LA EVOLUCIÓN DE LA MORAL CONTRACTUAL

The evolution of contractual morality

ALEJANDRO ROSAS*
Universidad Nacional de Colombia

RESUMEN

Las explicaciones evolucionarias del altruismo y la cooperación humana, inauguradas por pioneros como Darwin, Hamilton y Trivers, sugieren que la biología podría eventualmente construir una explicación científica plausible de un núcleo de la moralidad humana. Según este proyecto, la moralidad y la cooperación humanas emergen cuando los recursos son escasos y no pueden explotarse por individuos aislados; y cuando los individuos no pueden mantener a largo plazo una posición de dominio sobre otros para sus fines egoístas. Filosóficamente, una pregunta importante en relación con este proyecto concierne a los conceptos de *moral* y de *motivación moral* que allí se presuponen. El proyecto evolucionario no ha hecho claridad en este punto. Se argumenta en favor de dos tesis: 1) las explicaciones evolucionarias de la cooperación sugieren una moral de tipo contractual, pero son aún ambiguas en lo que concierne a las motivaciones favorecidas por la selección natural, y reflejan, sin resolverlo, un desacuerdo tradicional entre el contractualismo moral hobbesiano (de motivación egoísta) y el kantiano (de motivación altruista); 2) esas explicaciones no pueden, en su forma actual, resolver este desacuerdo, pero una reflexión sobre el papel que desempeña la capacidad de leer las motivaciones y el carácter de otros en la evolución de la moral puede suministrar argumentos a favor del contractualismo kantiano.

Palabras clave: altruismo, contractualismo moral, ética evolucionista, teoría de la mente.

ABSTRACT

Evolutionary explanations of altruism and human cooperation, first set forth by pioneers such as Darwin, Hamilton and Trivers, suggest that biology might be capable of offering a plausible scientific explanation of the core of human morality. According to this project, morality and human cooperation arise when resources are scarce; they cannot be exploited by isolated individuals; and individuals cannot maintain a long-term position of domination over others in order to advance their selfish ends. An important philosophical question that arises with respect to this project has to do with the concepts of de *morality* and *moral motivation* that it presupposes. The evolutionary project has not been clear in this respect. The article

Artículo recibido: 05 de abril de 2011; aceptado: 05 de julio de 2011

* *arosasl@unal.edu.co*

argues in favor of two theses: 1) evolutionary explanations of cooperation suggest a contractual type of morality, but they are ambiguous regarding the motivations favored by natural selection, thus reflecting, without resolving it, a traditional disagreement between Hobbes's moral contractualism (selfish motivations) and that of Kant (altruistic motivations); 2) in their current form, these explanations cannot resolve that disagreement, but a reflection on the role of the capacity to interpret the motivations and character of others in the evolution of morality could provide arguments in favor of Kantian contractualism.

Keywords: altruism, moral contractualism, evolutionary ethics, theory of mind.

1. Los beneficios de la cooperación

La cooperación es vista hoy como un fenómeno generalizado en el mundo biológico. Esta ha sido la posición dominante desde que los biólogos se dieron cuenta de que los primeros organismos multicelulares, y probablemente también los primeros genomas, resultaron de la agrupación de unidades –genes y células individuales– que antes subsistían independientemente y que se agruparon para conformar individuos a un nivel superior, porque cooperando sobrevivían mejor que aislados (Buss; Maynard Smith y Szathmáry).Nace así una perspectiva nueva sobre la evolución, en la cual el énfasis en el papel de la cooperación complementa y pone límites al énfasis tradicional darwiniano sobre la competencia y la lucha por la existencia.

Sin embargo, la cooperación plantea un problema si se concreta mediante comportamientos altruistas, en el sentido biológico del término, es decir, en comportamientos que implican una transferencia de beneficios de un organismo donante a uno receptor, de modo que el donante reduce su aptitud biológica y el receptor la aumenta. Desde Darwin, estos comportamientos se han considerado paradójicos, porque no parece posible que la selección natural favorezca rasgos que reducen la aptitud de su portador. No obstante, la paradoja puede resolverse. Existen condiciones bajo las cuales tanto el donante como el receptor pueden sacar provecho de rasgos altruistas o cooperativos. Trivers cimentó esta posibilidad para el caso de interacciones diádicas (entre dos individuos): si ambos individuos interactúan repetidamente y alternan los roles de donante y receptor, el resultado para ambos será positivo justo cuando el costo para el donante de su acto altruista sea menor que el beneficio para el receptor. Por ejemplo, supóngase que *A* salve a *B* de ahogarse a un costo muy bajo, por ejemplo, extendiéndole desde la orilla de un río torrentoso una rama suficientemente larga. El beneficio para *B* es la vida misma, mientras que el costo para *A* es una pequeña inversión de tiempo. Si la situación se repite con los roles cambiados y *B* reciproca con un acto altruista similar, entonces

tanto *A* como *B* obtienen del intercambio un beneficio que no podrían obtener aisladamente.

