DOI: 10.7213/revistadefilosofiaaurora.7778 ISSN 0104-4443

BOWLES, Samuel; GINTIS, Herbert. **A cooperative species**: human reciprocity and its evolution. Princeton: Princeton University Press, 2011.

Alejandro Rosas

Departamento de Filosofía, Universidad Nacional de Colombia, e-mail: arosasl@unal.edu.co

¿Cómo evolucionaron las preferencias sociales?[1]

Reciprocidad fuerte: Tesis centrales y dudas recurrentes

Explicar la cooperación humana, y en particular los mecanismos psicológicos que evolucionaron para sustentarla, es uno de los proyectos más desafiantes de la investigación científica interdisciplinaria en años recientes. *Una Especie Cooperativa* es un intento serio y reflexivo por aportar una explicación completa, sofisticado tanto en la teoría como en la evidencia empírica reunida para ese fin. Bowles y Gintis (B & G), dos economistas de renombre con amplio conocimiento en biología evolucionista y comprometidos con el poder explicativo de los modelos evolucionarios, pertenecen a un grupo de investigadores que en los últimos 10 años ha desarrollado una teoría sobre el carácter único de la cooperación humana, basados en una estrategia de interacción

[1] Una versión en inglés de esta reseña está en prensa en *Biological Theory*, v. 6, n. 2, p. 169-175, 2012. doi 10.1007/s13752-012-0013-y

que denominan "reciprocidad fuerte". La teoría conjuga dos tesis fundamentales: primero, el mecanismo próximo (psicológico) del comportamiento cooperativo humano incluye preferencias sociales y particularmente motivaciones altruistas. Esto contradice la tesis convencional en economía, según la cual la cooperación se basa en un egoísmo racional de largo plazo y en la mano invisible de Adam Smith. La tesis convencional fracasa en virtud del carácter necesariamente incompleto de los contratos y ante los retos planteados por interacciones del tipo del dilema de prisioneros (PD) y de la tragedia de los comunes (p. 5-6). Hardin (1968)[2] propuso el castigo como solución alternativa; pero si los castigadores (C) tienen motivos altruistas (p. 25-26, p. 31-32, p. 91-92, p. 165), el egoísmo no puede explicar la cooperación. Es preciso, entonces, concebir al agente económico como dotado de preferencias sociales y de emociones que apoyan el comportamiento cooperativo. El verdadero enigma a resolver es explicar qué procesos evolucionarios condujeron a este resultado (p. 6).

La otra tesis central de la teoría está ligada a la idea de que el sentido de equidad y las emociones morales son biológicamente costosas, es decir, altruistas (A) en el sentido de acarrear un costo neto en aptitud (*fitness*). Ello plantea una paradoja: la selección natural tenderá a favorecer a los no-altruistas (N), salvo que existan mecanismos evolucionarios especiales para favorecer los comportamientos que acarrean costos y pérdida en aptitud. Los autores proponen un mecanismo de selección de grupos (SG). Pero en los modelos de SG "el diablo está en los detalles". Así, su solución comprende al menos cinco componentes que se sostienen mutuamente: un modelo novedoso y empíricamente relevante de extinción selectiva de grupos por guerra inter-tribal (Cap. 7), que incorpora estimaciones empíricas de muertes por guerra y de la distancia genética entre tribus ancestrales. Dado los valores modestos de la distancia genética entre esos grupos (Cap. 6), el coeficiente de beneficio a costo de la conducta altruista debe ser muy alto. Esos valores altos se consiguen gracias a la extinción selectiva de grupos por la guerra, estimada empíricamente (Cap. 6). Dos componentes adicionales sirven

[2] HARDIN, G. The tragedy of the commons. *Science*, v. 162, p. 1243-1248, 1968.

de auxilio para reducir la desventaja en aptitud de los A frente a los N: una institución ancestral de compartir alimento soportada en motivos egoístas/ prudenciales y concebida como biológicamente egoísta; y una asociación entre altruistas dentro de los grupos o tribus, fenómeno que B & G llaman "segmentación", probablemente para recalcar el hecho de que, además de la estructura en grupos modestamente diferenciados de la población humana global, estos grupos a su vez están estructurados en su interior, lo cual aumenta la asociación positiva entre altruistas, para beneficio evolucionario del rasgo altruista (Cap. 7, p. 125). Los dos componentes restantes se entienden también como complementarios, a saber, normas e instituciones que promueven los bienes públicos (BP) y sancionadas con castigos (Cap. 9); y emociones morales como culpa y vergüenza, que incrementan la eficiencia de los castigos, pues una pequeña inversión en castigos moviliza esas emociones y produce una tasa alta de contribuciones al BP (Cap. 11).

