Boletín de la SESBE
Vol. 20(1) Marzo 2025

La moralidad como producto de la evolución

Heraclio Corrales Pavía, Universidad de Málaga, Departamento de Filosofía,
heracliocorrales@uma.es

Una faceta de la conducta normal de un ser humano es la capacidad de juzgar moralmente y de otorgar una significación ética a las acciones de otras personas. Se trata, sin duda, de un rasgo destacado del ser humano y han existido multitud de explicaciones sobre la razón de ser de esto, desde la fundamentación kantiana en la razón, hasta las ideas relacionadas con la superestructura social de la filosofía marxista. De todas ellas, hay una que ha obtenido un respaldo empírico creciente en las últimas décadas: que el origen de la moralidad está en nuestra historia evolutiva. El objetivo de este artículo es explicar qué características tiene esta teoría y cómo debe ser la mente para que sea una teoría viable. No se discutirán teorías rivales directamente, aunque esta teoría, como se comprobará, es incompatible con un origen social sobrevenido de la conducta moral.

Para este propósito, se desarrollarán dos secciones en las que se expondrán los resultados de algunas de las investigaciones con más trascendencia desde un punto de vista teórico e histórico, además de una adicional en la que se bosqueja una posible respuesta a la clásica objeción de que existen muchos códigos morales distintos, lo que parece ser un indicio de un origen aprendido de la moral. En la primera sección de este trabajo, se explicará en qué sentido puede ser útil, desde el punto de vista de la selección natural, que nuestra acción se oriente a ayudar a otros. Para ello, aludiremos a los desarrollos de teoría de juegos y a los experimentos simulados que se han desarrollado en este

ámbito, con particular énfasis en el ya clásico experimento de Axelrod (1981). En esta sección, se hablará del papel de la reputación y de la buena fama en este tipo de sistemas.

En la segunda sección, se explicará en qué sentido se puede hablar de tendencias de nuestra mente y de si acaso seguimos teniendo "instintos". En efecto, parece que somos seres racionales y que tenemos un papel consciente en nuestra toma de decisiones, pero, aún si esto no fuera falso, existe cierto margen para que nuestra mente procese la información al modo en el que lo harían generaciones anteriores a nosotros. Se desarrollará, por tanto, la teoría de la modularidad de la mente y la del *Welfare trade off ratio* (WTR) (Cosmides y Tooby, 2013), teoría novedosa y prometedora que permite entender cómo funcionan aquellas emociones que inducen o previenen conductas morales o emociones morales.

La posibilidad teórica de la evolución de la moralidad

Antes de comenzar, se deben realizar dos aclaraciones previas. En primer lugar, aunque, por mor de la simplicidad, hablemos en este trabajo solo de altruismo y de deseos de ayudar, no creo que la moralidad se agote en el altruismo y el daño. Como Tomasello ha señalado, que la moralidad surja como producto evolutivo implica que esta se relaciona con el incremento de la eficacia biológica agregada del grupo a través de una optimización de la cooperación, siendo esta última la justificación última de la existencia de la moralidad en nuestra especie y, de modo más primitivo, en otras (Tomasello, 2016). Lo que, en la práctica, se traduce en estados mentales de aprobación y desaprobación moral que tienen por objeto acciones, y en creencias sobre cómo deberíamos actuar tanto nosotros mismos como los demás. En efecto, a una gran cantidad de personas encuestadas les parece moralmente reprobable la historia de Mark y Julie, dos hermanos que, extremando las precauciones, mantienen una única relación sexual (Mark usa preservativo y Julie toma la píldora); después de la cuál nunca vuelven a tener un encuentro sexual y su relación mejora de modo significativo (Haidt, 2011). Se podría desarrollar, seguramente, una explicación en la que esta conducta se generalizara y aparecieran en el mundo una gran cantidad de niños con defectos congénitos por consanguinidad, pero no parece que esta sea la primera idea que se nos venga a la mente al juzgarlos… y, más importante, nos seguiría pareciendo mal incluso si no se entera nadie nunca.

