Lenguaje y Resguardo Social: la evolución cultural de la teoría de la mente

Language and Social Cover: the Cultural Evolution of the Theory of Mind

Víctor Fernández Castro
DOI: 10.26754/ojs_arif/arif.202024946

Resumen

De acuerdo con algunas posiciones en ciencia cognitiva y filosofía, la capacidad humana de invocar estados mentales para predecir y explicar la conducta humana es un producto de la evolución cultural; es decir, su aparición no se debe a la selección de estructuras heredadas genéticamente sino que son el fruto de la variación hereditaria mediante mecanismos de aprendizaje social. El objetivo de este artículo es proponer un modelo de la teoría de la mente como mecanismo lingüístico de resguardo social y promoción de la reputación dentro del marco de la evolución cultural. Esta posición, que denominaré *El Modelo de Resguardo Social*, se presenta y defiende en contraposición a otras tres posiciones que se evaluarán críticamente. Finalmente, se defenderá el modelo de dos posibles objeciones.

Palabras clave: Teoría de la mente, evolución cultural, lenguaje, reputación

Abstract

Recently several views in cognitive sciences and philosophy have argued that human capacity to ascribe mental states to predict and explain human behavior is a product of cultural evolution; that is, its emergence is not the product of the selection of genetically inherited structures but the product of hereditary variation through mechanisms of social learning. The aim of this article is to propose a model of the theory of mind as a linguistic mechanism of social cover and reputation management within the framework of cultural evolution. This position, called The Social Cover Model, is presented and defended in contrast to three other positions that will be critically evaluated. Finally, the model of two possible objections will be defended.

Keywords: Theory of Mind, Cultural Evolution, Language, Social Cover, Reputation

Recibido: 13/12/2020. Aceptado: 16/12/2020
Análisis. Revista de investigación filosófica, vol. 7, n.º 2 (2020): 191-215 ISSNe: 2386-8066

1. Introducción

En los últimos años una cantidad importante de investigación en filosofía, psicología y ciencia cognitiva se ha encargado de la capacidad conocida como *teoría de la mente*[1], la capacidad para para invocar actitudes proposicionales para explicar, predecir o moldear el comportamiento de otros agentes. Recientemente, varios autores han defendido que la teoría de la mente es producto de la evolución cultural (Fenici y Zawidzki 2020, Heyes 2019, Heyes y Frith 2014, Moore 2020). Al igual que el constructivismo clásico, y en contra del nativismo, estas posiciones consideran que la facultad para atribuir estados mentales es aprendida y no un producto de mecanismos seleccionados genéticamente. Sin embargo, a diferencia de las posiciones clásicas, estas autoras han enfatizado la idea de que la capacidad de teoría de la mente es adaptativa y fruto de la evolución mediante variación hereditaria, aunque la transmisión de generación en generación sea a través del aprendizaje social y no de los genes. Al mismo tiempo, esta posición, que denominaré *constructivismo evolutivo*, ha abrazado algunos de los estudios empíricos a favor del lingüalismo, la tesis de que atribuir estados mentales depende causal o constitutivamente del lenguaje. El constructivismo evolutivo, además, no sólo intenta hacerse cargo de la evidencia empírica relevante sino apuntar a *estrategias evolutivamente estables* (Maynard-Smith 1982), es decir, plantear hipótesis sobre cómo las atribuciones de estados mentales llegaron a invadir una población de organismos lingüísticos y como se ha mantenido estable frente a la invasión de mutantes que carecen de ese rasgo.

El principal objetivo de este ensayo es discutir críticamente tres modelos constructivistas evolutivos sobre cómo la psicología popular evolucionó como una herramienta lingüística. Al mismo tiempo, defenderé un modelo que denominaré el "Modelo de Resguardo social", de acuerdo con el cual, la atribución de estados mentales es una capacidad lingüística orientada a promover nuestra reputación a través de dar razones que justifiquen o apoyen nuestros patrones de conducta. El artículo se estructura de la siguiente manera: en la sección 2 caracterizaré el constructivismo de manera general junto con la evidencia a favor de la idea de que la teoría de la mente es una capacidad que depende del lenguaje. En la sección 3, 4 y 5 discutiré tres modelos constructivistas sobre por qué evolucionó a través