Todos los componentes son importantes, pero prestaré aquí atención especial a la extinción selectiva de grupos y a las normas sancionadas con castigo altruista (estrategia C) para promover la provisión de BP, porque esos mecanismos cargan un peso mayor en el argumento a favor del carácter biológicamente altruista de la cooperación. El argumento tendría su mayor poder de convencimiento si estos dos mecanismos evolucionarios estuvieran apropiadamente coordinados. Sorprendentemente, los autores los presentan en dos modelos que explican, con independencia mutua, la evolución de la cooperación biológicamente altruista en la provisión de BP. Esa independencia es un problema, pues el modelo de normas y castigos para la provisión de BP no es, en mi opinión, un modelo de SG, aunque los autores así lo presenten. Y si las preferencias sociales evolucionan en el modelo de normas y castigos sin la necesidad de SG, eso es una seria objeción a la tesis del carácter biológicamente altruista de la cooperación humana, en particular porque el modelo de normas y castigos pretende recoger la evidencia empírica a favor de las preferencias sociales que se deriva de experimentos económicos (Cap. 3). Con todo, este contratiempo es bienvenido y apunta a una teoría alternativa, a saber, que las preferencias sociales y las emociones que sustentan la cooperación pueden

evolucionar sin la ayuda de la SG y, felizmente, sin atribuir un rol causal considerable a guerras las inter-tribales entre nuestros ancestros.

Además de la doble tesis descrita hasta aquí: 1. Los humanos cooperan sobre la base de preferencias sociales, y 2. Estas preferencias son biológicamente altruistas y evolucionan por SG, el libro contiene una tesis adicional 3. Las preferencias sociales no pueden evolucionar por selección en encuentros diádicos (reciprocidad directa e indirecta). B & G defienden firmemente la tesis que, dondequiera que una estrategia cooperativa evolucione por el beneficio a individuos en lugar de a grupos, el mecanismo psicológico que evoluciona excluye preferencias sociales o altruistas a favor de preferencias ego-céntricas (*self-regarding*) (p. 60, 76). Esta manera de conectar el egoísmo evolucionario o biológico con la motivación egoísta se atribuye también a Trivers (1971)[3], con lo cual "altruismo recíproco" aparece así necesariamente como un equívoco (p. 52). Más adelante argumento que B & G malinterpretan a Trivers (1971).

A pesar de los esfuerzos serios de los autores por probar las tesis 2 y 3, este reseñista no se da por convencido. La idea de que la reciprocidad fuerte incurre en un costo biológico neto (altruista), ha sido convincentemente criticada en Guala (2012)[4]. Los experimentos económicos muestran que la cooperación humana descansa sobre preferencias sociales y emociones morales, que subyacen a la cooperación y al castigo contra desertores en dilemas sociales. Pero los experimentos en el laboratorio no pueden demostrar que los casos de castigo "en el campo" son biológicamente costosos para la estrategia C que los aplica. De hecho, los reportes antropológicos que describen las reglas sociales en pequeños grupos de cazadores y recolectores no aportan evidencias en favor de castigo costoso o altruista en el campo. Lo que los mismos autores dicen en su breve reseña de la evidencia antropológica en la Secc. 6.4 no sugiere que el castigo sea costoso para C. En particular, el acoso en grupo a los ofensores y la incitación a sus familiares a asumir la aplicación de los castigos, son mecanismos ingeniosos, propios de esas sociedades pequeñas, para eliminar

[3] TRIVERS, R .The evolution of reciprocal altruism. *Quarterly Review of Biology*, v. 46, n. 1, p. 35-57, 1971.

[4] GUALA, F. Reciprocity: weak or strong? What punishment experiments do (and do not) demonstrate. *Behavioral and Brain Sciences*, v. 35, p. 1-59, 2012.

los costos relativos del castigo. Con esos mecanismos se busca impedir el surgimiento del problema del gorrón (*free-rider*) de segundo orden, que coopera pero se evita los costos del castigo. La defensa del carácter biológicamente altruista del castigo de los desertores empeora con el modelo computacional y matemático presentado por los autores (cap. 9). Argumento más adelante que no es un modelo de SG y que no presenta al castigo contra desertores como biológicamente altruista, a pesar de que B & G así lo creen. Por eso es razonable explorar una concepción alternativa, según la cual las preferencias sociales y las motivaciones altruistas evolucionan como mecanismos próximos de comportamientos como el altruismo recíproco, que los autores no ven como casos de altruismo biológico. La teoría biológica actual contiene profundos desacuerdos sobre lo que cuenta como altruismo biológico (ROSAS, 2010)[5]. Pero como no puedo entrar aquí en esas controversias, presentaré la concepción alternativa como aquella según la cual las preferencias sociales, que sustentan la escala única de la cooperación humana, evolucionan por medio de mecanismos biológicos que no son casos de SG en ningún sentido aceptado.

