nacional de Colombia, Colombia.
arosasl@unal.edu.co

derado un rasgo paradójico desde el punto de vista de la teoría de la evolución por selección natural. El propósito de este ensayo es explicitar estas conexiones teóricas y abogar por la inclusión de las emociones morales en un tratamiento potencialmente novedoso y plausible de la evolución de la moral.

La teoría sobre la evolución de la moral parte de la "paradoja del altruismo". El mismo Darwin fue el primero en señalar una potencial contradicción con su teoría (Darwin 1989, cap. 5). La paradoja consiste en que el comportamiento egoísta tiene intuitivamente una ventaja biológica, reproductiva, frente al comportamiento moral, de modo que, si se trata en ambos casos de rasgos heredables, la selección natural debería eliminar al segundo. Paradójicamente, esto no ha sucedido. Para enfrentar la paradoja, la biología evolucionista contemporánea sugiere "quitarle el altruismo al altruismo" (Trivers 1971). Sin embargo, esta frase alude sólo al altruismo definido *biológicamente* como la transferencia de aptitud de un individuo a otro. El altruismo es aparente, biológicamente hablando, si la donación retorna al altruista con intereses, pero ello no implica que el comportamiento sea egoísta *psicológicamente* hablando. Desde esta perspectiva, es posible que las *emociones* morales sirvan a los intereses del individuo (sean egoístas en sentido biológico) sin dejar por ello de ser auténticas emociones morales y, por tanto, sirvan también a los intereses del grupo social. Esta es la posibilidad teórica que exploro en este ensayo.

Darwin afirmó la existencia de la moral como rasgo paradójico y ancestral en la especie humana sobre la base de anécdotas acerca del comportamiento de individuos en sociedades tribales (sin autoridad política centralizada), ya sea por relatos de terceros o por lo que él había observado, por ejemplo, en Tierra de Fuego. Afirmó, por ejemplo, que hay individuos que se sacrifican por su tribu en confrontaciones bélicas y se resisten a traicionarla por un beneficio personal si son tomados prisioneros. Hoy, gracias a técnicas experimentales, tenemos evidencias de la paradoja que superan lo anecdótico. Un conjunto importante de evidencias viene del terreno de la economía experimental. Aunque son menos espectaculares que el sacrificio de los individuos en épocas de guerra (también de esas manifestaciones tenemos hoy suficientes), su carácter paradójico o anómalo no ha escapado a atención de los investigadores. Además, dan testimonio de la ubicuidad y cotidianeidad del fenómeno.

2. LA ANOMALÍA DE LA COOPERACIÓN

Partamos de una caracterización de egoísmo que nos permita entender la paradoja. Una conducta egoísta es aquella por la que un agente busca su propio beneficio. Pero esto no es necesariamente contradictorio con la moral. Al contrario, a menudo la moral lo exige. Sin embargo, el término

"egoísmo" tiene una connotación de reproche moral en el lenguaje cotidiano, y la misma connotación debe presuponerse cuando se afirma que la tesis de la moral como adaptación biológica es paradójica, porque el egoísmo es más adaptativo. No habría paradoja si no fuese porque el término egoísmo en esa afirmación connota conductas contrarias a la moral. En razón de esto, debemos caracterizar el egoísmo de modo que recojamos esa connotación: no es simplemente buscar el propio beneficio, sino buscarlo violando normas diseñadas para la interacción cooperativa y para la competencia limpia.

Con esta definición de egoísmo, notemos el paralelo entre la paradoja del altruismo en biología evolucionista y la paradoja o anomalía de la cooperación en economía y en teoría de juegos. Así como en biología evolucionista la selección natural predice un comportamiento egoísta contrario a la moral, así también la predicción de la teoría de juegos es que los humanos nos comportaremos como maximizadores egoístas en situaciones con estructuras de interacción propias de dilemas de prisionero (DP). Los niveles de cooperación experimentalmente confirmados en situaciones que suelen caracterizarse como dilemas sociales (Kollock 1998) son paradójicos según esta predicción. Especialmente en interacciones de una sola jugada, la estrategia dominante en dilemas de prisionero es la defección o *free-ride*, pues cuando cada jugadora se enfrenta a la elección entre cooperar (C) o no-cooperar (D), puede deducir fácilmente de la matriz de utilidades en esas situaciones que, cualquiera que sea la elección de las demás jugadoras, a ella le conviene en cualquier caso no-cooperar.

Matriz de pagas del DP. Para cualquier elección del otro, gano más si no coopero

Pagas para un C y un D del DP de n-jugadores. *I* es el capital inicial, *c* el costo de la contribución y *b* el beneficio producido por un C para *n* individuos.

Veamos en detalle la lógica de esta predicción cuando se trata de un DP de n jugadores, también conocido en economía como el problema de la provisión de bienes públicos. En el laboratorio, los psicólogos o los economistas experimentales diseñan el DP de varios jugadores de la siguiente manera: cuatro jugadoras comienzan con un capital, pongamos 10 unidades monetarias (UM), y son invitadas a contribuir una cantidad *C*, la que ellas decidan, en un fondo común. El monto depositado por todos los

contribuyentes en el fondo común se multiplica por un número mayor a 1 pero menor al número de jugadoras, pongamos 2; el valor resultante se reparte por partes iguales entre las cuatro jugadoras. Es decir, cada UM que una jugadora contribuye al fondo se convierte en 2 UM en el fondo común, pero la jugadora que la contribuyó sólo recibe 0.5 UM ($C \cdot 2/4$, ó $C/2$). Si todas y cada una contribuyen C, el depósito total del fondo es de $4C$ y cada jugadora obtiene $4C \cdot 2/4 = 2C$. Así, todos obtienen una ganancia neta. Pero como ninguna jugadora sabe con certeza si las demás contribuirán, y dado que de su propia contribución siempre le retorna sólo la mitad, no contribuir (D) es la estrategia dominante. Lo mejor que cada una puede hacer es quedarse con su capital inicial y esperar que las demás contribuyan, pues puede ganar algo de las contribuciones de ellas, pero nada con las propias. La jugadora racional-egoísta obtiene así su capital inicial más la mitad de cualquier suma que en total hayan contribuido las demás. La teoría de juegos asume que los seres humanos somos racionales y egoístas, es decir, que en las interacciones con los demás calculamos el comportamiento que nos trae el mayor beneficio y no dudamos en perseguirlo, incluso a expensas de otros. Sin embargo, si todos los jugadores se comportan según la predicción y juegan D en estas interacciones, todos obtendrán un resultado deficiente, pues podrían obtener más si todos cooperasen. Pero, racionalmente, no parecen tener otra opción en esas situaciones, de ahí su carácter de dilemas.