La posibilidad de obtener beneficios mediante la cooperación es similar cuando se trata de bienes públicos en grandes grupos. En este caso el esfuerzo colectivo es más productivo que la suma de esfuerzos individuales aislados. Por ejemplo, ante el ataque de un grupo enemigo, cada miembro del grupo atacado podría emprender una defensa solitaria, pero un esfuerzo coordinado de defensa grupal es más efectivo, no sólo para el grupo sino para cada individuo particular. La cooperación, entonces, en el sentido de una acción colectiva y coordinada, produce beneficios que exceden la suma de las acciones individuales. La idea es que en interacciones tanto diádicas como de *n*-personas (donde $n > 2$), la acción reciprocada o coordinada produce para cada individuo beneficios superiores a los que se pueden obtener aisladamente. Visto desde la perspectiva de la evolución, la cooperación evoluciona cuando la selección natural favorece rasgos conductuales que enlazan las acciones de la forma requerida.

II. Cooperación contractual

Sin embargo, el mero hecho de que un individuo gane cooperando más de lo que gana sin cooperar no garantiza que la cooperación evolucione. Tanto en las interacciones diádicas como en las de *n*-personas, la estructura de la interacción es a menudo un *dilema de prisionero*. Es bien sabido que en un dilema de prisionero de un solo período la cooperación no es ni racional ni adaptativa; aunque los jugadores se beneficien de la cooperación mutua, cada uno gana más si escatima la contribución propia y se beneficia de las contribuciones de otros. Se trata del problema clásico de la provisión de bienes públicos. Existe una tentación de vivir de la contribución de los demás, sin aportar lo propio. Hume se ingenió una hipótesis sobre el origen de las promesas que se basa precisamente en esa tentación. Un individuo enfrentado a la posibilidad de cooperar con otro en una interacción diádica podría razonar de este modo: "me gustaría poder ayudarte hoy con la expectativa de que tú me ayudes mañana, pero necesito que me des una señal clara de que reconoces mi expectativa y que es tu intención satisfacerla (una promesa). Así, si te sientes tentado a engañarme y de hecho lo haces, el haber proferido una promesa (una fórmula verbal) me permite castigarte dañando tu reputación por tu falta de palabra" (*cf.* THN, III, II, V).

El castigo es crucial en la evolución de la cooperación en los dilemas de prisionero. Si los jugadores interactúan repetidamente en un dilema de prisionero, el jugador *A* puede castigar al desertor *B* respondiendo a su vez con deserción en el período siguiente. *B* podría

entonces cooperar en períodos subsiguientes, pues de lo contrario perdería la oportunidad de beneficiarse en el juego repetido. Trivers y después Axelrod y Hamilton argumentaron que la interacción repetida entre los mismos dos jugadores es una condición necesaria para la evolución de la cooperación como reciprocidad entre individuos no emparentados. Sin embargo, Trivers también señaló que deben cumplirse otras condiciones adicionales. Se requiere, por ejemplo, que los jugadores tengan igual poder, pues una relación de dominio no favorece la cooperación basada en la reciprocidad: si un individuo domina a los demás, los puede forzar a aceptar acuerdos inequitativos. Esto es importante porque las ganancias distribuidas desigualmente en los modelos evolucionarios no se consideran cooperación, aunque ambos jugadores ganen. La literatura y los modelos desarrollados a partir de Trivers asumen que la cooperación implica un principio de equidad en la distribución de los beneficios. En este sentido, la cooperación debe alcanzarse sin que ninguna de las partes coaccione a la otra. Esto coincide con concepciones contractuales de la moral humana como la de Locke, o más recientemente la de Gauthier, en las que el problema de la provisión de bienes públicos y el dilema de prisionero constituyen la estructura básica de la situación humana, cuya solución exige la instauración por acuerdo de normas de equidad.