En segundo lugar, que, como ha señalado el filósofo Philip Kitcher (2011), el término "altruismo" es ambiguo y se refiere, dependiendo del contexto, a tres aspectos distintos del mundo: i) altruismo biológico, que es aquel que implica un beneficio directo en la eficacia biológica de otro individuo, junto con un perjuicio para la propia; ii) altruismo conductual, en donde el individuo satisface los *intereses* del otro (no necesariamente aumentando la eficacia biológica o, al menos, no siendo esto central, como en el caso del consuelo en los bonobos), al margen de que la acción estuviera dirigida a beneficiar a un tercero o no, y iii) altruismo psicológico, que se refiere al deseo genuino de prestar ayuda a otro como fin en sí mismo, sin ningún beneficio propio en mente (Kitcher, 2011). Esta apreciación es fundamental, porque pareciera que si, al ser altruistas, ganamos eficacia bilógica, entonces siempre hay de fondo cierto maquiavelismo; pero no, puede ser que sintamos genuino altruismo psicológico y deseos de ayudar a los demás, aunque el origen de ese deseo sea que ese tipo de conductas han aumentado la eficacia biológica neta de nuestros ancestros.

Una razón bastante sencilla de por qué nos ayudamos los unos a los otros es que, en el fondo, los genes que perduran en el tiempo son aquellos con un fenotipo asociado que hace más probable la supervivencia de las copias del gen, incluso de aquella que no están en el propio organismo. De modo que cuidamos de esas copias que aparecen en otros organismos (de nuestros hijos, por ejemplo). Esta es una explicación de selección de parentesco explicada en término de selección de genes, tal y como lo hace Dawkins (2017), lo que es una forma bastante ilustrativa de comprender el fenómeno. Pero hemos de comprender que se trata de una teoría independiente y que la aceptación de la selección de parentesco, i. e. la idea de que la selección natural opera sobre familias que se apoyan mutuamente, no compromete con la teoría de la selección de genes (que señala que el objeto de la selección son, en última instancia, los genes) (Diéguez-Lucena, 2012). Algunos autores han apuntado que el afecto y el cuidado puede ser algo que nace en el seno de la relación maternofilial y que esa capacidad se aplica a un número progresivamente mayor de personas a medida que evoluciona la especie (Kitcher, 2011).

Ahora bien, ¿cómo es posible que una conducta que explícitamente tiene por objeto satisfacer los intereses de un tercero triunfe sobre una egoísta en la que velemos por nuestras propias necesidades? Para resolver esto, se han usado simulaciones en términos de teoría de juegos. Supongamos que se le ha detenido junto a su cómplice y se le está sometiendo a

un interrogatorio. Se le ofrece, para que coopere, un trato: si *defrauda* a su compañero y confiesa, *quedará libre de la cárcel* siempre que su compañero no confiese, yendo su amigo *3 años* a la cárcel. Si ambos confiesan, irán *2 años* cada uno. Si ninguno confesara, solo estarían en la cárcel por *1 año*. ¿Qué ha de hacer? Si ordenamos nuestras prioridades, vemos que lo óptimo sería *defraudar nosotros y que el otro coopere*, de este modo, no pisaríamos la cárcel. En segundo lugar, que *cooperemos* los dos (entre nosotros, entiéndase), por lo que iríamos 1 año. Luego, que defraudemos los dos, iríamos así dos años cada uno, y, por último, ser defraudados al tiempo que cooperamos, que es lo que se llama "hacer el primo" (Poundstone, 2018).

Si lo pensamos un momento, parece que lo *único* razonable es defraudar. Si defraudamos, siempre vamos a estar en una situación mejor que si no lo hacemos: si el otro coopera, estamos en nuestra situación preferida; pero es que, si el otro defraudara, estamos en la tercera situación en lugar de hacer el primo. Visto desde otro punto de vista, es como si nos restaran un año de prisión solo por defraudar. La situación se agrava si consideramos que el otro está pensando exactamente lo mismo que nosotros (Poundstone, 2018). ¿Cómo puede ser esto un argumento en favor del origen evolutivo de la moralidad?