En la década de los ochenta, psicólogos y economistas experimentales comenzaron a realizar experimentos en los que ponían a sujetos humanos a interactuar en dilemas de este tipo. Encontraron que las predicciones de la teoría de juegos (TJ) son falsas. Aproximadamente un 50 por ciento de los sujetos cooperan en el DP de una sola jugada. En el caso de la provisión de bienes públicos, los primeros experimentos mostraron que la contribución promedio oscilaba entre el 40 y el 60 por ciento del capital inicial para distintas condiciones experimentales (Dawes y Thaler 1988). Una desviación interesante se dio cuando los sujetos eran estudiantes de posgrado en economía: su contribución promedio fue del 20 por ciento. Se utilizaron también otros juegos, como *ultimatum,* que miden la disposición a distribuir o repartir equitativamente un bien. En *ultimatum* juegan dos sujetos, una proponente que recibe todo el dinero y hace una oferta de repartición, y una receptora que tiene el poder de aceptar o rechazar la oferta. Si la receptora acepta, el bien se divide según la oferta propuesta; si la rechaza, ninguno de las dos jugadoras obtiene nada. Estos experimentos se practicaron también en sociedades de pequeña escala, o premodernas, y muestran rangos de variación en la oferta de los proponentes similares a los encontrados en estudiantes universitarios (Henrich et alia 2006).

El siguiente diagrama, tomado de Henrich et alia (2006), consigna resultados del juego *ultimatum* en diversas culturas. Las burbujas negras muestran, con su tamaño, el porcentaje de rechazos (indicado en algunas como guía) para cada cultura representada en el eje Y y para cada oferta de repartición del monto monetario indicado en el eje X. Así por ejemplo, los Gusli y los Maragoli (ambos de Kenya, África) rechazan en porcentajes altos las ofertas por debajo del 40 por ciento de la suma a repartir; en cambio los Samburu, de la misma región, rara vez rechazan ofertas positivas, no importa su magnitud. Las líneas verticales sólidas indican el promedio de las ofertas de repartición para cada cultura. Se observa que oscilan entre el 30 por ciento y el 50 por ciento. Las líneas verticales punteadas indican la oferta que hubiese maximizado la ganancia de la jugadora proponente, teniendo en cuenta el promedio de rechazos de su cultura. Se observa que las proponentes, en general (salvo los Gusli, los cima

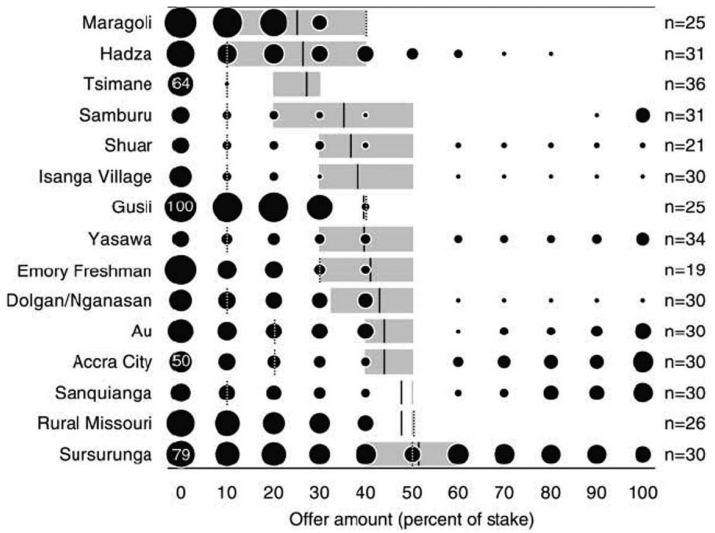


FIGURA 1

Niveles de oferta en el juego *ultimatum* en diversas culturas. Las burbujas negras muestran, con su tamaño, el porcentaje de rechazos (indicado en algunas como guía) para cada cultura representada en el eje Y, y para cada oferta de división del monto monetario indicado en el eje X. Ver explicación en el texto. Fuente: Henrich et alia 2006.

En suma, los experimentos han mostrado, consistentemente, que la predicción de la teoría de juegos junto con el axioma de conducta egoísta es falsa. Esta "anomalía" en el comportamiento humano en juegos estratégicos pone de manifiesto el mismo fenómeno al que se alude con la paradoja

del altruismo en biología evolucionista cuando se trata de humanos. Transferir beneficios a otros en interacciones cooperativas parece tan irracional desde el punto de vista económico, como maladaptado o "ecológicamente irracional", desde el punto de vista evolutivo.

Es necesario explicitar que la TJ predice cooperación en juegos repetidos. Cuando el juego se repite indefinidamente, y se trata de un DP de dos jugadores, Axelrod y Hamilton (1981) mostraron analíticamente que las estrategias cooperativas condicionales como *tit for tat* (TFT) obtienen ganancias superiores a las estrategias no-cooperativas (aunque hay razones para pensar que TFT es una estrategia demasiado rígida (Kollock 1993). Los mismos resultados se obtuvieron con simulaciones computacionales (Axelrod 1984). En este caso, la egoísta que juega D en un juego de un solo periodo no tiene incentivos para jugar D en lugar de C en un juego repetido, si la jugadora adversaria responde a D con D y a C con C. Esto es lo que hace la estrategia condicional y cooperativa TFT.

Pero TFT no es exitosa en juegos de varios jugadores o juegos de bienes públicos. Aquí, de nuevo, la estrategia dominante es D (Boyd & Richerson 1988). Y sin embargo, los experimentos con juegos repetidos de n-jugadores en el laboratorio muestran niveles paradójicos de cooperación inicial (aprox. 50 por ciento), aunque la cooperación decae a medida que el juego a caída de la cooperació

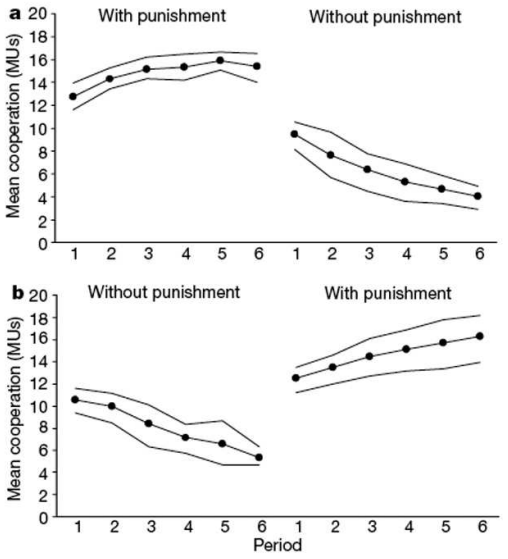


FIGURA 2

Niveles de cooperación promedio en juego de bienes públicos repetido por 12 periodos, 6 sin castigo y 6 con castigo alternando el orden (a primero sin castigo, b en orden inverso). Sin castigo los niveles decaen, pero suben o se mantienen cuando hay castigo. Fuente: Fehr & Gächter 2002.