Preferencias sociales

El *homo oeconomicus* de la teoría económica clásica es un agente quien se interesa sólo en sí mismo, es decir, no da ningún peso en sus deliberaciones, o sólo un peso instrumental, al efecto positivo o negativo de sus acciones sobre el bienestar de otras personas. Tener una preferencia social es tener lo que precisamente el *homo oeconomicus* no tiene: ''Las preferencias sociales incluyen un interés, positivo o negativo, en el bienestar de otras personas, así como el deseo de cumplir con normas éticas'' (p. 3). Si obtienes satisfacción al cooperar con cooperadores o crees que es tu deber hacerlo, tienes preferencias sociales. Si crees que es tu deber castigar a los desertores en dilemas sociales o te satisface hacerlo, tienes preferencias sociales. Y también las tienes si te sientes culpable o

[5] ROSAS, A. Beyond inclusive fitness? On a simple and general explanation for the evolution of altruism. *Philosophy and Theory in Biology*, v. 2, 2010.

avergonzado por haber desertado con quienes han cooperado contigo. Sostener que los humanos tenemos preferencias sociales es equivalente a sostener que somos agentes morales y no sólo agentes auto-interesados, y que la moralidad no se puede reducir al auto-interés.

Las preferencias sociales permiten coordinar las estrategias en dilemas sociales para obtener resultados óptimos. Cuando se analiza la estructura de una interacción entre dos jugadores, es relevante tener en cuenta si los jugadores tienen o no preferencias sociales. Si a los jugadores sólo les importa su propia paga, el equilibrio de Nash es la mutua deserción en un DP de un solo período. Pero si se interesan por la paga de otros, la cooperación mutua es un equilibrio de Nash en el mismo juego, que se transforma por ello en un juego de seguridad (*assurance game*), también conocido como "caza del ciervo": cada jugador prefiere cooperar si tiene la seguridad de la cooperación del otro (p. 12).

El capítulo 3 revisa una multitud de experimentos que proporcionan evidencia de preferencias sociales en los humanos. Por ejemplo, sujetos castigados responden a castigos simbólicos, es decir, castigos que no acarrean costos materiales para el castigado. Al responder al castigo simbólico, los castigados muestran que no los motiva primariamente un interés en reducir costos. Más bien, los castigados "intentan reparar un daño ante los otros miembros del grupo" (p. 31). Los experimentos donde un observador no involucrado en el juego, es decir un tercero, castiga al desertor, muestran que hay castigo aunque el castigador no espere beneficios de su acción. Es interesante que en estos experimentos el tercero observa un juego *diádico* y lo concibe como regido por normas de cooperación; un hecho importante sobre el que regresaremos en la sección 4.

Cuando los escépticos respecto de la existencia de preferencias sociales discuten evidencias de este tipo, señalan que en la vida real (a diferencia del laboratorio) el comportamiento en cuestión se presenta en contextos en los que la reputación está en juego y el castigo trae beneficios a C. Esto es correcto, pero sólo cuestiona los costos objetivos del comportamiento, es decir, el altruismo biológico. La tesis en favor de las preferencias sociales no es una tesis sobre costos biológicos netos, sino sobre los motivos e intenciones del agente, el lado psicológico del

comportamiento. La posibilidad interesante, ignorada por los autores y por algunos críticos, es que una preferencia social altruista pueda tener, incluso dentro de los confines de un mismo grupo, una utilidad biológica positiva y mayor a la obtenida por una preferencia egoísta y calculadora.

B & G sostienen que "si el único mecanismo que condujo a la evolución de la cooperación hubiese sido el altruismo recíproco o la señal costosa (*costly signaling*) los motivos próximos para ayudar y los procesos cognitivos que los activan se habrían derivado [de la representación, AR] de beneficios personales" (p. 76). Nuestro principal objetivo en esta reseña es cuestionar la concepción que vincula las preferencias sociales al altruismo biológico y busca mecanismos especiales de SG para explicar su evolución. B & G se adhieren a la tesis de que los motivos desinteresados sólo pueden evolucionar por mecanismos que implican SG y particularmente favoritismo intra-grupal y hostilidad y guerra inter-grupal. El castigo que se implementa en la reciprocidad directa e indirecta negando o retirando la cooperación a los desertores, es motivado según B & G de manera egoísta y difiere por ello del castigo en la reciprocidad fuerte: "La retaliación contra desertores con el expediente de retirarles la cooperación puede forzar a los individuos auto-interesados a cooperar. Esta literatura culmina en los teoremas folk [de la teoría de juegos, AR]" (p. 80). Pero como veremos más adelante, es sorprendente que B & G sostengan que las preferencias egoístas pueden explicar la cooperación en interacciones diádicas repetidas del tipo de la reciprocidad directa. Su exposición y discusión de los "teoremas folk" en el Cap. 5 concluye con la tesis que la cooperación en los juegos diádicos repetidos – y por supuesto también en juegos de bienes públicos – no puede darse sin las normas sociales (aquellas que implican preferencias sociales). Haré un breve análisis de esa discusión en la última sección. Pero antes, debemos examinar y comparar sus dos modelos independientes para la evolución del altruismo biológico en humanos.