[1] En este artículo, usaré los términos, 'teoría de la mente', 'lectura de mente' y 'psicología popular' (folk psychology) de manera intercambiable para referirme a la atribución de estados mentales.

de ciertas prácticas lingüísticas y qué función adaptativa podría haber supuesto. Argumentaré que esos modelos presentan varios problemas para dar una visión plausible sobre la evolución de la psicología popular dentro del marco de la evolución cultural. En la sección 6 presentaré el modelo de Resguardo Social y explicaré cómo esta posición evita los problemas de las otras posiciones y refuerza alguno de sus puntos fuertes. Finalmente, en la sección 7 defenderé el modelo de dos posibles objeciones relacionadas con su plausibilidad evolutiva y sus consecuencias metafísicas.

2. La Evolución Cultural de la Teoría de la Mente

En 2014, Heyes y Frith publicaron un artículo titulado "The Cultural Evolution of Mind Reading" en el que proponían que, al igual que la capacidad de leer letra escrita, la teoría de la mente es una capacidad que no depende de mecanismos evolucionados genéticamente. Lo interesante de la propuesta de Heyes y Frith se encuentra en que, trazar esta analogía con la lectura, nos permite ver cómo algunas características de ambas facultades que parecen apuntar al innatismo no son necesariamente prueba ello. Por ejemplo, la lectura, al igual que la teoría de la mente, depende de mecanismos corticales específicos, está sujeta a desordenes genéticamente heredados y muestra cierta variación cultural dentro de unos márgenes universales. Sin embargo, la lectura no es una facultad innata, ya que sabemos que su existencia data de la historia reciente del *Homo Sapiens*. De manera similar, podemos concluir, nos dicen Heyes y Frith, que dichas características aplicadas a la teoría de la mente deben ponernos sobre la pista de que es producto de la evolución cultural; es decir, aunque son fruto de la evolución mediante variación hereditaria, dicha evolución no se transmite de generación en generación a través de los genes sino del aprendizaje social.

El constructivismo, entendido como la tesis de que la teoría de la mente es una capacidad adquirida mediante aprendizaje, no es nueva. Una de las teorías más establecidas sobre la teoría de la mente, la denominada *Teoría-Teoría* (Gopnik y Meltzoff 1997, Wellman 2014) proponen, en una de sus versiones, que los niños actúan como científicos que adquieren los conceptos mentales mediante el desarrollo de teorías explícitas de cómo los estados mentales causan el comportamiento a través de la observación y la prueba de hipótesis. Sin embargo, la versión evolutiva de Frith y Heyes sugiere que el conocimiento del niño sobre la mente se deriva principalmente, no de la observación, sino de la instrucción que el niño recibe de los expertos en psicología popular en su mundo social (Frith y

Heyes 2014, p. 4). La adquisición del conocimiento y competencia que se necesita para atribuir estados mentales de manera adecuada a otros agentes no deriva de la formulación y evaluación de hipótesis sino del conocimiento compartido por otros agentes.

Existen, además, dos fuentes importantes de evidencia empírica a favor del constructivismo evolutivo. Por un lado, estudios recientes apuntan a una gran variedad cultural en el uso de la psicología popular (Lavelle 2019). Por ejemplo, mientras los niños australianos y estadounidenses entienden que dos personas pueden tener creencias diferentes antes de aprender que las personas pueden tener acceso diferente al conocimiento de un hecho, los niños chinos e iraníes exhiben el patrón contrario (Shahaeian y otros 2011). Otro ejemplo es el caso de las diferencias en los mismos conceptos mentales en diferentes culturas. El concepto 'Kokoro' en japonés, por ejemplo, no parece tener correspondencia en ningún otro idioma y recoge aspectos tanto emocionales y corpóreos como racionales (Lebra 1993); o la distinción entre 'believing' y 'thinking' en inglés que refleja diferentes grados de creencia y parece ser idiosincrática de la cultura anglosajona (Wierzbicka 2006).