Resulta que la situación cambia si jugamos varias veces. Robert Axelrod diseñó un concurso de programación en el que los aspirantes tendrían que presentar una estrategia para jugar un *número indefinido* de rondas al dilema del prisionero. Tendrían que jugar contra todos los demás aspirantes, contra una copia de ellos mismos y contra uno que coopera o defrauda aleatoriamente (llamado *Random*). La puntuación final es el resultado agregado de todas las rondas (Axelrod, 1981). Entendemos por estrategia a "la descripción completa de una forma determinada de jugar […]. Una estrategia debe prescribir las acciones a realizar tan detalladamente que nunca haga falta tomar una decisión al seguirla" (Poundstone, 2018). Nótese que es fundamental que el número de rondas sea indefinido, pues, como se ha dicho, lo más beneficioso si jugamos una sola vez es defraudar; cooperar puede servir para mandar el siguiente mensaje al otro jugador: "si nos ayudamos, podemos estar en una situación mejor que la situación 3", por lo que no tiene mucho sentido hacerlo cuando no hay más ocasión de cooperar. Pero, más aún, si sé que el otro no tiene incentivo para cooperar en la última, ¿por qué iba a cooperar en la penúltima? Podemos reproducir este razonamiento hasta la primera ronda, por lo que el juego carecería de sentido.

La estrategia que ganó fue bautizada como *Tit for tat* (Donde las dan, las toman), que consistía en i) cooperar siempre en la primera ronda, y, ii) a partir, de ahí imitar lo que hubiera hecho el otro jugador en la ronda anterior, sea cooperar o defraudar. Esta estrategia tenía la particularidad de ser amable (cooperar al principio), vengativa, conciliadora y muy sencilla, pues constaba solo de dos normas. Algo muy interesante del caso es que, cuando Axelrod repitió el experimento, aun sabiéndose quién había ganado, *Tit for* tat volvió a ganar.

Como es natural, no todas las situaciones que se dan en el reino animal siguen esta lógica, Trivers ha señalado una serie de condiciones que modulan el modo en el que se dan esta clase de interacciones: i) duración de la vida; ii) grado de dispersión; iii) nivel de dependencia mutua; iv) cuidado parental, y v) existencia o no de jerarquía de dominancia donde primen las relaciones verticales (Trivers, 1971). Esto tiene sentido: es más rentable cooperar con aquellos con los que van a tener ocasión de devolvernos el favor en el futuro, y nada de esto va a ocurrir si no necesitamos la ayuda de nuestros congéneres, como puede ser el caso, verbigracia, de una lapa. Dawkins propone un ilustrativo ejemplo de pájaros que tienen que decidir si desparasitar o no la cabeza de sus compañeros, situación cuya clasificación de los resultados potenciales se corresponde con el dilema del prisionero (Dawkins, 2017).

Lo dicho solo sirve para mostrar que, en algunas situaciones, puede ser beneficioso no actuar de un modo burdamente egoísta, lo que es poco probable que se le haya pasado por alto a la selección natural. Pero la verdad es que *Tit for tat* es una estrategia de lo más extraordinaria en el mundo animal, lo que no debe extrañarnos, porque no solemos interactuar con congéneres aleatorios y distintos cada vez. Más bien, solemos interactuar con gente que quiere interactuar con nosotros de vuelta, y las personas nos parecen adecuadas o no para nuestro proyecto vital en función de las cualidades que estas hayan exhibido (Krebs, 2011).

Lo dicho solo sirve para mostrar que, en algunas situaciones, puede ser beneficioso no actuar de un modo burdamente egoísta, lo que es poco probable que se le haya pasado por alto a la selección natural

Boletín de la SESBE
Vol. 20(1) Marzo 2025

Se ha comprobado que, cuando el emparejamiento puede basarse en la reputación previa de partidas anteriores, aquellos con más prestigio sacaban más beneficio

Los sistemas como los descritos por Axelrod se llaman de "reciprocidad directa" y aquellos en los que los jugadores no se emparejan aleatoriamente, sino en función de sus puntuaciones, son de "reciprocidad indirecta". Se ha comprobado, en el juego del bien común, que las personas con mejor nombre tienden a obtener mejores beneficios. En este juego, se da una suma inicial a cada persona y se ponen a jugar en grupos. Cada miembro del grupo tenía dos opciones: poner sus ingresos en un fondo común, en el que se multiplica y se divide en partes iguales entre los jugadores o quedárselo. El bote se reparte igualmente entre todos los jugadores, al margen de que hayan cooperado o no. Se ha comprobado que, cuando el emparejamiento puede basarse en la reputación previa de partidas anteriores, aquellos con más prestigio sacaban más beneficio. Existe una correlación positiva también con la posibilidad de castigar, lo que es un indicio sobre los beneficios de que se desarrolle la conducta punitiva (Hauser, 2007).