¿Es la cooperación humana realmente altruista en sentido biológico?

Los autores sostienen que la cooperación humana apoyada psicológicamente en preferencias sociales, en especial cuando se trata de

la provisión de BP, es biológicamente altruista y evoluciona como tal. Esto implica que su explicación evolucionaria debe recurrir a modelos de SG. Los autores presentan dos modelos de SG en los Cap. 7 y 9 respectivamente. Voy a presentar y discutir ambos modelos en esta sección. Pretendo mostrar que el modelo de normas sancionadas con castigos (Cap. 9) no convence como modelo de SG y no permite defender la tesis de que la cooperación humana es biológicamente altruista.

El modelo desarrollado en los Cap. 6 y 7 presenta la evolución de la cooperación altruista gracias a la extinción selectiva de grupos en guerras inter-tribal ancestrales. Llamemos a este modelo el *modelo-E*. El *modelo-E* hace esfuerzos novedosos por ser empíricamente relevante. Usando datos arqueológicos y etnográficos, incorpora estimaciones empíricas tanto de la distancia genética como de la ocurrencia de conflicto letal entre grupos ancestrales. La distancia genética entre grupos mide el grado de asociación (*assortment*) entre los A. La asociación entre los A determina que ellos reciban los beneficios del altruismo más que los N. Las estimaciones empíricas muestran que la distancia genética entre grupos es pequeña (Tabla 6.1, p. 100). Es decir, la asociación entre A no es lo suficientemente grande para que evolucione el altruismo por selección de parentesco (p. 102). Pero los modelos de SG también necesitan valores significativos de asociación. Valores reducidos o mínimos implican que el altruismo sólo puede evolucionar si el coeficiente de beneficio a costo del altruismo es extraordinariamente elevado (p. 102). Los valores estimados de asociación en bandas de cazadores y recolectores indican que los beneficios deben ser aprox. 15 veces los costos (p. 76). Los autores defienden que estos coeficientes extraordinarios se dan gracias a la competición a nivel de grupos en la guerra inter-tribal, que ellos también estiman empíricamente con datos arqueológicos sobre porcentajes de muerte violenta (evidencias óseas de muerte por proyectiles o armas corto-punzantes) (Tabla 6.2, p. 104). Grupos en donde prevalecen los A tienen mayor probabilidad de ganar conflictos letales entre grupos. En el modelo que construyen, grupos con predominio de A sobre N ganan y se toman el territorio de los vencidos, doblando su número. Los autores calculan entonces cuál es el costo del altruismo que le permitiría aún evolucionar. Como ejemplo, para

algunos parámetros que caen bajo el rango estimado, podría evolucionar un altruismo cuyo costo haría que, en ausencia de SG, una población con 90% de A quedase con sólo el 10 % de A en 150 generaciones (p. 123). Es justo decir que los autores, a pesar de construir su *modelo-E* con la guerra inter-tribal, también contemplan las posibilidad de que la extinción se produzca por el fracaso de grupos de N para enfrentar exitosamente (mediante cooperación) las frecuentes crisis ambientales a las que estuvieron sometidos nuestros ancestros (p. 97-98, p. 146-147).

En el *modelo-E* el castigo no juega un papel en la evolución de la cooperación altruista. Cuando comienzan a construir su segundo modelo (Cap. 9), que contiene la estrategia C (castigadores), los autores se refieren al *modelo-E* como un modelo de SG para el "altruismo indiscriminado" (p. 148). El A no discrimina entre cooperadores y desertores y no vincula su altruismo a la reciprocidad o al castigo. Sin embargo, el *modelo-E* se ayuda con la inclusión de instituciones como la de compartir alimento, que reduce la diferencia en aptitud entre A y N. Los autores defienden que esa institución no habría necesitado apoyarse en preferencias sociales o en castigos (p. 130). Pero su breve revisión de los reportes etnográficos acerca de bandas de forrajeadores, muestra que en ellas la institución de compartir el alimento se apoya en la psicología del reproche público: la tacañería se castiga primero simbólicamente, indicando que la receptividad psicológica propia de las preferencias sociales ya tiene que estar activa en quienes no comparten de acuerdo a las expectativas del grupo (p. 107-109). Quizás los autores quieren enfatizar que esas instituciones y los modos de hacerlas cumplir no son casos de altruismo *biológico.* Me uno decididamente a ese veredicto, pero insisto que se trata de comportamientos apoyados en preferencias sociales; y por tanto, estamos ante un caso claro en donde las preferencias sociales operan en un contexto ancestral que no es biológicamente altruista. Esto contradice la tesis 3. de los autores, formulada en la primera sección.