Por otro lado, evidencia empírica importante muestra una relación estrecha entre lenguaje y la teoría de la mente. Una de las principales fuentes de esta evidencia son los experimentos que intentan elucidar la conexión entre la exposición a cierto tipo de vocabulario o contextos lingüísticos y la capacidad de pasar el test de creencia falsa[2] (Astington y Jenkins 1999, Happé 1995, Taumoepeau y otros 2008). Estos estudios muestran, por ejemplo, una correlación fuerte entre la exposición al vocabulario mental por parte los progenitores y la precocidad de los niños para pasar el test de creencia falsa. Además, Astington y Jenkins (1999) encontraron que el éxito en la teoría de la mente no era un predictor de las habilidades lingüísticas, pero sí al revés, el dominio del lenguaje era un buen indicador de la competencia de la teoría de la mente. Otro grupo de experimentos que respaldan la influencia

[2] Wimmer y Perner (1983) diseñaron lo que ahora se conoce como la tarea de falsa creencia explícita. En esta tarea, un niño es expuesto a un escenario donde una marioneta, Maxi, pone chocolate en un armario x. Cuando Maxi no está presente, su madre cambia el chocolate de x en un armario y. Los niños tienen que indicar la caja donde Maxi buscará el chocolate cuando regrese. Sólo cuando el niño es capaz de representar la creencia errónea de Maxi, es capaz de señalar correctamente la caja x. La tarea prueba si los niños tienen una "representación explícita de lo erróneo de la creencia de esta persona en relación con su propio conocimiento" (p. 103). Wimmer y Perner (1983) descubrieron que los niños menores de 3 años a menudo fallan en esta tarea.

del lenguaje en la adquisición de la lectura de mentes son los de De Villiers y otros (De Villiers y De Villiers 2000, De Villiers y Pyers 2002). En esto experimentos se muestra una fuerte correlación entre el dominio de la competencia lingüística para usar complementos oracionales y la teoría de la mente. En uno de esos experimentos, por ejemplo, las autoras presentaban a los sujetos preguntas del tipo "La mamá dijo que compró manzanas, pero mira, realmente compró naranjas. ¿Qué dijo la mamá que compró?", cuya respuesta requiere comprender el uso de la clausula-que. De Villiers y sus colegas descubrieron que los niños que responden correctamente a este tipo de preguntas tienen más éxito a la hora de enfrentarse con la tarea de la falsa creencia frente a los niños que no lo hacen. Además, otro estudio mostró que entrenar a niños en ejercicios de aprendizaje con complemento oracionales mejora sus puntuaciones en las tareas de falsa creencia (Hale y Tager-Flusberg 2003).

En resumen, tanto la variabilidad cultural asociada a la teoría de la mente como los experimentos sobre la conexión entre lenguaje y el test de creencia falsa parecen apuntar a que la esta es aprendida y, presumiblemente, producto de la evolución cultural. Sin embargo, esto abre un interrogante fundamental ¿Cómo y por qué produjo la evolución cultural la teoría de la mente?

3. Atribuciones de Estados Mentales y Enseñanza

De acuerdo con Heyes (2019), la respuesta se encuentra en cómo la lectura de mentes refuerza los mecanismos de aprendizaje dentro de una población, los cuales, a su vez, retroalimentan la evolución cultural. Heyes (2019, pp. 198-203) argumenta que algunos de nuestros mecanismos cognitivos más importantes son producto de la evolución cultural en tanto que están sujetos a la selección cultural de grupo. Es decir, estos mecanismos emergieron en tanto que facilitaban el aprendizaje de estrategias y tecnologías que suponían una ventaja adaptativa para el grupo o mejoraban la transferencia de esas estrategias. En este sentido, la teoría de la mente emergió como un mecanismo que mejoraba el aprendizaje de los individuos de un grupo de modo que tanto las estrategias adaptativas como el conocimiento que suponía una mejora de la vida se transmitía mejor y de manera más fiable entre los miembros, mejorando su aptitud (*fitness*) frente a otros grupos que no tuvieran dicho mecanismo. Concretamente, la teoría de la mente facilita la dinámica de enseñanza en tanto que sirve para desentrañar lo que otros individuos piensan en el mismo contexto de interacción, lo que ayuda al maestro a detectar errores o aciertos en el aprendizaje del pupilo.