3. EL JUGADOR EMOCIONAL

Ante estos resultados experimentales, que desmienten el axioma del egoísmo utilizado por la economía y la TJ, los psicólogos buscaron una explicación en la psicología de los jugadores. Hay dos razones por la cuales un individuo puede decidir no cooperar en un dilema del prisionero: ya sea porque teme que el otro jugador decida no cooperar, en cuyo caso le conviene no cooperar; o ya sea porque desea obtener un beneficio a costa de la cooperación del otro. Al primer motivo se lo denominó temor (*fear*), y al segundo codicia (*greed*) (Coombs 1973). Advirtamos, primero, el contraste entre esta concepción del jugador y la que suele presuponer la TJ tradicional. En la TJ, el jugador es un maximizador que calcula en "frío" que, hagan lo que hagan los demás jugadores, siempre le irá mejor en un DP de una sola jugada jugando la estrategia D (no-cooperar): pues si los demás no cooperan, le conviene jugar D (no-cooperar); y si los demás cooperan, también le conviene jugar D para obtener mayores ganancias. En una aproximación psicológica, en cambio, se presume que los jugadores no calculan fríamente, sino que emociones como el temor o la codicia juegan un papel. Creo que esa presunción es correcta. Básicamente, lo que ocurre es que cuando las personas se enfrentan a estos juegos no pueden impedir que normas de cooperación se activen y demanden autoridad sobre su conducta. Y al activarse las normas se activan también las emociones. Temor y codicia se refieren precisamente a esas emociones. Siendo esto así, me parece justo precisar que llamar "temor" a una de las motivaciones en cuestión es una manera indirecta e insuficiente de referirse a una motivación más compleja, pues es lógico asumir que quien decida no cooperar por temor a que no haya suficientes cooperadores, es alguien que preferiría cooperar si estuviese segura de la cooperación de los demás. Así, su estructura de preferencias puede expresarse del modo siguiente:

Preferencias Cooperativas: Si los demás cooperan, prefiero cooperar, es decir, tengo aversión a ser la codiciosa que gana a expensas de otros, pero si los demás no cooperan, prefiero no cooperar, pues tengo aversión a ser el simplón (*sucker*) que es explotado por otros. En cambio, detrás de lo que se llamó "codicia" está la negación de la preferencia 1.

Preferencias Oportunistas: Si los demás cooperan, prefiero no cooperar con el fin de aprovechar la oportunidad de ganar a sus expensas.

El codicioso se diferencia del frío del sujeto maximizador porque obra bajo la percepción emocional de los beneficios obtenidos a expensas de otros y probablemente tiene conciencia de violar sus expectativas de cooperación. La conducta no-cooperativa del codicioso está gobernada

por esta percepción "caliente" o codiciosa de los beneficios obtenidos, aunque se deban violar normas y expectativas sociales.

Robyn Dawes y un grupo de psicólogos hicieron experimentos para discriminar entre la codicia y el temor como explicaciones del comportamiento no-cooperativo en juegos de bienes públicos (Dawes et alia 1986). Hasta donde puedo ver, era la primera vez que se investigaba experimentalmente el papel de motivaciones emocionales en la conducta de los sujetos en esos experimentos. Utilizaron un juego de bienes públicos de un solo periodo para 7 jugadoras, quienes reciben cada una 5 dólares al inicio del juego y reciben 10 dólares cada una, hayan o no contribuido, si al menos tres jugadoras (o 5 en otra versión) contribuyen todo su capital. En la forma estándar del juego, si se cumple la condición las contribuyentes saldrán con 10 y las *free-riders* con 15 dólares. Pero si no se cumple, las contribuyentes saldrían con 0 y las *free-riders* con los iniciales 5 dólares. En este juego, al igual que en el DP de dos jugadoras, hay dos posibles motivos para no contribuir: el temor a perderlo todo si no hay suficientes contribuyentes, y/o la codicia de querer ganar a costa de las contribuyentes en caso de que el mínimo requerido se logre sin la propia contribución.

Para investigar el rol de estas posibles motivaciones, introdujeron modificaciones al juego con el fin de evocarlas por separado. Una de las modificaciones consistía en introducir una garantía de devolución de dinero si no se alcanzaba el número requerido para la bonificación. De esa manera eliminaron del juego el temor a perder la contribución (ser el simplón o *sucker*), pero mantuvieron un contexto propicio para la codicia, pues si se alcanzaba el número requerido de contribuyentes las no-contribuyentes salían con 5 dólares más que las contribuyentes. Observaron que el porcentaje de contribuyentes no variaba significativamente con relación a la forma estándar del juego. El temor a ser el simplón no podía ser, entonces, el que impedía que hubiese un número mayor de contribuyentes. La otra modificación consistía en uniformizar el saldo final. Si se alcanzaba el mínimo requerido de contribuyentes, todos obtendrían un saldo final neto de 10 dólares, independientemente de si habían contribuido sus 5 dólares o no. Esta modificación elimina el incentivo para la codicia, pues nadie puede ganar a costa de las contribuciones de los demás: todos terminan con 10 dólares al final si se logra el mínimo requerido de contribuyentes. Observaron que la contribución subió significativamente —del 51 por ciento al 86 por ciento en un experimento, del 64 por ciento al 93 por ciento en otro— y concluyeron que la codicia, y no el temor a ser el simplón, es el motivo principal que impide que los niveles de contribución suban por encima de un promedio del 50 por ciento en los juegos estándar.

Recientemente, algunos economistas experimentales han propuesto explicar los niveles paradójicos de cooperación en estos juegos con una

clasificación de los sujetos experimentales según sus motivaciones y preferencias (Fischbacher y Gächter 2006; Gächter 2006). La idea es que las personas se pueden clasificar en distintas categorías, según su estructura de preferencias sobre cómo comportarse en interacciones económicas como las que se simulan en esos juegos. En particular, distinguen dos tipos definidos: los cooperadores condicionales y los *free-riders*. Sus preferencias

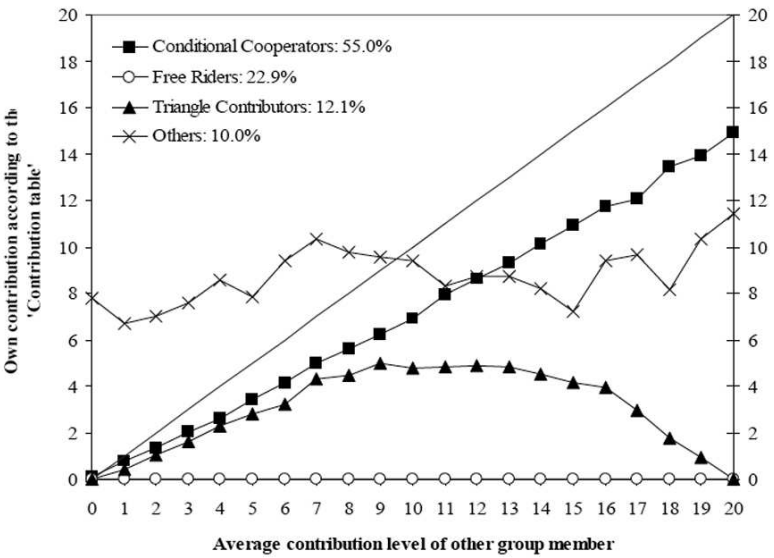


FIGURA 3
Funciones de contribución promedio de distintos tipos de jugadores: en especial interesan los *free-riders* que nunca contribuyen nada (línea de círculos vacíos que coincide con el eje X) y los cooperadores condicionales que contribuyen si los demás contribuyen y en montos semejantes a la contribución esperada (línea de cuadrados sólidos). Fuente: Fischbacher y Gächter 2006.