En el modelo desarrollado en la Cap. 9, el castigo de los desertores es esencial para la evolución de la cooperación; llamémoslo el *modelo-C*. El *modelo-C* es crucial para el argumento del libro, porque modela nada menos que la evidencia empírica proveniente de experimentos económicos

con sujetos humanos en dilemas sociales. En ellos se pone de manifiesto que la cooperación alcanza niveles altos gracias a disposiciones a cooperar y a castigar a desertores. Esas disposiciones expresan preferencias sociales, pues los C castigan en el laboratorio aunque sepan que sus costos no serán compensados por interacciones futuras. La disposición que lleva a C castigar se apoya en el deseo intrínseco de hacer cumplir normas de cooperación (a diferencia del deseo de beneficiarse de tal cumplimiento). En el *modelo-C*, además, los motivos punitivos y las conductas que ellos activan acarrean un altruismo biológico, es decir, reducen la aptitud de C (p. 20). La estrategia de la reciprocidad fuerte defendida por B & G es al mismo tiempo psicológica- y biológicamente altruista y no podría haber evolucionado sin SG. El *modelo-C*, por tanto, pretende ser un modelo de SG, en donde la estrategia C evoluciona e invade desde frecuencias bajas en una población sin la ayuda de la extinción selectiva de grupos, aunque el *modelo-C* también supone que la población esta estructurada en grupos. Dado que la distancia genética entre grupos no puede ser mayor en el *modelo-C* que lo que se estimó para el *modelo-E*, los lectores pueden empezar a sentir curiosidad: cómo puede evolucionar la cooperación biológicamente altruista sin los coeficientes extraordinarios de beneficio a costo proporcionados por la extinción selectiva?

En lo que sigue, explico el *modelo-C* con la intención de convencer a los lectores de que, a pesar de lo que piensan su propios creadores, no es realmente un modelo de SG. Como ocurre en el *modelo-E*, hay dos estrategias en la población: los N que no contribuyen a los BP salvo que sean castigados, y los cooperadores A que contribuyen espontáneamente a los BP. Pero como los A están dispuestos a castigar a N y esa disposición es esencial para su evolución, la estrategia A en el *modelo-C* es en realidad la estrategia C. N y C juegan un juego de BP iterado en grupos, donde C castiga y fuerza a N a cooperar. Pero el castigo se aplica de manera coordinada, sólo cuando el número de C en cada grupo alcanza un quórum predefinido. El castigo coordinado es una forma inteligente y eficiente de castigo que refleja las prácticas reportadas en la etnografía de sociedades de pequeña escala. En la primera ronda del juego, los C emiten una señal pública a un costo que anuncia su disposición a castigar. Si el número de señales alcanza o

supera el quórum, los C cooperan, y castigan a los N que desertan en esa ronda; de lo contrario los C ni cooperan ni castigan. El costo de la señal marca su desventaja en aptitud frente a N. N coopera en todas las rondas posteriores a la ronda en que es castigado. En el modelo analítico, los autores muestran que la población evoluciona hacia uno de dos equilibrios estables, dependiendo de los valores para el quórum de castigadores y en la frecuencia y distribución de C en la población. Los dos equilibrios estables son, o todos N, o un polimorfismo estable de N y C (Figura 9.1, p. 152).

En las simulaciones que se basan en el modelo analítico, la población inicia en un estado en que todos los agentes son N. Los grupos tienen en promedio 30 individuos. El quórum de castigadores es de 6 en todos los grupos. El quórum se alcanza por azar: hay mutación azarosa de N a C y también migración azarosa, de manera que, por azar, 6 castigadores se encuentran en algunos grupos. Este proceso azaroso puede llevar a la población al punto en donde las zonas de atracción de ambos equilibrios estables se encuentran en una frontera, el "filo de la navaja", que representa un equilibrio inestable. La población puede o no cruzar el filo hacia la zona de atracción del equilibrio polimórfico. En esa zona los C tienen, en promedio, una ventaja selectiva sobre los N. Si la población cruza y entra a la zona de atracción del equilibrio polimórfico, la ventaja de los C sobre los N se incrementa primero, y luego decrece hasta alcanzar el punto en que C y N tienen la misma aptitud y permanecen en un equilibrio polimórfico. Es conveniente simplificar su *modelo-C* eliminando de la primera ronda la fase de emisión de una señal pública. En su lugar, los C cooperan en la primera ronda de BP y esa acción costosa sirve de señal de su disposición a cooperar. C coopera en la primera ronda a un costo **c**, publicando así su disposición a castigar. Pero C castiga sólo si observa que el número de señales (acciones cooperativas) alcanza o supera el quórum. Si no hay quórum, C no coopera en rondas subsiguientes y tampoco castiga. N no coopera si no es castigado, pero si lo es coopera en todas la rondas que siguen al castigo. Todos, N y C, se benefician **b/n** de cada acto de cooperación, donde **n** es el número de individuos en un grupo. Cuando la población alcanza estocásticamente el "filo de la navaja", la ventaja selectiva de