Las interacciones repetidas pueden llevar a la cooperación si ningún agente practica o puede practicar la coacción. Pero incluso si admitimos interacciones repetidas y excluimos la coacción, la cooperación podría no emerger. Si la cooperación se apoya en intereses egoístas, hay dos caracteres propios de la filosofía moral británica que aun en contextos de interacciones repetidas y libres de coacción plantean un problema a la moralidad, entendida como instauración y cumplimiento de un acuerdo contractual equitativo. Se trata del "necio" (*fool*) de la filosofía hobbesiana y del "bribón astuto" (*sensible knave*) de la humeana. En su característico espíritu egoísta, ellos plantean que la honestidad es una buena política en general, pero que hay excepciones que un egoísta sabio puede y debe aprovechar (Hume 1902 §232). Estos caracteres egoístas no están en contra de obtener beneficios unilaterales; al contrario, los prefieren y sólo renuncian a ellos si existe un peligro real de que quede al descubierto que prefieren explotar a sus contrapartes en lugar de cumplir incondicionalmente el acuerdo cooperativo. Es decir, ellos no constriñen su egoísmo por razones morales y para beneficio de todos, sino para su propio beneficio, y sólo si es estrictamente necesario. Su estrategia es mantener una apariencia de moralidad, es decir, de cumplimiento del pacto, cuando en realidad prefieren el beneficio unilateral. En este sentido recurren al engaño deliberado de sus contrapartes (Sayre-McCord). El necio

y el bribón harán lo que sea necesario para aparecer como honestos ante los demás y esto incluye actuar de manera honesta y equitativa si es preciso; pero serán deshonestos en todas las ocasiones en las que puedan serlo sin poner en peligro su reputación.

El necio y el bribón ponen de manifiesto que el egoísta racional no tiene por qué preferir el cumplimiento incondicional del pacto. Adoptan más bien la estrategia de aparentar moralidad y por tanto no se adhieren incondicionalmente al acuerdo moral. Ellos son la objeción crucial contra la idea hobbesiana según la cual un agente egoísta racional que busca sobrevivir en un mundo de enemigos potenciales es suficiente para explicar y justificar el comportamiento moral. Contra Hobbes se puede argumentar que no parece que el necio o el bribón sean irracionales. Contra Hobbes también se puede afirmar que si el egoísmo racional es el único motivo disponible para sustentar la moral, es más factible que los egoístas racionales sean como esos personajes en lugar de ser genuinamente morales. Si el necio y el bribón son lo suficientemente astutos y cuidadosos, pueden muy bien sobrevivir y competir exitosamente con individuos que se disponen a cumplir incondicionalmente la moral por acuerdo, precisamente porque vivirían de ellos como sus parásitos, explotándolos. Pero aun concediendo su éxito biológico, esos personajes mostrarían, en virtud de su carácter de parásitos, que la moral no puede apoyarse en motivos puramente egoístas (*ibid.*). Más bien, los genuinos agentes morales serían aquellos agentes explotados que cumplen incondicionalmente el pacto porque albergan un respeto a las personas como tales. Este respeto sería una motivación primitiva, que una teoría contractual alternativa adopta como irreductible al egoísmo racional (Hampton).

En todo caso, sea una moral basada en el egoísmo racional una idea viable o no, ella y la idea opuesta, que sostiene que la moral necesita de un respeto a las personas como tales, representan dos versiones distintas y plausibles de contractualismo. ¿Es posible decidir cuál es más acertada desde un punto de vista descriptivo y explicativo? Es decir, ¿cuál describe mejor la naturaleza de la moral humana? Aquí presentaré una defensa de la moral contractual de corte kantiano (haciendo un uso laxo y sin pretensiones exegéticas de este calificativo), basándome en una reflexión de tipo evolucionario. La idea básica es que los egoístas que recurren al engaño, como el necio y el bribón, sólo pueden tener éxito como parásitos, es decir, cuando en una población hay agentes morales kantianos. Si los egoístas lograran desplazar por completo a los agentes kantianos, presionando su extinción en una dinámica evolucionaria en la que se reproduzcan los más aptos, la consecuencia más probable sería la extinción de toda forma de cooperación cuando se torne evidente que todo agente es egoísta y falso.

III. Una teoría de la moral como egoísmo disfrazado

Llegamos así a las preguntas principales: ¿pueden subsistir la moralidad y la cooperación apoyadas en el egoísmo racional? ¿O es necesario suponer motivos altruistas, en el sentido de juego limpio y respeto por las demás personas? ¿Se extinguiría la cooperación si los agentes egoístas dominaran totalmente la población?