La modularidad de la mente y la psicología evolucionista

Para que la evolución haya podido dejar su huella en nuestra conducta, es necesario que se den algunas características en la mente humana. En efecto, si tuviéramos un control libre y directo sobre cada uno de nuestros procesos cerebrales, si todo fuera aprendido, no podríamos hablar de que la moralidad tiene un origen evolutivo. Como es evidente, podemos aprender y este aprendizaje puede afectarnos a la hora de elegir un curso de acción u otro. Pero esto no significa que el cerebro sea un folio en blanco en el que se puede escribir cualquier cosa. Más bien, lo que se defiende es que a través del aprendizaje se modula el funcionamiento de estructuras preexistentes a través de medios que son también fruto de la selección natural.

La teoría sobre la naturaleza de la mente preferida por los psicólogos evolucionistas es la de la modularidad de la mente. Según este modelo, "la mente es un sistema complejo de muchas partes que interactúan" (Pinker, 2018), es decir,

Boletín de la SESBE
Vol. 20(1) Marzo 2025

Nuestro cerebro, al igual que el resto de los órganos de nuestro organismo, tiene una función, que es responder al entorno del modo que este requiere: la respuesta que tenemos ante una apetitosa hamburguesa es distinta a la que tenemos cuando vemos a alguien que nos parece atractivo; y esto no es en algo que hagamos a nivel consciente

existen diferentes programas que generan diferentes respuestas ante distintos estímulos. Cierto es que podemos aprender, pero podemos aprender precisamente porque estamos "programados" para que podamos hacerlo (Pinker, 2018). Es decir, nuestro cerebro, al igual que el resto de los órganos de nuestro organismo, tiene una función, que es responder al entorno del modo que este requiere: la respuesta que tenemos ante una apetitosa hamburguesa es distinta a la que tenemos cuando vemos a alguien que nos parece atractivo; y esto no es algo que hagamos de modo consciente (Cosmides y Tooby, 2013). Nuestro cerebro procesa la información, la categoriza, la asocia a un programa concreto, genera una respuesta fisiológica y, en algunos casos, hace que aparezca en la consciencia como un deseo, aversión, asociación, etcétera.

De este modo, la respuesta ante un estímulo dado no es algo que decidiríamos de modo consciente, solo vemos el producto del módulo. Pero ¿es aprendido? De nuevo, si estos módulos fueran muy maleables o su funcionamiento se basará en un aprendizaje social en su práctica totalidad, tampoco podríamos decir nada muy interesante de nuestra vida moral desde la psicología evolucionista. Esto es una cuestión, como es natural, de grados. Los estudios de heredabilidad con gemelos, que tienen por objeto explicar qué porcentaje de las diferencias entre la personalidad de los individuos se explica por la genética concluyen que los *Big five* (los rasgos de la personalidad principales cuya combinación nos sirve para describir a cualquier persona de un modo bastante adecuado) son heredables en torno al 50%. Estos rasgos, que en inglés forman el acrónimo OCEAN son apertura a nuevas experiencias, responsabilidad, extroversión, afabilidad y neuroticismo (Bueno, 2020; Pinker, 2018).