C sobre N se debe a los grupos donde el número de C está en o por encima del quórum. En estos grupos hay una ronda donde los N son castigados por primera vez y C tiene una ventaja dada por

$$p - c - k/n_c^2$$

El primer término es el costo de ser castigado **p**, que es pagado por los N cuando el quórum de C es alcanzado; el segundo término es el costo de cooperar **c**, pagado por los C; y el tercer término es el costo de castigar **k**, que es compartido por el número de castigadores $\mathbf{n_c}$ y pagado con la probabilidad $\mathbf{1/n_c}$ de que la amenaza de castigo no disuada y tanto el castigador como el castigado paguen costos (p. 150). Según los valores del modelo, donde $\mathbf{p} = \mathbf{k} = 0.015$ and $\mathbf{c} = 0.01$, C tiene una ventaja selectiva sobre N en los grupos donde hay castigo cuando $\mathbf{n_c} \geq 3$. En la simulación esto se satisface en todos los grupos donde se castiga, pues el quórum para castigar se fijó en $\mathbf{n_c} = 6$. Esta ventaja de C sobre N sólo se da en una ronda, en todas las rondas subsiguientes C y N empatan pues ambos cooperan y ya nadie castiga ni es castigado. En los grupos en donde no hay quórum C tiene una desventaja de **c**, que es el costo de cooperar o contribuir al BP, pagado en la ronda en la que C anuncia su disposición a castigar (en el modelo simplificado se hace cooperando).

B & G dicen que C es altruista tanto en grupos por debajo como en grupos por encima del quórum. En los grupos que están *por debajo* del quórum, C paga **c** en una ronda, y luego deserta. C es altruista. Pero en los grupos que están *por encima* del quórum, a pesar de que todos los C pagan $\mathbf{k/n_c^2} + \mathbf{c}$ en una ronda, no es verdad que C sea altruista. Supongamos una población en donde $\mathbf{n_c} > 6$ en todos los grupos. En todos los grupos, algunos C pagan costos de castigo innecesariamente, porque la cooperación de los N puede asegurarse con menos castigadores. Así las cosas, parece que algunos C se beneficiarían cambiando a N. Sin embargo, los C que pagan innecesariamente costos de castigo tienen una ventaja en aptitud sobre los N en todos los grupos donde N es castigado, dada por $\mathbf{p - c - k/n_c^2}$ en una ronda, asumiendo que hay una ronda en donde esos C que no necesitarían castigar cooperan y castigan. Pero esto sucede solo si muchos C migran simultáneamente a un grupo que estaba sin quórum y sobrepasan

el quórum. En la ronda siguiente a su entrada, los recién llegados cooperan y todos los C del grupo castigan porque hay quórum. Pero si varios C entran a un grupo que ya tenía quórum, estos recién llegados cooperan pero no castigan, pues en el grupo ya todos los N cooperan. Además, hay que tener en cuenta que un C que cambia a N sólo se beneficia del ahorro en costos de castigo si coopera inmediatamente, pues en su grupo por hipótesis hay castigo para los que no cooperan. Si no coopera y es castigado debe pagar **k** – el costo de ser castigado – que es más costoso que el costo de castigar dado por $\mathbf{k/n_c}^2$. Intuitivamente, este no es un modelo de SG. Aunque se puede decir que algunos C están malgastando recursos y por tanto aptitud, esos C superan, con todo y su malgasto, la aptitud de los N de su grupo. Un rasgo no se dice altruista por el hecho de malgastar recursos, si a pesar del malgasto supera en aptitud a los desertores del grupo. En efecto, la única razón por la que el castigo es altruista es el problema del gorrón de segundo orden, donde los C son explotados por los cooperadores que no castigan. Pero esto no sucede en el *modelo-C* de B & G, porque allí los C aplican el castigo en grupo y de manera coordinada para hacerlo eficiente. B & G son conscientes de ello (p. 156), pero inexplicablemente no sacan las consecuencias obvias. Al parecer fueron víctimas de la definición de altruismo con la que operan: un rasgo es altruista si aumenta su aptitud dejando de ser altruista (p. 153, 161). Esta definición captura muchos casos, pero da algunos falsos positivos, como lo muestra aquí el caso de los C que malgastan recursos en castigo innecesario. Una definición más precisa podría construirse con la idea de asociación positiva entre altruistas: la sección 4.8 de su libro apunta en esa dirección (ver también ROSAS, 2010). Pero esta manera de definir altruismo arroja serias dudas sobre el concepto estándar de SG que manejan los autores. En todo caso, C no es altruista en un modelo en donde C castiga sólo si el grupo alcanza un umbral de Cs que asegura que los beneficios del castigo superen los costos, y donde los C no compiten con gorrones de segundo orden. En este caso no se cumple que, *dentro* de los grupos con castigo, los N superen en aptitud a los A (que son siempre C), que es lo que exige el concepto estándar de SG.