Veamos primero si la cooperación se mantiene en un mundo de agentes puramente egoístas, o si desaparece cuando estos agentes recurren a una estrategia de apariencia y engaño, como el necio y el bribón astuto. Supongamos que una población inicialmente mixta (agentes egoístas y kantianos o altruistas) haya evolucionado hacia una población de puros agentes egoístas que recurren al engaño. El uso generalizado del engaño por parte de los egoístas impide que, al menos inicialmente, se sepa con certeza si ya no hay agentes kantianos en la población, puesto que todo agente egoísta debe aparentar honestidad para atraer a los demás agentes a interacciones cooperativas, y para ello tendrá que "invertir" en algunos actos cooperativos que usará como cebo. Es interesante que algunos biólogos evolucionistas piensen que esta es una imagen acertada de las sociedades humanas (*cf.* Alexander). Según estos, excepto quizás por la existencia de motivaciones altruistas entre parientes, los humanos hemos evolucionado hacia el egoísmo: nos centramos en nuestro propio beneficio y cooperamos sólo cuando la coerción o el engaño no son posibles, y esto sólo con miras a algún acto de explotación final tan pronto como creamos que podemos hacerlo sin ser detectados. Y para asegurar que nuestra naturaleza tramposa sea eficiente, la selección natural nos ha moldeado como maestros del autoengaño (*id*. 123). Creemos ser altruistas y honestos, porque si así no lo creyésemos, fácilmente evidenciaríamos nuestra falsedad y el intento de engañar a otros fracasaría. El altruismo genuino no existe entre humanos; es sólo una mascarada que personificamos tan convincentemente, que los actores hemos llegado a creer que somos el personaje. En un mundo así será siempre verdad que en algún momento clave de la interacción entre personas aparecerá la conducta egoísta y explotadora. Y aunque por momentos haya apariencia de cooperación, como ya se mencionó, esta sólo será con miras a un acto de explotación final.

Si nuestra naturaleza social es como la describe esta teoría, se sigue que la cooperación humana descansa sobre fundamentos muy frágiles. Esta tesis dice que los humanos cooperamos sólo cuando no es factible engañar y salir airosos, mientras solapadamente esperamos mejores oportunidades para la persecución despiadada del interés egoísta. Creo que esta visión de nosotros mismos debería perturbarnos profundamente: un mundo así no es un mundo donde nos

sintamos en casa; sólo podemos soportarlo si vivimos autoengañados, es decir, si no sabemos que es así. Pero si la tesis de que el egoísmo es señor en el mundo social se difunde e impone como una verdad obvia, los pesimistas tienen todo el derecho a sus predicciones sombrías sobre el futuro de la especie humana.[1]

Esta teoría sobre el altruismo humano como disfraz del egoísmo surge en un intento de derivar nuestro talante social-psicológico profundo a partir de una teoría biológica sobre la evolución del comportamiento social. Según esta teoría, la selección natural actúa primordialmente sobre los genes y selecciona a aquellos genes que se benefician a sí mismos en grado máximo, ya sea beneficiando a sus portadores o a sus parientes cercanos. Si nuestros genes fueron seleccionados por predisponer a sus portadores a conductas que los benefician, aunque sea explotando a otros, nuestras motivaciones sociales profundas deben también ser egoístas, o al menos así razona la teoría. Nuestras intuiciones morales esconden esta verdad, porque de ese modo el autoengaño nos lleva a creer en nuestro altruismo para así poder engañar y explotar más efectivamente a los demás. La teoría supone que los humanos, tanto ahora como en el pasado ancestral, preferimos interactuar con agentes dispuestos a respetar a los demás con espíritu de equidad. Dado que somos egoístas por diseño biológico, no hay, ni ha habido nunca, altruistas con quienes interactuar. Este hecho ha tenido que ser camuflado, pues de lo contrario la cooperación entre humanos no habría siquiera despegado. La cooperación depende y ha dependido entonces de disfrazar las motivaciones egoístas, y como la conciencia de estar engañando a otros puede llevarnos involuntariamente a traicionarnos, el autoengaño es necesario para un engaño eficiente (*ibid.*). De este modo la selección natural favoreció la evolución del autoengaño. Un elemento fundamental de este es que denigramos del egoísmo y ensalzamos el altruismo. Nuestras actitudes valorativas conscientes son, así, producto del autoengaño. Si esta concepción de la naturaleza social humana nos produce malestar, este malestar es sólo una manifestación más del autoengaño. La teoría ofrece así una explicación de la intuición que otorga un valor positivo al altruismo, mientras que al mismo tiempo niega que los humanos seamos, o podamos ser, genuinamente altruistas. De este modo es neutralizado el testimonio de la intuición moral cotidiana por la tesis de un egoísmo humano fundamental.