Boletín de la SESBE
Vol. 20(1) Marzo 2025

Este es el marco teórico en el que se desarrolla el estudio de la huella evolutiva en nuestra moralidad, pasemos a ver cómo se ha defendido el funcionamiento de los módulos implicados en el desarrollo de la conducta social. El inconsciente funciona usando unos parámetros llamados *"variables regulatorias internas"*, que "no son exactamente creencias ni conceptos", sino valoraciones que hace nuestra mente de las aferencias que recibe de los sentidos y asociaciones, y que utiliza para modular la respuesta al entorno que ha causado el estímulo (Cosmides y Tooby, 2013). Estas variables pueden afectar a diferentes módulos de modo simultáneo, v.g. si nuestra mente interpreta que una persona tiene un alto grado de parecido genético con nosotros por haber visto a nuestra madre cuidarlo, experimentaremos más simpatía hacia esta persona al tiempo que nos sentiremos menos atraídos sexualmente hacia ella. Todo esto sin perjuicio de que los módulos estén "encapsulados" (i. e. no se influyen el uno al otro). Diferentes módulos usan la misma información para distintas cosas (Cosmides y Tooby, 2013).

Entre todos los módulos, para lo que aquí se trata, hay uno que tiene un interés especial para el estudio de la conducta moral: la ratio de compensación de bienestar (WTR en adelante, por sus siglas en inglés) que es, intuitivamente, la variable regulatoria interna que modula lo mucho que nos preocupamos por el bienestar y los intereses de otra persona en concreto. Así, cuando el WTR que experimentamos hacia una persona es más alto, tenderemos a favorecerla en mayor medida, incluso en casos donde haga falta un gran empeño por nuestra parte para satisfacer un poco de sus intereses (Cosmides y Tooby, 2013; Sell, 2017). El WTR, como se mostrará inmediatamente, es fundamental para decidir si ayudar o no a otra persona a satisfacer sus intereses. Asignamos un WTR a cada persona en base a una serie más compleja de parámetros, y este es bastan-

Entre todos los módulos hay uno que tiene un interés especial para el estudio de la conducta moral: la ratio de compensación de bienestar que es, intuitivamente, la variable regulatoria interna que modula lo mucho que nos preocupamos por el bienestar y los intereses de otra persona en concreto

te más consistente, coherente y transitivo con respecto el asignado a otros de lo que esperaríamos por mero azar, siendo esto así en todas las culturas estudiadas (Delton *et al.*, 2023).

¿Cómo decide nuestra mente si priorizar los deseos de otra persona sobre los nuestros? Comparando dos cursos posibles de acción, uno previsiblemente altruista y otro previsiblemente egoísta. Se actuará de modo altruista (nos referimos, naturalmente, al altruismo psicológico) exclusivamente si:

$$WTR \times B_{\text{de la otra parte}} > B_{\text{propio}}$$

Donde B representa el beneficio previsible del curso de acción altruista. Así, esta fórmula compara la prioridad que le asignamos a los intereses de otras personas en nuestro cálculo mental inconsciente con la propia. Naturalmente, lo habitual es que sea un valor entre 0 y 1, pero podemos concebir casos en los que sea mayor que 1 o menor que 0. Nótese que el hecho de que una persona sea un desconocido para nosotros no significa que nuestro WTR hacia ella sea 0 (i. e. que no nos importe en absoluto).

Esto no solo es relevante porque muestra un mecanismo por el que nos ayudamos por el mero deseo de colaborar: sirve para explicar la existencia de las emociones morales, esto es, aquellas emociones que nos conducen a acciones que pueden ser interpretadas desde un punto de vista moral. Para esto es necesario que podamos figurarnos cuál es el WTR que otras personas nos asignan a nosotros. De acuerdo con este modelo, nos sentimos como nos sentimos por la discordancia entre el WTR real y el estimado. Si una persona nos trata mejor de lo que esperábamos, nos sentimos agradecidos y estamos más dispuestos a actuar en su beneficio; si, por el contrario, vemos que no hemos tenido a alguien en la consideración debida, nos sentimos culpables como mecanismo para tratarla mejor en el futuro, etcétera (Cosmides y Tooby, 2013).

Una emoción que ha recibido particular atención es la ira, que se produce cuando percibimos que no nos tratan con la consideración que creemos que nos merecemos (de ahí que los narcisistas sean bastante irritables). Frente a la teoría rival de que nos enfadamos porque se nos impone una desigualdad, la teoría del WTR nos permite entender por qué nos enfadamos más cuando alguien nos deja plantado por un motivo trivial que por uno trascendente (Sell, 2017).