Esta clasificación nos permite hacer algunas conjeturas: La tasa de cooperación observada en los experimentos de DP de una sola movida entre dos o más jugadores corre en buena medida por cuenta de los cooperadores condicionales. El incremento que observaron Dawes y sus colegas en sus experimentos de hace veinte años cuando quitaron el incentivo para la codicia debe atribuirse a los *free-riders*, que al verse sin incentivo para la

codicia deciden contribuir también. En los juegos repetidos entre varios jugadores, es decir, los que simulan la provisión de bienes públicos, la tasa de cooperación decae porque los cooperadores condicionales reaccionan a las jugadas codiciosas o a las menores contribuciones de otros jugadores reduciendo su propia contribución.

4. EMOCIONES SOCIALES Y NORMAS

Las emociones que se involucran en la estrategia de los jugadores en estos juegos están estrechamente ligadas a normas sociales de conducta. Decimos que una conducta está guiada por una norma cuando un mecanismo normativo juega un papel explicativo especial en la producción de la conducta. Particularmente, el mecanismo normativo genera una disposición a actuar conforme a la norma a través de un proceso que se puede describir de la siguiente manera [i]:

1. Cuando el sujeto se encuentra en una situación a la que se aplica la norma, se activa un mecanismo de monitoreo de la adecuación de sus conductas a la norma;
2. Si no hay adecuación, se genera una emoción negativa (displacer) hacia sí mismo;
3. El displacer da lugar a un esfuerzo por corregir la conducta para que se ciña a la norma;
4. Los puntos anteriores (2 y 3) son independientes de sanciones externas producto del incumplimiento (castigos de la autoridad, por ejemplo).

Este conjunto de propiedades coincide con las condiciones que se cumplen cuando decimos que una persona *internalizó* la norma. Ellas distinguen la conducta guiada por normas de la conducta guiada por hábitos y por planes que no implican normas. Ni las conductas guiadas por planes predeterminados ni las guiadas por hábito cumplen los puntos anteriores, pues si uno nota *a posteriori* que no cumplió con un plan predeterminado o que no actuó como habitualmente lo hace, no siente displacer (2), ni trata de ajustar la conducta al plan prestablecido (3) si no ha habido consecuencias negativas externas (4). Todas las conductas guiadas por normas producen un displacer particular cuando se nota la propia falta de conformidad y motivan un esfuerzo por ajustar la conducta. El displacer puede ser del tipo del arrepentimiento o de la vergüenza. No siempre se trata de normas morales: los cuatro puntos valen también para las conductas guiadas por normas de etiqueta.

Según esta concepción, la vergüenza y el arrepentimiento, por ejemplo, son emociones básicas ligadas a un mecanismo normativo que se activa cuando el sujeto se reconoce enfrentado a una situación a la que se aplica la norma. Este análisis es compatible con los modelos de psicólogos y neurocientíficos. La respuesta emocional está preprogramada para cursos de acción que incumplan la norma. Ekman (1999) acuñó para esto el término

'programa de afecto' (*affect program*). Adicionalmente, la representación de ese curso de acción está marcada somáticamente de manera negativa para facilitar al sujeto la elección de un curso de acción beneficioso (Damasio 1996).

El antropólogo Daniel Fessler, de la UCLA, propuso un esquema innato y universal para la emoción de la vergüenza, que a diferencia del arrepentimiento, requiere que los sentimientos negativos hacia sí mismo sean mediados por sentimientos negativos de los miembros del colectivo social hacia el individuo transgresor. Aunque su propuesta tiene plausibilidad intuitiva para nuestra cultura urbana moderna, la menciono precisamente porque no se basa en una intuición sobre el uso de la palabra 'vergüenza' en esta cultura, sino en el análisis de 305 casos de la cultura Bengkulu, Indonesia, en los cuales los entrevistados o informantes espontáneamente utilizaron la palabra *malu* para describir su propio estado emocional o el de otros en su aldea. Fessler traduce *malu* como vergüenza [ii]. El esquema tiene esta estructura:

1. Ego viola una norma
2. Ego es consciente de su falta
3. Otro es también consciente de la falta de Ego
4. Ego sabe que Otro sabe de su falta
5. Otro despliega hostilidad y repulsión hacia Ego, o
   Ego asume que Otro experimenta hostilidad y repulsión hacia Ego
6. Ego experimenta *malu*, una emoción aversiva (Fessler 1999).

En toda emoción social el estímulo implica ya una percepción de una situación a la que aplica una norma social. Es decir, el estímulo que dispara las emociones se presenta en situaciones sociales complejas, que no "existen" sino para los organismos que pueden atribuir estados mentales, tales como emociones dirigidas hacia otras personas y representaciones de normas colectivas de conducta. Por otro lado, el reconocimiento de la situación a la que aplica la norma está usualmente mediado por valoraciones y creencias culturalmente determinadas, pues no en todas las culturas se clasifican los mismos hechos como infracciones a una misma norma. Normas de generosidad pueden ser comunes a muchas culturas, pero la generosidad no se mide de la misma manera en todas. Una mediación semejante puede existir entre el "programa de afecto" y la expresión conductual del mismo, que puede presentar variaciones culturalmente determinadas (Ekman 1999, Lazarus 1991). Este complejo sistema, en donde interactúan factores innatos y factores culturales, se muestra en el diagrama siguiente (tomado de Mallon & Stich 2000):

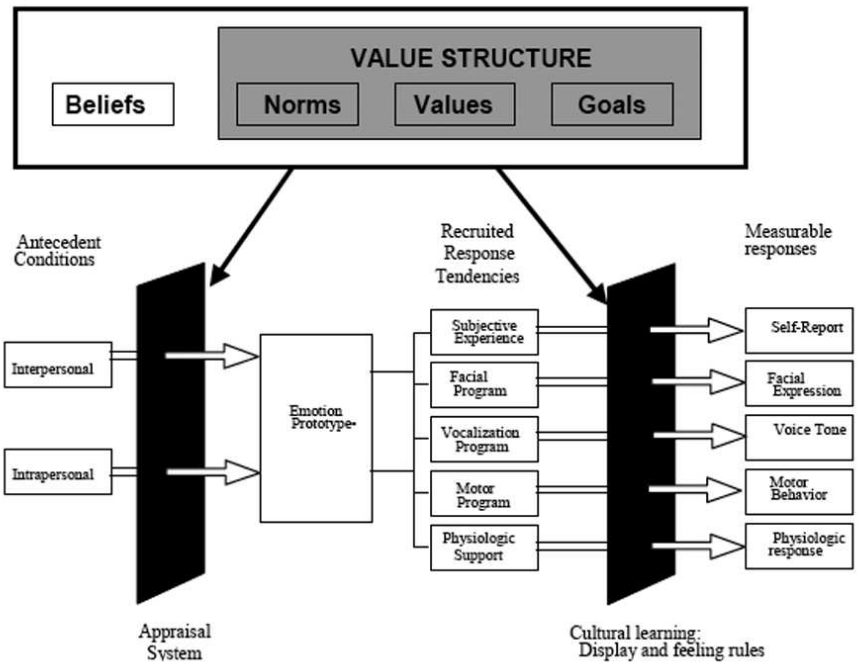


FIGURA 4
Elaboración del modelo de Levenson, en el que se explicita el rol de las creencias y las estructuras de valor en la generación de una emoción.