1 Un árbitro anónimo plantea que el altruismo está garantizado así sea como autoengaño. Es importante reconocer que si la teoría de Alexander es verdad, toda cooperación humana es una fachada que esconde una agenda de manipulación y lucha por el poder. Una vez puesto esto al descubierto, es probable que la lucha solapada por el poder se torne en una lucha abierta y despiadada.

Sin embargo, esta teoría tiene una falla lógica fatal en la concepción del proceso evolucionario particular que condujo a la cooperación humana. Visto de cerca, su intento de subvertir nuestra valoración positiva del altruismo es contradictorio. En su escenario ancestral para la evolución del autoengaño, el punto de partida consiste en que los agentes se aproximan unos a otros no sólo con motivaciones inevitablemente egoístas por diseño biológico, sino también con una valoración negativa del egoísmo que amenaza con frustrar toda interacción cooperativa. Esta valoración negativa del egoísmo es fundamental en su concepción del proceso evolucionario. Pero, ¿de dónde provienen la valoración negativa del egoísmo y la positiva del altruismo *antes de* la evolución del autoengaño? Esas valoraciones evolucionarían, de acuerdo a la teoría, como una estructura subsidiaria para consolidar el autoengaño. La lógica de la teoría consiste en suponer que nuestros ancestros buscaban personas altruistas con quienes cooperar pero que no los había; y, precisamente por esa razón, tuvieron que evolucionar al autoengaño –que consiste básicamente en creer que uno es altruista y, con el fin de hacer creer a otros que uno lo es, en declarar públicamente que el altruismo es valioso–. Pero entonces, ¿por qué buscaban nuestros ancestros interactuar con altruistas al inicio del proceso si la valoración positiva del altruismo sólo fue requerida como una forma de consolidar el autoengaño? ¿De dónde viene esa valoración inicial del altruismo que hizo necesario el engaño y el autoengaño? Esto es algo que la teoría no puede explicar. La teoría termina contradiciéndose: sostiene que valoramos el altruismo como consecuencia de una necesidad profunda de autoengañarnos; pero para poder llegar a esa conclusión, describe un proceso evolucionario cuyo punto de partida es que los agentes egoístas sólo quieren interactuar con altruistas porque de algún modo los valoran. No podemos explicar la existencia de esos valores como producto del fenómeno cuya evolución los mismos agentes egoístas presionaron. Debemos buscar una explicación alternativa en donde el valor del altruismo no sea un simple producto del autoengaño.

IV. Cooperación en un mundo de agentes egoístas

Vimos que la teoría expuesta contiene una contradicción en su escenario ancestral. Para comenzar a corregir esta visión del escenario, quiero mostrar ahora cómo, dependiendo de las circunstancias, en unos casos el egoísmo puede ser un soporte fiable de la cooperación y, en otros casos, dar lugar a una estrategia de mendacidad y parasitismo moral. Lo primero se puede visualizar imaginando un mundo hipotético en donde sólo hay agentes egoístas y en el cual el engaño es imposible porque las intenciones y disposiciones de cada

quien son transparentes para los demás. En un mundo así, todos seríamos lectores infalibles de las mentes de los otros. En este caso, la intención de desertar en una situación moral sería instantáneamente conocida por los involucrados, lo que los induciría a desertar también. Pero si fuéramos todos racionales y buscáramos nuestro mayor beneficio, sería preferible formarse la intención de cooperar y además llevarla a cabo. Así se produciría el beneficio de la recompensa mutua, mientras que cualquier intención de deserción sería leída por los otros y los induciría a desertar a su vez. Ya que en un dilema de prisionero la mutua deserción es peor para ambos jugadores que la recompensa mutua por cooperar, si los egoístas transparentes fueran racionales, siempre escogerían la recompensa por cooperar. Este es básicamente el argumento de Gauthier para derivar el constreñimiento moral a partir de la racionalidad instrumental. Gauthier construye su argumento con agentes translúcidos, es decir, no transparentes del todo, por ser la translucidez más realista con respecto al ser humano que la transparencia. En un momento mostraré por qué la derivación de la cooperación a partir del egoísmo racional no funciona con agentes translúcidos. Ahora es importante notar que en un mundo de egoístas con intenciones transparentes, y asumiendo que deciden racionalmente, la cooperación sería completa. El egoísmo transparente puede producir una imitación perfecta de un mundo moral. Aquí no existiría el altruismo, pero tampoco existiría la explotación. Es más, el altruismo no sería valorado como algo precioso, ni el egoísmo sería vilipendiado; y así no existiría el sistema de valores propio de nuestra moral. Nuestras actitudes valorativas hacia el altruismo y el egoísmo se deben al hecho de que no somos transparentes y de que no siempre podemos leer correctamente las intenciones y disposiciones de los demás. En un mundo donde los egoístas son translúcidos, en lugar de perfectamente transparentes, la cooperación está destinada a desaparecer, como explico a continuación.