La enseñanza puede entenderse como un tipo de aprendizaje social donde el agente modelo o maestro no sólo permite al agente receptor o pupilo observar cómo despliega una cierta estrategia, sino que actúa con la intención de producir un cambio perdurable en la mente receptor para que sea capaz de desplegar la estrategia (Byrne y Rapaport 2011). Aunque algunos investigadores han puesto en duda que la noción de enseñanza deba definirse en términos de psicología popular, parece inevitable pensar que esta y el aprendizaje están estrechamente relacionados. La atribución de estados mentales, nos dice Heyes, permite a los maestros representarse la extensión o límite del conocimiento de sus pupilos y, por tanto, inferir a cada estadio del proceso de aprendizaje lo que se necesita mostrar o decir a ese pupilo en particular para superar su ignorancia, corregir sus creencias falsas o construir su corpus de conocimiento (Heyes 2019, p. 145). En este sentido, parece razonable pensar que, aunque la psicología popular pudiera hacer mejorar diversas capacidades sociales y no sociales, su evolución cultural pudiera estar ligada a el aprendizaje y, por tanto, que fuera seleccionada por su capacidad pare mejorar la transmisión fidedigna de estrategias y conocimientos[3].

Existen varios problemas a la hora de entender la evolución cultural de la atribución de estados mentales en los términos en los que lo hace Heyes (Morin 2019). Para nuestro objetivo, el primero de estos problemas está estrechamente relacionado con la visión seleccionista de la evolución cultural de Heyes, la cual no hace justicia a algunas dinámicas evolutivas culturales. Muchas tradiciones, estrategias culturales o normas sociales sencillamente no son seleccionadas por dinámicas ciegas que atienden a mejoras en la calidad de vida de los individuos o ayudan a mejorar la transmisión sino simplemente por cambios guiados que atienden a razones individuales como, por ejemplo, que son más atractivas desde un punto de vista psicológico (Claidière y otros 2014, Sperber 1996). Esto mismo podría aplicarse al caso de los mecanismos cognitivos en general y teoría de la mente en particular. Uno podría imaginar que las atribuciones de estados mentales aparecieron porque, por ejemplo, nos permitieron llevar a cabo actos de habla vicarios o citar a otras personas (Van Cleave y Gauker 2010, Geurts en prensa).

[3] Heyes argumenta que los aparatos cognitivos específicamente humanos como el aprendizaje selectivo, la imitación o la misma teoría de la mente no sólo fueron heredados mediante mecanismos sociales, sino que además fueron seleccionados precisamente porque mejoraban esa herencia. Es decir, la lectura de mentes fue seleccionada por su capacidad para mejorar la fidelidad de la información transmitida. Esta fidelidad provocó que la evolución cultural se *darwinizara* siendo más seleccionista.

Esta función podría ser atractiva desde el punto de vista comunicativo o incluso psicológico ya que nos permitiría lleva a cabo actos de habla por boca de otros y por tanto comunicar contenidos de los que no tenemos que hacernos cargo o con los que no tenemos que comprometernos. En este escenario, uno podría imaginar que la lectura de mentes podría dispersarse sin que fuera seleccionada de manera ciega, sino simplemente porque permite a los agentes ciertos usos comunicativos que les parecen útiles (ver sección 6).

Por otro lado, Heyes asume que la evolución cultural de la teoría de la mente se produce porque ayuda a facilitar la transmisión de información relevante de unos individuos a otros de manera fidedigna a través de la enseñanza. Sin embargo, no queda claro que la teoría de la mente necesariamente tenga un impacto positivo en la monitorización o comprensión de los otros agentes que se traduzca en una mejora del aprendizaje. Parte de la razón es que existe evidencia empírica importante que muestra que nuestras atribuciones de estados mentales están sesgadas y reflejan varios prejuicios y limitaciones (Spaulding 2018, Westra 2017). Por poner algunos ejemplos, tendemos clasificar a las personas según sus categorías sociales, como la edad, la raza y el género que facilitan la atribución de rasgos de la personalidad, como la la agresividad y la confianza (Olivola y Todorov 2010, Rule y otros 2009). Además, el interés propio (por ejemplo, la reducción de la ansiedad, la confirmación de nuestra visión del mundo) también afecta a la atribución (Dunning 1999). Este tipo de sesgos muestran que nuestra imagen de la mente de los demás no es necesariamente fidedigna, lo que pone en duda la idea de que esta mejoraría la capacidad de los maestros para detectar los problemas de aprendizaje de los pupilos.