5. EL RETO DE LA COOPERACIÓN EN LOS DILEMAS SOCIALES

Los experimentos económicos descritos arriba sugieren la existencia de mecanismos normativo-emocionales que explican los patrones de cooperación observados. ¿Es factible afirmar que se trata de adaptaciones individuales? ¿Pudieron estos mecanismos ayudar a nuestros ancestros a sobrevivir en el pasado evolutivo? Robert Frank, economista de Cornell, esbozó en 1988 una teoría de las emociones morales que implica una respuesta positiva a estos interrogantes. Frank tuvo en cuenta los resultados de la economía experimental naciente y trató de mostrar que la estructura de interacción social típica de los DP de dos o más jugadores es el problema para el que las emociones morales son la respuesta. En su teoría, las interacciones con esta estructura reciben el nombre de *commitment problems* (problemas de compromiso) pues Frank los ve desde el reto que la estructura dilemática le plantea a un actor que quiere beneficiarse de las oportunidades de cooperación allí presentes. En la estructura de

interacción de estas situaciones hay una tensión entre la racionalidad individual y la colectiva, o entre el beneficio individual y el del grupo: hay un incentivo para que el individuo se aparte de la conducta que lleva al óptimo para el grupo, en aras de obtener un óptimo para él. Si no se introducen mecanismos sociales o psicológicos especiales, la cooperación no tiene perspectivas de realizarse, con lo cual pierden tanto el individuo como el grupo. Frank ilustra esos dilemas o "problemas de compromiso" con cuatro ejemplos, todos de interacciones entre dos personas. Pero su teoría puede también hacerse extensiva a dilemas entre muchos jugadores.

1. P y R (Pedro y Rosa o Patricia y Raúl) quieren abrir un restaurante. P sabe cocinar, R sabe administrar. Pero a ambos les preocupa que el socio podrá hacer trampa, apropiándose a sus espaldas, de manera sutil e indebida, de algunos bienes producto de la cooperación. Eso los disuade de comenzar una empresa común que los beneficiaría a ambos.

2. P y R consideran la posibilidad de formar una familia, pero ambos temen que el otro abandone la empresa común en algún momento futuro. De ser así, se habrá perdido el tiempo y esfuerzo invertido en esa empresa común.

3. P y R pueden ganar de una actividad conjunta pero P, que es el (la) más necesitada(o), se enfrenta a la codicia de R, quien se quiere aprovechar de la necesidad de P forzándolo(a) a acordar previamente una repartición desigual de las ganancias.

4. P y R tienen hijos comunes. Las actividades de R perjudican a P, quien por esa razón debe dedicar al cuidado de los hijos más tiempo que R. R, en su codicia, sabe muy bien que P prefiere renunciar a su tiempo libre que descuidar a sus hijos.

Es importante notar que en todos los ejemplos la cooperación abre oportunidades de beneficios que no son accesibles a cada actor aisladamente. Además, los beneficios se obtienen gracias a una interacción sostenida en el tiempo: no son inmediatos sino diferidos a mediano y largo plazo. Por otro lado, esas oportunidades de mutuo beneficio se pueden frustrar por el comportamiento racional-egoísta de cada individuo. Es decir, existe la oportunidad de beneficios mutuos, pero también un obstáculo a su realización. Se presentan dos tipos de obstáculos, que conviene explicitar.

En 1 y 2 el problema es que el éxito de la cooperación depende de que ambos actores mantengan consistentemente en toda la duración de la interacción un deseo de interactuar cooperativamente. Pero ambos prevén que en algún punto futuro el abandono de la interacción cooperativa podría prometer más beneficios. Esto podría suceder haciendo trampa sutil y clandestinamente, como en el caso 1; o simplemente abandonando abiertamente la empresa común a mitad de camino por otra más beneficiosa, como en el caso 2. Ambos temen, entonces, que el otro adopte un

comportamiento codicioso cuando se presente la oportunidad. Este problema subsiste aun cuando ambos jugadores sean cooperadores de corazón, pues lo que pesa es la incertidumbre que cada uno tiene sobre el carácter del otro. Lo que se requiere es una señal inequívoca y manifiesta del talante cooperativo.

En 3 y en 4, el problema es distinto. Sucede que puede ser racional en algunas circunstancias ceder a la codicia de otro. Por ejemplo, cuando el afán de castigar una actitud codiciosa lleva a terminar una interacción cooperativa, los costos pueden ser más altos para quien más necesita los beneficios de la cooperación. Al menos desde ese punto de vista (el de los costos), el castigo es irracional. Uno de los jugadores, actuando con base en preferencias oportunistas, puede decidir sacar partido de este hecho, ofreciendo al más necesitado un beneficio menor, que para él sería "irracional" rechazar.

Si los actores en estas situaciones pudieran comprometerse con el óptimo para el grupo y expresar confiable y convincentemente ese compromiso, podrían cosechar los beneficios de la cooperación. El compromiso tendría una doble faz: en 1 y 2 se trataría del compromiso de renunciar a la codicia, es decir, a beneficios personales que se obtendrían a costas de la interacción cooperativa, ya sea haciendo trampa a espaldas del otro jugador (caso 1); ya sea abandonando la empresa cooperativa a mitad de camino por otra opción mejor (caso 2). En los casos 3 y 4 se trataría del compromiso del actor que se enfrenta a una contraparte codiciosa de acarrear con los costos que conlleve castigar su codicia y mantenerse firme en las exigencias de la norma de cooperación. Este compromiso debe ser también, como en el caso anterior, convincente y manifiesto para el jugador codicioso. Sólo así se lo puede disuadir de la tentación de la codicia cuando el castigo acarrea costos que no parece racional asumir. Resumiendo, la solución es que cada jugador tenga un *doble compromiso*, tanto con la actitud cooperativa como con la disposición a castigar la codicia, sean cuales sean los costos de ello, y que ese doble compromiso sea manifiesto y creíble para la contraparte.