El punto crucial es que en una población de agentes egoístas y racionales, la translucidez da origen a una estrategia de engaño que frustra la cooperación. Imaginemos que en un mundo de agentes egoístas racionales y transparentes, donde todos cooperan, los agentes pierden su transparencia y se vuelven translúcidos. La translucidez se define en términos de la probabilidad de identificar correctamente las intenciones de los otros antes de la acción: la probabilidad de acertar es mayor que la de errar, pero hay una probabilidad de errar; es decir, los agentes no son infalibles al leer otras mentes. Es de esperarse, entonces, que los egoístas deliberadamente explotarán la translucidez para influir en la probabilidad de que sus contrapartes se equivoquen en una dirección particular: manipularán la información de mane-

ra que los demás detecten en ellos disposiciones cooperativas que en realidad no existen (Sayre-McCord). En otras palabras, la translucidez le abre las puertas a una nueva estrategia: el engaño. Nótese, sin embargo, que engañar no consiste aquí en hacer creer en motivaciones altruistas, pues esas motivaciones no existen en ese mundo de egoístas racionales. El engaño consiste en hacer creer a los demás que la intención es cooperar cuando no lo es. A medida que el engaño se propaga, se llega a un punto en que la probabilidad de identificar correctamente a alguien realmente dispuesto a cooperar cae por debajo del azar. Llegado este punto, ya no conviene cooperar cuando uno cree haber detectado intenciones cooperativas, pues la probabilidad de errar es mayor que la de acertar. Los agentes que cultivan el engaño lo hacen precisamente porque, en contra de lo que sugiere la translucidez, ellos logran engañar, las veces que lo intentan, en un porcentaje alto. Eso lleva gradualmente a que en casos dudosos nadie crea en las intenciones cooperativas de los otros; y finalmente lleva a la convicción de que toda acción cooperativa es usada como cebo, lo cual es bien plausible en un mundo de egoístas tramposos. La posibilidad de engañar es lo que acarrea la desaparición de la cooperación en un mundo de agentes egoístas. Y tan pronto cesa la cooperación, el engaño deja de tener sentido.

v. ¿Cómo emerge la valoración positiva del altruismo?

Para explicar la emergencia de nuestro sistema de valores es necesario mantener el carácter translúcido de todo agente, pero cambiar la composición de la población en proceso de evolución. Supongamos que en una población de egoístas translúcidos como la que acabamos de imaginar, en la que apenas se comienza a propagar la estrategia del engaño, emerge una proporción considerable de agentes kantianos. Como tales, están dispuestos a constreñir su egoísmo por razones de equidad y tienen una aversión natural al engaño. Una aparición súbita de agentes kantianos es biológicamente improbable y sólo podría darse en un proceso gradual, pero aquí, para visualizar una dinámica distinta a la desaparición de la cooperación que inferimos en la sección anterior, nos ubicamos en el punto en el que la población contiene ya una proporción significativa de esos agentes. En un mundo translúcido marcado por el engaño, en donde los egoístas disfrazan su carácter, un acto de cooperación es ambiguo: podría ser una jugada estratégica de un agente tramposo que finge ser cooperativo para explotar a los incautos en el momento oportuno. Pero si en la población existen agentes kantianos que cooperan por equidad y que no buscan engañar, un acto cooperativo puede efectivamente indicar la existencia de la cooperación. Su comportamiento cooperativo está respaldado por

una motivación o disposición a la equidad, que resulta entonces de mucho valor. Como a todo agente le interesa participar en los beneficios de la cooperación y evitar ser explotado por los tramposos, todos quieren interactuar exclusivamente con agentes kantianos. Sondean el carácter de sus contrapartes buscando detectar allí un principio de equidad, que es la única garantía de cooperación beneficiosa. Incluso los egoístas tramposos tienen una oportunidad de sobrevivir como parásitos en una sociedad que promueve la existencia de agentes kantianos y los valora.