¿Por qué no invaden los C hasta fijarse en la población y llegan más bien a un equilibrio polimórfico con N? La razón es que siempre quedan algunos grupos en donde C no alcanza el quórum. En esos

grupos C no castiga y tiene menor aptitud que N. En sus simulaciones esos grupos son responsables de una tasa de deserción del 15% (p. 159), pues en ellos C nunca castiga y N nunca coopera. En esos grupos C es altruista, pero el número de C no aumenta en la población en virtud de lo que allí sucede. No hay un proceso de SG que lleve a los C al "filo de la navaja". El proceso que los lleva es el azar; y una vez cruzado el filo, los C aventajan en promedio en aptitud a los N, en virtud de que los aventajan dentro de los grupos donde hay castigo.

Preferencias sociales y altruismo recíproco

El *modelo-C* es un modelo en donde un comportamiento cooperativo biológicamente egoísta está motivado, según lo que muestran los experimentos económicos, por una psicología altruista de preferencias sociales. Es por tanto un mero prejuicio seguir sosteniendo que las motivaciones altruistas no pueden evolucionar por un mecanismo biológicamente egoísta. Este prejuicio se expresa en varios pasajes del libro. Algunos son directos (p. 52, 76) mientras que otros son ambiguos. Un ejemplo de ambigüedad se da en su discusión de Tit for Tat. Cuando Tit for Tat está rodeado de sus semejantes, Tit for Tat no es altruista, "porque maximiza la utilidad esperada del actor" (p. 60). La observación es exacta en sentido biológico, porque si los vecinos interactuantes son Tit for Tat, la paga de un Tit for Tat es mayor que la de un desertor en ese contexto, *independientemente de cuál sea su móvil psicológico.* Pero inmediatamente los autores dicen que si un Tit for Tat está rodeado de desertores, entonces sus móviles psicológicos no pueden ser egoístas, sino altruistas (p. 60). Aquí B & G asumen que si un Tit for Tat sacrifica recursos (aunque sólo en la primera movida) en un vecindario de desertores, debe estar motivado por un altruismo psicológico. Esto sugiere que cuando niegan que Tit for Tat es altruista en un vecindario dominado por Tit for Tat, lo están diciendo en sentido psicológico y no sólo biológico. Sin embargo, una lección que deberíamos sacar de su discusión en el Cap. 5 del teorema folk, es que la existencia de estrategias cooperativas en un dilema de prisioneros iterado no nos permite inferir motivos egoístas en

los agentes, salvo que la sombra del futuro sea indefinidamente larga, la información sea perfecta y los agentes sean perfectamente racionales. Pero estas condiciones no se dan en la vida real. Por ello no es posible inferir el egoísmo psicológico en interacciones repetidas en las que, objetivamente, cooperar da más utilidad que desertar. Esto, recordemos, pone en duda la tesis 3 defendida por los autores (ver Sección 1.)

El argumento de los autores en el Cap. 5 es que el teorema folk prueba la existencia de múltiples equilibrios de Nash para estrategias cooperativas que obtienen mayor utilidad que la deserción mutua en juegos iterados con información imperfecta o privada; pero el teorema no da "ninguna razón para creer que los jugadores se pueden coordinar en alguno de los muchos equilibrios posibles demostrados por el teorema" de modo que esos equilibrios son "evolucionariamente irrelevantes" (p. 87). El peor caso es cuando la información es privada, "porque los jugadores no concuerdan siquiera sobre lo que sucedió en el pasado y no pueden coordinar su conducta" (p. 89). Los autores señalan que los equilibrios evolucionariamente relevantes para favorecer la cooperación se alcanzan por medio de normas e instituciones "que evolucionaron a lo largo de milenios por ensayo y error" (p. 91). Estas normas son impensables sin suponer compromisos "con el cumplimiento de estándares de conducta altruistas y éticos" (p. 92). La conclusión de esto es que no es posible explicar la cooperación humana sin apelar a las preferencias sociales. Este resultado vale tanto para interacciones diádicas iteradas como para juegos de BP, en los que el número de jugadores es muy superior a 2.