Podemos llamar a la solución recién esbozada la solución del *doble compromiso manifiesto*. Cuando las interacciones cooperativas se prolongan en el tiempo, como suele ser el caso en los humanos, siempre se presentarán situaciones que, desde el punto de vista de los beneficios inmediatos, son "tentaciones", tanto para la codicia como para la tolerancia con la codicia. Si los humanos cediésemos a esas tentaciones de manera regular, la cooperación típicamente humana, a gran escala y diferida, no existiría. Es, pues, importante que los actores puedan superar las tentaciones gracias al compromiso. Pero también es importante que su capacidad para el compromiso sea visible y manifiesta a los demás. La señal natural de la capacidad de comprometerse con un curso de acción cooperativo es la que

se da a través de la reputación, es decir, del comportamiento consistente a lo largo del tiempo, aunque existen señales más sutiles y quizás más frágiles.

6. UN MECANISMO NORMATIVO-EMOCIONAL PARA SUPERAR LOS DILEMAS

Si la especie humana se caracteriza por la realización de la cooperación diferida y a gran escala, los problemas señalados tienen que haberse solucionado en el curso de la evolución. La tesis principal de Frank (1988) es que las emociones morales y las normas de cooperación asociadas son los elementos clave de la solución. Nos permiten mantener los compromisos requeridos, así como también dar señales visibles y confiables de los mismos a través del comportamiento consistente (y, sostiene Frank, a través de otras señales, que yo consideraría más frágiles, derivadas de las signaturas fisiológicas y observables de cada emoción particular). Los elementos explicitados arriba pueden considerarse robustos, es decir, no están abiertos a contrastación empírica. Por un lado, está la concepción de la estructura de interacción típica de la cooperación a gran escala. La estructura consiste en un conflicto entre el interés último del agente (que coincide con el interés colectivo), en aprovechar los beneficios diferidos e individualmente inaccesibles de la cooperación, por un lado, y su interés inmediato, que consistiría en aprovechar beneficios inmediatos obtenidos a expensas de la actitud cooperativa de los demás, por otro. Si no se resuelve de alguna manera este conflicto, los humanos nunca habríamos evolucionado para cooperar a gran escala, es decir, entre no-parientes y con beneficios diferidos (Trivers 1971). Sólo los insectos sociales han desarrollado un nivel de cooperación comparable, que se explica por el alto grado de parentesco genético entre los individuos que conforman las comunidades cooperativas (Alexander 1987). El otro elemento robusto es una concepción *general* de la solución requerida para cooperar en estas situaciones. Se trata de la solución expuesta arriba como el *doble compromiso manifiesto* con actitudes cooperativas y disposiciones a castigar la codicia.

La tesis del doble compromiso manifiesto supone que durante la evolución humana se presentaron de manera recurrente posibilidades de cooperación en interacciones dilemáticas, en donde existía un conflicto entre el interés inmediato y el interés colectivo o de largo plazo. Esta hipótesis requeriría para su contrastación un tratamiento cuidadoso de la evidencia antropológica y paleoantropológica disponible sobre la evolución humana y sobre sus aptitudes sociales y cooperativas ancestrales [iii]. Frank presenta, además, una hipótesis particular sobre el mecanismo motivacional concreto que hace posible el doble compromiso manifiesto. Se trata de una hipótesis sobre las emociones morales como soportes de la cooperación en interacciones dilemáticas. Como hipótesis empírica, está abierta a confirmación o refutación; y aunque resultase errada, merece atención, así sea

sólo porque los errores de la ciencia nos acercan a la verdad. A continuación expongo esta hipótesis.

El comportamiento humano, al igual que el animal, se rige por un mecanismo de recompensa que utiliza sensaciones subjetivas de placer y displacer. El hambre o apetito que impulsa a cualquier animal a comer es una sensación de displacer que se dispara en el sistema nervioso central cuando los índices nutricionales del cuerpo (por ejemplo, el nivel de azúcar en la sangre) indican valores por debajo de un umbral. De esta manera, los incentivos materiales, la necesidad objetiva de calorías en este caso, juegan sólo un papel indirecto, a través de su correlación con las sensaciones subjetivas, que son los motores directos de la acción. El ajuste o correlación entre las sensaciones subjetivas y el incentivo material es imperfecto: funciona bien en el entorno en el que evolucionó. En el caso del mecanismo que gobierna la ingestión de alimento, el entorno evolutivo presentaba típicamente periodos de escasez regulares. De ahí que el mecanismo favorece sobrealimentarse cuando hay alimento disponible, con el fin de acumular reservas para los periodos de escasez. En el entorno actual este diseño es maladaptativo, al menos para aquellos que dispongan de un acceso constante a los alimentos.

Este mecanismo de recompensas gobernado por sensaciones de placer y displacer es filogenéticamente muy antiguo. Tiene un sesgo presentista en el sentido de que el atractivo de una recompensa crece exponencialmente a medida que el tiempo que nos separa de su obtención se aproxima a cero si están próxin nómeno se den as espe-cies ar

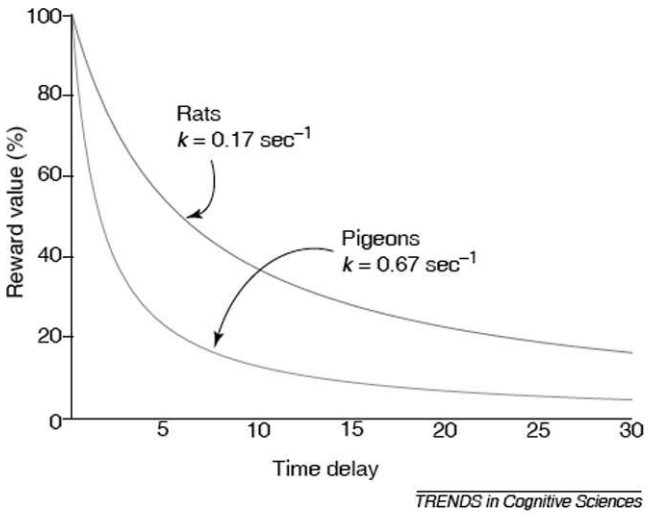


FIGURA 5
La curva de descuento temporal en ratas y palomas. La recompensa se devalúa en un 50 por ciento en 2 segundos en palomas y en 6 segundos en ratas.
Fuente: Stevens & Hauser 2004.

El diseño del mecanismo de recompensa con descuento temporal tiene una utilidad funcional indisputable para la supervivencia en contextos hostiles o altamente competitivos, pues quien no sobrevive al presente no tiene futuro. Todavía se puede constatar su influencia en humanos. Experimentos de este tipo lo demuestran: si ofrecemos a cada una de un conjunto de personas 100 pesos ahora o 120 dentro de 3 días, la mayoría preferirá 100 ahora. Pero si ofrecemos 100 pesos dentro de 28 días y 120 dentro de 31, aunque las cantidades y las distancias temporales entre ellas son idénticas a las del primer ofrecimiento, la mayoría preferirá 120 dentro de 31 días. La razón está en que el valor de una recompensa depende de su distancia relativa al presente. En el primer caso la distancia relativa entre ahora y dentro de 3 días es inmensa; opaca el hecho de que 120 pesos es objetivamente mayor a 100. El monto ofrecido dentro de tres días se descuenta comparado con el que se recibe ahora y preferimos este último. Pero la m do a 28 y otro situado a al presente, de ahí que el recido.