Para que en este nuevo escenario los agentes kantianos tengan una oportunidad de mantenerse o de aumentar en la población, es necesario que ellos puedan discriminar y esquivar a los egoístas. Esto no es un problema si el rasgo que predispone a la equidad se liga por diseño genético a un rasgo que discrimina entre los que tienen ese rasgo y los que no. En el contexto de la selección de parientes (Hamilton), los biólogos pueden fácilmente imaginar una presión selectiva que favorezca ese ligamiento, pues en los altruistas por parentesco la conducta altruista debe ligarse y quedar condicionada a la discriminación de parientes. Si se toma la conducta altruista como el rasgo que indica el parentesco, podría evolucionar un diseño que condicione la expresión del altruismo a la detección del comportamiento altruista en otros. Este altruismo condicional se extendería luego a los no-parientes en la medida en que sean altruistas (Axelrod y Hamilton). Así se garantizaría que los beneficiarios de la cooperación sean predominantemente altruistas. Si un altruista carece de la habilidad de discriminación, es presa fácil de los egoístas y pierde aptitud biológica. De ahí que la discriminación sea condición necesaria para un altruismo adaptativo. Más precisamente, el altruismo y su reconocimiento son rasgos que se retroalimentan positivamente y constituyen una combinación evolutivamente exitosa.

El mecanismo para la discriminación no es como el que sirve en algunas especies para el reconocimiento de parientes, basado, por ejemplo, en una sola cualidad sensorial, procesada automáticamente y conectada a un acto reflejo. La discriminación entre altruistas y egoístas es, en el ser humano, una habilidad cognitiva de nivel superior y puede haber evolucionado junto con una "teoría de la mente". La existencia de agentes egoístas que ponen todo su empeño en imitar a los altruistas, pone retos al mecanismo e impulsa su refinamiento. Los egoístas tejen una trama de conductas que sugieren motivaciones y disposiciones altruistas. Como en una carrera armamentista, se requieren capacidades cada vez más finas para detectar apariencias. La existencia de ese reto social en el entorno de nuestros ancestros originó o consolidó procesos de evaluación cognitiva independientes de

conductas disparadas como actos reflejos. Es posible que así se haya impulsado la representación de acciones como procesos complejos que pueden analizarse en elementos separados y cuya consistencia puede evaluarse mientras se suspende la acción. Una evaluación de este tipo sopesa un número indeterminado de evidencias y puede demorar en dictaminar un veredicto final. Y al tiempo que la representación de la acción y de la motivación subyacente se vuelve sofisticada, la relación con la conducta es mediada por decisiones sobre el valor epistémico de una o varias conductas, como signo de un carácter equitativo y por tanto valioso. Así, la capacidad de reconocer a otros altruistas se transforma en una actitud valorativa del carácter altruista como tal.

Esta actitud valorativa emerge gradualmente en la filogenia a medida que la habilidad de discriminar altruistas de egoístas se vuelve sofisticada, consciente e independiente de conductas reflejas, es decir, a medida que la mente se vuelve más compleja. El término del proceso es que la cooperación se torna condicional y mediada por un sistema de valores. Nuestra explicación de este sistema no lo presenta como artificio del autoengaño. Ambos tipos de agentes, tanto los morales como los egoístas, valoran el altruismo porque es una cualidad que garantiza los beneficios de la cooperación. En cambio, el egoísmo se convierte en blanco de aversión, porque es el vehículo de estrategias indecentes y tramposas de interacción social. Cuando el engaño entra a jugar, una disposición hacia la justicia y la equidad es la única garantía confiable de interacción cooperativa, y por ello es altamente valorada.

En esta explicación asumimos que nuestra habilidad para leer otras mentes es imperfecta, pero suficientemente acertada como para sostener una presión selectiva (natural y psicosocial a la vez) en contra de los egoístas. Es decir, que la probabilidad de identificar correctamente a un altruista o agente kantiano es mayor que la probabilidad de errar. ¿Pero no es esto, otra vez, la translucidez que habíamos rechazado en la sección anterior? La respuesta es: no. Recordemos que rechazamos la translucidez en un mundo compuesto únicamente de agentes egoístas. En ese mundo, tan pronto como un egoísta adopta la estrategia del engaño, todos tienen una razón para adoptarla. Tras la adopción universal del engaño, la translucidez no puede ya garantizar que las intenciones cooperativas sean identificadas correctamente con una probabilidad superior al azar. Pero en una población con una proporción substancial de agentes kantianos, la translucidez puede desempeñar un papel análogo a la transparencia. La razón de ello es que los agentes kantianos no adoptarán la estrategia del engaño, precisamente por su compromiso con la equidad. La tasa de