Como es natural, esto solo es la génesis del impulso moral, constamos de módulos de carácter general que permiten razonar y gestionar las acciones con implicaciones

morales. Sigo pudiendo contar hasta diez antes de gritar, pero ahora comprendemos que la indignación y la ira viene de un sistema profundamente arraigado en nuestra mente que tiene que ver con el altruismo, la reciprocidad y, en último término, con el tipo de conductas que nos suscitan emociones por una u otra razón. Naturalmente, podemos discutir qué es bueno y qué debemos perseguir, pero el acceso a la comprensión del bien y el mal viene, en primera instancia, de lo aquí expuesto.

Es importante notar que nada de lo aquí expuesto pretende sugerir que la moralidad sea un producto rígido de la evolución, por el contrario, hay razones para pensar que se trata de una adaptación que exhibe plasticidad fenotípica i. e. cuya expresión varía en entornos distintos. Esto es perfectamente coherente con la observación antes referida de que la razón última es la cooperación: diferentes grupos en diferentes circunstancias necesitan conductas diferentes, parece difícil que la moralidad sea todoterreno e independiente del aprendizaje social.

Conclusión

En este artículo se ha bosquejado una defensa sobre el origen evolutivo de la moralidad, aludiendo para ello a las tesis de más relevancia sobre teoría de juegos y psicología evolucionista. A través de la exposición de los modelos simulados de cooperación representados por el dilema del prisionero, se ha justificado en qué sentido es posible que sea beneficioso para un individuo experimentar altruismo psicológico. Naturalmente, esta justificación se ha usado para explicar el porqué del altruismo biológico, pero parece claro, que al menos en nuestra especie, el altruismo espontáneo está mediado por la experiencia del altruismo psicológico en este sentido, por lo que es un indicio también de por qué sentimos este tipo de estados mentales.

En segundo lugar, se ha defendido la naturaleza modular de la mente y los mecanismos más relevantes de producción de actos morales y de emociones morales. Si bien se ha dicho que esto no agota ni la reflexión ética ni el razonamiento moral, sí que podemos entender que es el origen de lo que nos hace capaces de captar la bondad o maldad de las acciones que percibimos. A través del estudio del WTR, podemos comprender qué es lo que subyace a nuestras ideas intuitivas y espontáneas sobre la cooperación, el altruismo, la moralidad y las expectativas que tenemos sobre las acciones de los demás. Creo que un estudio de la moralidad que pasara por alto el origen de estas intuiciones sobre el bien y el mal estaría irremediablemente incompleto.

Referencias

Axelrod, R. y Hamilton, W.D. 1981. The evolution of cooperation. Basic Books: New York.

Bueno, D. 2020. Genética y aprendizaje: Cómo influyen los genes en el logro educativo. *Journal of Neuroeducation* 1(1), 38–51.

Cosmides, L. y Tooby, J. 2013. Evolutionary psychology: New perspectives on cognition and motivation. *Annual Review of Psychology* 64(1), 201-229.

Dawkins, R. 2017. El gen egoísta extendido. Madrid: Bruño.

Delton, A. W. *et al.* 2023. Cognitive foundations for helping and harming others: Making welfare tradeoffs in industrialized and small-scale societies. *Evolution and Human Behavior* 44(5), 1–17.

Diéguez Lucena, A. J. 2012. La vida bajo escrutinio. Una introducción a la filosofía de la biología. Barcelona: Biblioteca Buridán.

Haidt, J. 2001. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review* 108(4), 814–834.

Hauser, M. D. 2007. Moral Minds. HarperCollins e-books.

Kitcher, P. 2011. The ethical project. Massachusetts: Harvard University Press.

Krebs, D. L. 2011. The Origins of Morality. Oxford: Oxford University Press.

Pinker, S. 2018. La tabla rasa. La negación moderna de la naturaleza humana. Barcelona: Paidós.

Poundstone, W. 2018. El dilema del prisionero. Madrid: Alianza.

Sell, A. N. 2017. The grammar of anger: Mapping the computational architecture of a recalibrational emotion. *Cognition* 168, 110–128.

Tomasello, M. 2016. A Natural History of Human Morality. Cambridge: Harvard University Press.

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