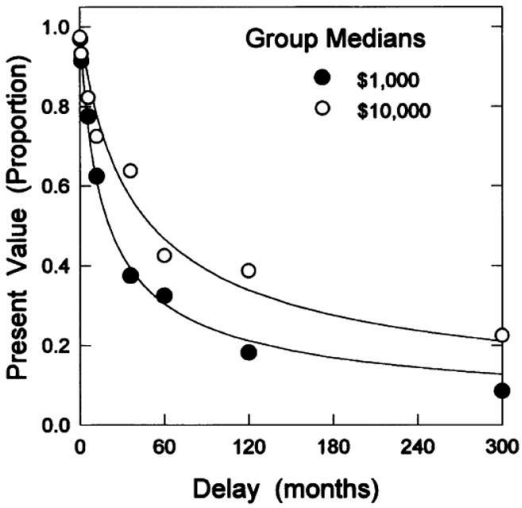


FIGURA 6
Curva de descuento temporal en humanos. La curva es similar a la de la figura 4, pero el descuento se mide en meses en lugar de segundos.
Fuente: Myerson & Green 1995.

El sesgo presentista del mecanismo de recompensa lo hace inadecuado para lidiar con las situaciones sociales que tienen la estructura antes mencionada, es decir, en las que el agente se enfrenta al conflicto entre un beneficio inmediato y un beneficio diferido, si bien mayor. El mecanismo de recompensa crea un problema de control de impulsos, pues presenta los beneficios inmediatos de tal modo que los diferidos no pueden competir con ellos. Como la cooperación a gran escala depende de poder controlar esos impulsos, este mecanismo es inadecuado para facilitarla.

En teoría, la selección natural pudo haber enfrentado esta situación tratando de modificar el diseño del mecanismo de recompensa, especialmente en lo relativo a su sesgo presentista. Obviamente, no podía simplemente sustituir el sesgo presentista por uno futurista, pues la supervivencia es principalmente asunto del presente. La única vía era modificar el diseño introduciendo una fina capacidad discriminatoria que le dijese al mecanismo cuándo dar más peso a los beneficios lejanos y cuándo más a los cercanos o inmediatos. Pero no cualquier diseño que uno pueda imaginarse estuvo verdaderamente disponible a la selección natural. Frank sostiene que emociones como la ira, el desprecio, la simpatía, la vergüenza y la culpa juegan un papel como "ensayos burdos de afinar el mecanismo de recompensa" (*crude attempts to fine-tune the reward mechanism*) para permitir a los humanos cooperar en situaciones dilemáticas.

Estas emociones no funcionan, sin embargo, independientemente y paralelamente al mecanismo de recompensa, sino que se sirven de él. Particularmente, ellas introducen un placer o displacer *presente* ante la perspectiva de ciertas acciones que conducen a beneficios o perjuicios futuros. La teoría es compatible con la hipótesis de Damasio, según la cual, marcamos somáticamente representaciones de cursos de acción con base en sus desenlaces negativos o positivos en la experiencia pasada (Damasio 1996). Precisamente el carácter *presente* de ese placer y displacer permite que las anticipaciones de daños o beneficios lejanos compitan con las sensaciones que el mecanismo de recompensa activa con correlación a beneficios o daños inmediatos. De ese modo, los sentimientos morales logran superar el sesgo presentista del mecanismo, *sirviéndose de él*. Un agente que tenga a su disposición estos sentimientos o emociones podrá controlar sus impulsos hacia beneficios inmediatos, que le impiden aprovechar los beneficios diferidos de la cooperación. Si se enfrenta a la tentación de hacer trampa en una relación cooperativa, la simpatía por la otra persona, así como la anticipación de la culpa o la vergüenza, le permiten controlar su impulso hacia el beneficio que promete la trampa. Por otro lado, un fuerte sentimiento de desaprobación como la indignación moral le ayuda a superar la tentación de evitarse los costos de castigar a una persona codiciosa. A largo plazo esto redunda en su beneficio, pues

obtiene así una reputación que disuade la agresión o codicia de otros en futuras interacciones.

De aquí se deriva el valor de la reputación como señal de un carácter comprometido con la cooperación. En principio, un bribón sensato (traduciendo la expresión de Hume: "*sensible knave*") podría pensar que puede mantener su reputación siguiendo la máxima de actuar honestamente cuando esté expuesto a la luz pública, y de hacer trampa con moderación y en secreto, es decir, en todas aquellas ocasiones en que pueda hacerlo sin ser expuesto. Hume ya había intentado explicar las razones por las que esta máxima del bribón no funciona (ENM, §233). Frank nos da una explicación equivalente y quizás más clara. El bribón se engaña si cree que puede gobernarse por esa máxima. La máxima supone que es posible tener una conducta cooperativa sobre la base de un cálculo racional *sin tener una aversión emocional* a violar las reglas de cooperación. Si el bribón quiere aprovecharse de las oportunidades de hacer trampa, tiene que arreglárselas sin esa aversión, pues si tuviera esa aversión y las emociones morales, no sería un bribón. Sin embargo, en ausencia de las emociones morales, el bribón queda a merced del mecanismo de recompensa con su sesgo presentista. Y ya vimos que el mecanismo de recompensa no permite controlar los impulsos hacia beneficios inmediatos, es decir, el bribón no será capaz de actuar consistentemente con su máxima.

Podemos incluso complementar esta explicación con otra reflexión que ni Hume ni Frank aportan. El cálculo racional en el que se quiere apoyar el bribón supone la disponibilidad de creencias ajustadas a la realidad sobre las probabilidades de ser descubierto. Pero muy probablemente, por un mecanismo que conocemos como racionalización, cada vez que el mecanismo de recompensa lo impulse a violar las reglas de cooperación a pesar de su probable exposición pública, el bribón remplazará en sus creencias la probabilidad real por una amañada, con el fin de creer que actúa en coherencia con su máxima. Sin sentimientos morales, el atractivo de los beneficios inmediatos es irresistible, y el bribón, a pesar suyo, ajusta sus creencias sobre las consecuencias de la violación a su deseo de beneficios inmediatos, en lugar de ajustarlas a la realidad, como debería. Con ello racionaliza su acción, pero termina arruinando su reputación.