identificaciones correctas aumenta por la sola existencia de agentes kantianos. También es de esperarse que aparezcan signos objetivos y eventualmente detectables de la diferencia entre egoístas y altruistas; así, los altruistas tendrán la posibilidad real de interactuar predominantemente con otros altruistas y de repartirse equitativamente los beneficios de la cooperación. El proceso cooperativo equivale entonces a un proceso de selección psicosocial en favor de motivaciones altruistas (Nesse; Rosas; Trivers 50-51). Como los altruistas se buscan para interactuar entre sí, porque su disposición hacia la equidad es el único signo confiable de una actitud cooperativa consistente, los beneficios de la cooperación se reparten entre ellos y así su aptitud aumenta.

He identificado dos variables importantes en el proceso selectivo que llevó a la evolución de nuestra valoración positiva de agentes altruistas o kantianos y a su selección: una de ellas, descubierta por Trivers, es la diferencia entre agentes con un principio de equidad para su conducta cooperativa y agentes que cooperan por un egoísmo calculador; la otra es la diferencia entre una capacidad infalible y una falible para inferir las motivaciones y disposiciones de los demás. Debido a que somos lectores falibles de las mentes de otros, susceptibles al engaño, no podemos confiar en egoístas racionales hobbesianos, que eventualmente adoptarán el engaño como estrategia, sino sólo en agentes kantianos, cuya conducta cooperativa se basa en el respeto a los demás y en el reconocimiento de su derecho a participar con equidad en los beneficios de la cooperación.

VI. Conclusión

Un argumento basado en el papel de la lectura de mentes al evaluar los caracteres morales y al escoger las contrapartes en una interacción cooperativa puede ayudar a resolver la ambigüedad entre dos versiones del contractualismo moral. En una teoría posible, la capacidad de leer mentes favorece la evolución del autoengaño. Nuestros valores conscientes, como el elogio del altruismo y la desaprobación del egoísmo, son productos del autoengaño, porque este nos protege de evidenciar involuntariamente nuestras motivaciones egoístas e intenciones de engaño. En la teoría que defiendo, el rol de la lectura falible de mentes favorece la evolución tanto de la motivación y el carácter equitativo, como de la capacidad de detectar ese carácter en otros con el fin de interactuar preferentemente con quienes revelen tenerlo. Esto crea una presión selectiva de tipo psicosocial en favor de lo que llamamos un agente kantiano o una disposición a la equidad que, operando en tiempo evolucionario, selecciona efectivamente este tipo de agentes.

Bibliografía

Alexander, R. The *Biology of Moral Systems*. New York: Aldine de Gruyter, 1987.

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

Buss, L. *The Evolution of Individuality*. Princeton: Princeton University Press, 1987.

Gauthier, D. *Morals by Agreement*. Oxford: Oxford University Press, 1986.

Hamilton, W. D. "The Genetical Evolution of Social Behavior I and II", *J. Theor. Biol*. 7 (1964): 1-52.

Hampton, J. "Two Faces of Contractarian Thought". *Contractarianism and Rational Choice*, Vallentyne, P. (ed.). New York: Cambridge University Press, 1991. 31-55.

Hume, D. *A Treatise of Human Nature* [1739] [THN], Selby-Bigge, L. A. (ed.). Oxford: Clarendon Press, 1896.

Hume, D. *Enquiries Concerning the Human Understanding and Concerning the Principles of Morals*, Selby-Bigge L. A. (ed.). Oxford: Oxford University Press, 1902.

Maynard Smith, J. & Szathmáry, E. *The Major Transitions in Evolution*. New York: Oxford University Press, 1997.

Nesse, R. M. "Runaway Social Selection for Displays of Partner Value and Altruism", *Biological Theory* 2/2 (2007): 143-155.

Rosas, A. "Beyond the Sociobiological Dilemma: Social Emotions and the Evolution of Morality", *Zygon* 42/3 (2007): 685-699.

Sayre-McCord, G. "Deception and Reasons to be Moral". *Contractarianism and Rational Choice*, Vallentyne, P. (ed.). New York: Cambridge University Press, 1991. 181-195.

Trivers, R. "The Evolution of Reciprocal Altruism", *Q. Revue of Biology* 46 (1971): 35-57.

Vallentyne, P. (ed.). *Contractarianism and Rational Choice: Essays on David Gauthier's Morals by Agreement*. New York: Cambridge University Press, 1991.