Pero a pesar de este resultado, los autores nos sorprenden sosteniendo decididamente que la reciprocidad directa e indirecta en interacciones diádicas lleva y ha llevado en el pasado a la evolución de conductas cooperativas basadas en motivaciones egoístas y totalmente ajenas a las preferencias sociales. Pero su propia discusión del teorema folk nos enseña que una psicología egoísta no podría sustentar la cooperación, ni siquiera en interacciones diádicas, si la información es imperfecta o privada. Precisamente esta misma tesis se encuentra ya en Trivers (1971), por razones muy similares a las que los autores esgrimen en su discusión del teorema folk. Agentes psicológicamente egoístas son los que Trivers llama "tramposos sutiles" (1971, p. 51), es decir, agentes que "inician

actos altruistas desde una disposición calculadora más bien que generosa". Ellos no son confiables, como sí lo son los agentes movidos por la generosidad o por un sentido de equidad. "La selección [natural, AR] promueve la desconfianza hacia quienes actúan de manera altruista sin las bases emocionales de la generosidad o la culpa, porque sus tendencias altruistas son menos confiables a futuro" (TRIVERS, 1971, p. 50-51)[6].

De hecho, la experiencia común enseña que los individuos egoístas tienden a usar trucos para explotar a los cooperadores en interacciones diádicas; manipulan la información para ocultar su carencia de generosidad o equidad y su carácter calculador y egoísta; fingen emociones que no tienen; consistentemente bloquean y distorsionan los canales que procuran convertir la información privada en pública; y manipulan la información sobre la sombra del futuro, ocultando a sus contrapartes que secretamente han decidido el último período de la interacción, en el cual darán su golpe tramposo. Después de eso desaparecen. En el argumento de Trivers, los egoístas que hacen trampa y engañan impiden que haya una inferencia del egoísmo a la cooperación en interacciones diádicas iteradas. Debido a que agentes enteramente egoístas (sin preferencias sociales) no tienen razones para ser veraces, el altruismo recíproco sólo es posible si las preferencias sociales han evolucionado ya en este contexto. En suma, la cooperación evoluciona en interacciones diádicas no gracias a motivos egoístas, sino sólo si esas interacciones han moldeado la psicología humana para albergar también preferencias sociales. Toda cooperación humana debe concebirse basada en las normas que son facilitadas por la preferencias sociales. El hecho de que los experimentos muestren que las terceras partes castigan a los desertores cuando observan un juego diádico, indica que ellas ven el juego diádico como regido por normas de cooperación. No hay razón para pensar que los jugadores lo vean de otra manera. Trivers argumentó que el altruismo recíproco debía ser biológicamente egoísta para evolucionar, pero negó que pudiera apoyarse en una psicología enteramente egoísta. La emociones sociales son necesarias. Desafortunadamente, esta es una alternativa que los autores desatienden por completo.

[6] TRIVERS, R .The evolution of reciprocal altruism. *Quarterly Review of Biology*, v. 46, n. 1, p. 35-57, 1971.

Un razonamiento similar se puede aplicar al fenómeno de la reputación. B & G señalan con razón que la reputación no debe concebirse meramente como mecanismo para obtener mejores utilidades en interacciones futuras (p. 44-45). Esto corresponde al sentido psicológico egoísta de reputación, que es el cuadra con los agentes puramente egoístas – los que Trivers denomina impostores egoístas. Pero hay también un sentido altruista de reputación que es imposible sin preferencias sociales. Los autores dicen : ''A las personas les importa mucho cómo otros las evalúan, independientemente de las recompensas materiales o los castigos que se deriven de esa evaluación.'' (p. 44). Adam Smith (2000)[7], citado por B & G, lo expresó de este modo: ''Cuando la naturaleza moldeó al ser humano para la sociedad, lo dotó con un deseo original de complacer…a sus prójimos. Ella hizo que la aprobación de sus semejantes sea halagüeña y agradable **por sí misma**… (parte III, Sect. I, Paragr. 13, resaltado añadido)'' (p. 45). ¿Qué razón tenemos nosotros para negar que esta forma altruista de reputación haya evolucionado en el contexto de la reciprocidad directa e indirecta en interacciones diádicas? No veo ninguna razón pertinente.

He sido crítico con la concepción defendida por los autores porque la alternativa aquí esbozada parece más parsimoniosa. Pero aun queda mucho por decir en torno a este fascinante proyecto. Uno desearía que los autores se animen a recoger algunos de los puntos de crítica expresados aquí y los usen para seguir desarrollando su investigación, para beneficio de la investigación que busca una explicación evolucionaria de la cooperación humana.

Recibido: 20/06//2012
Received: 06/20/2012

Aprobado: 06/11/2012
Approved: 11/06/1012

[7] SMITH, A. *The theory of moral sentiments*. New York: Prometheus, 2000. (Originalmente publicada em 1759).