La reputación, en suma, es un signo confiable del doble compromiso que se requiere para facilitar la cooperación en situaciones dilemáticas. Las emociones ligadas a normas sociales lo hacen posible. El agente anticipa su emoción negativa dirigida hacia sí mismo en caso de incumplimiento y también la emoción negativa de los afectados (o terceros) hacia él como transgresor. La explicación plantea varios interrogantes, que han de quedar aquí abiertos: ¿En qué momento y debido a qué presiones selectivas concre tas adquirieron nuestros ancestros la capacidad de percibir las oportunidades de beneficios cooperativos en situaciones dilemáticas? ¿En

qué momento pudieron proyectar consecuencias de largo plazo de acciones presentes? ¿Qué presiones selectivas concretas determinaron la adquisición de la capacidad de diseñar y representarse conscientemente normas colectivas de conducta?

Son preguntas que no podemos abordar aquí, pero es obvio que es necesario responderlas para completar la teoría evolutiva de la moral aquí sugerida, que asume que nuestros ancestros necesitaron cooperar en situaciones dilemáticas y evolucionaron un mecanismo psicológico especial para lograrlo.

NOTAS

i Ver por ejemplo, el análisis del filósofo Peter Railton de la Universidad de Michigan en Railton (2006).
ii Fessler llegó a la aldea sabiendo hablar indonés y se dedicó el primer año a aprender el dialecto bengkulu.
iii Ver una hipótesis al respecto en Rosas (2007).

BIBLIOGRAFÍA

Axelrod, R. (1984), *The Evolution of Cooperation*. New York: Basic Books.

Axelrod, R., Hamilton, W.D. (1981), "The evolution of cooperation," *Science* 211: 1390–1396.

Bowles, S., Gintis, H. (2002), "Prosocial emotions," Santa Fe Institute Working paper # 02-07-028. Accesado el 24 abril de 2004 en http://www.santafe.edu/research/publications/workingpapers/02-07-028.pdf.

Boyd, R., Richerson, P. J. (1988). "The evolution of reciprocity in sizable groups," *Journal of Theoretical Biology* 132: 337-356.

Boyd, R., Richerson, P. J. (1992), "Punishment allows the evolution of cooperation (or anything else) in sizable groups," *Ethology and Sociobiology* 13: 171-195.

Coombs, C. A. (1973), "A reparameterization of the prisoner's dilemma game," *Behav. Science* 18: 424-28.

Damasio, A. R. (1994), *Descartes' Error: Emotion Reason and the Human Brain*. New York, NY: Gossett/Putnam.

Damasio, A. R., Everitt, B. J., Boshop, D. (1996), "The somatic marker hypothesis and the possible functions of the prefrontal cortex," *Philosophical Transactions of the Royal Society of London B* 351: 1413-1420.

Darwin, C. (1989), *The Descent of Man, and Selection in Relation to Sex*. Second Edition, revised and augmented (1877), edited by P. H. Barrett and H. B. Freeman. New York: New York University Press.

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

Dawes, R., Orbell, J. M., Simmons, R., Van Kragt, A.J.C. (1986), "Organizing groups for collective action," *American Political Science Review* 80: 1171-1185.

Ekman, P. (1999), "Basic emotions," in T. Dalgleish and M. Power (Eds.) *Handbook of Cognition and Emotion*, pp. 45-60. Chichester: John Wiley and Sons Co. Accesado el 25/4/2005 en http://www.paulekman.com/pdfs/basic_emotions.pdf

Fehr, E., Gächter, S. (2002), "Altruistic punishment in humans," *Nature* 415: 137–140.

Fessler, D. M.T. (1999). "Toward an understanding of the universality of second order emotions," in *Beyond Nature or Nurture: Biocultural Approaches to the Emotions*, A. Hinton (ed.), pp. 75-116. New York: Cambridge University Press. (http://www.sscnet.ucla.edu/anthro/faculty/fessler/reprints.htm)

Fischbacher, U., Gächter, S. (2006), "Heterogeneous social preferences and the dynamics of free-riding in public goods," *CeDEx Discussion Paper* num. 2006–01. Accesado el 2/11/06 en http://www.nottingham.ac.uk/economics/cedex/papers/2006-01.pdf

Frank, R. (1988), *Passions Within Reason. The Strategic Role of the Emotions*. New York: W. Norton and Co.

Gächter, S. (2006), "Conditional cooperation: Behavioral regularities from the lab and the field and their policy implications," *CeDEx Discussion Paper 03*

http://www.nottingham.ac.uk/economics/cedex/papers/2006-03.pdf. Accesado el 27/10/06.

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., Cardenas, J. C., Gurven, M., Gwako, E., Henrich, N., Lesorogol, C., Marlowe, F., Tracer, D., Ziker, J. (2006), "Costly punishment across human societies," *Science* 312 (23 June): 1767-1770.

Hume, David (1751, 1902), *Enquiries Concerning the Human Understanding and Concerning the Principles of Morals*. L. A. Selby-Bigge (ed.), 2da ed. Oxford: Clarendon Press.

Kollock, P. (1993), "An eye for an eye leaves everyone blind: Cooperation and accounting systems," *American Sociological Review* 58(6): 768-786.

Kollock, P. (1998), "Social dilemmas: The anatomy of cooperation," *Annual Review of Sociology* 24: 183-214.

Lazarus, R. (1991), "Progress on a cognitive-motivational-relational theory of emotion," *American Psychologist* 46(8): 819-834.

Mallon, R., Stich, S. (2000), "The odd couple: The compatibility of social construction and evolutionary psychology," *Philosophy of Science* 67(1): 133-154.

Myerson, J., Green, L. (1995), "Discounting of delayed rewards: models of individual choice," *Journal of the Experimental Analysis of Behavior* 64(3): 263-276.

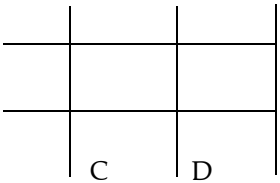
Railton, P. (2006), "Normative guidance," in *Oxford Studies in Metaethics: Volume 1*. Edited by Russ Shafer-Landau, New York: Oxford University Press. Accesado el 28/06/2006 en http://fds.oup.com/www.oup.co.uk/pdf/0-19-929189-6.pdf.

Rilling, J. K., Gutman, D. A., Zeh, T. R., Pagnoni, G., Berns, G. S., Kilts, C. D. (2002), "A neural basis for social cooperation," *Neuron* 35 (July 18): 395–405.

Rosas, Alejandro (2007), "El entorno ancestral de las normas de equidad," en A. Rosas (ed.) *Filosofía, Darwinismo y Evolución*, Bogotá: Universidad Nacional de Colombia.

Stevens, J. R., Hauser, M. D. (2004), "Why be nice? Psychological constraints on the evolution of cooperation," *Trends in Cognitive Sciences* 8: 60–65.

Trivers, R. (1971), "The evolution of reciprocal altruism," *Quarterly Review of Biology* 46 (1): 35-57.



	C	D
C	3 3	4 0
D	0 4	1 1

	D
C	$I+\sum b/n$ $I\text{-}c+\sum b/n$