4. Comparación de estados mentales, toma de perspectiva e iteración

La idea de que la evolución de las adscripciones de estados mentales está ligada a la función de generar una imagen más adecuada de otras mentes no sólo aparece asociada a un mejor aprendizaje si no también a la emergencia de nuevas capacidades. Moore (2020) ha argumentado que el lenguaje nos permitió crear modelos más fidedignos de la mente de otros lo que hizo aparecer nuevas capacidades como como la de comparar diferentes estados mentales o poder tomar la perspectiva de otros agentes. La gramaticalización del lenguaje permitió que surgieran diferentes estructuras gramaticales como marcadores temporales o modales (Pogovac 2015), pero también verbos perceptivos, los cuales, ya proporcionaban

los ingredientes necesarios para empezar a representar estados epistémicos más abstractos como conocimiento o creencia.

Una vez que aparecieron estos ingredientes, el uso de las atribuciones de estados mentales se dispersó a través de la población porque nos permitía tener modelos más fidedignos de la mente de los demás, lo que dotó a los agentes de tres habilidades novedosas: (1) contrastar estados mentales, (2) iterar estados mentales y (3) toma de perspectiva de segundo nivel 2 (*'Level-2' Perspective taking*). En primer lugar, la teoría de la mente nos permitía contrastar estados mentales, es decir, nos permitiría evaluar las consecuencias conductuales de actitudes diferentes hacia las mismas proposiciones. En este sentido se nos permite clarificar la diferencia entre que Ricardo *crea que* el tren sale a tiempo y que María lo *dude*. Así, podríamos explicar porqué Ricardo podría ir deprisa a la zona de vías, mientras que María sigue comiendo tranquilamente a pesar de que ambos cogen el mismo tren (ejemplo del propio Moore). En segundo lugar, los verbos de actitudes proposicional nos permitirían crear representaciones no sólo de segundo si no de tercer y cuarto orden (María pretende que Antonio crea que se ha acabado la cerveza). En tercer lugar, las atribuciones mediadas lingüísticamente nos permitirían "comprender cómo se presentan las cosas a los demás cuando se perciben desde diferentes perspectivas" (Moore 2020, p. 13). Mientras que los chimpancés y los niños pre-lingüísticos son capaces de representar lo que otros agentes ven o no, sólo los humanos pueden representarse lo que ven desde la perspectiva del agente mismo, es decir, son capaces de percibir que un objeto puede percibirse por uno mismo de manera diferente a como lo percibe otro agente[4].

Moore considera que la conexión entre estas capacidades y las estructuras gramaticales involucradas en las atribuciones, junto con la evidencia de la variación cultural (ver sección 2), son razones suficientes para sospechar que la teoría de la mente es fruto de la evolución cultural. Sin embargo, el autor no explica cómo estas tres habilidades que permite un mayor refinamiento de los modelos de la mente de los demás suponen una ventaja evolutiva que permitiría la dispersión, y es aquí precisamente donde la posición de Moore plantea algunos problemas.

[4] Moll y Metlzoff (2011) colocaron a varios individuos (niños de 3 años) en frente del experimentador y dos objetos azules idénticos entre los dos. Después, se colocaba un cristal amarillo entre el experimentador y uno de los objetos de manera que se veía verde para el experimentador y azul para el sujeto. Cuando se les pedía a los sujetos el objeto azul o verde, el sujeto escogía el objeto correcto de acuerdo a la perspectiva del experimentador.

En principio, existen dos hipótesis plausibles de cómo tal refinamiento podría suponer una ventaja adaptativa. En primer lugar, como parece sugerir Moore, se podría pensar que tal refinamiento mejora la coordinación entre individuos lo que supondrían una mejora para el grupo. Desde este punto de vista, aquellos grupos de agentes que fueran capaces de coordinarse mejor tendrían una ventaja adaptativa con respecto a otros grupos. En segundo lugar, como defienden algunos de los autores que vincula la teoría de la mente con la iteración de estados mentales (p.ej. Sperber 2000), se podría pensar que la ventaja consiste en ser capaz de manipular los estados mentales y el comportamiento de los demás a través del engaño. Es decir, la psicología popular permitiría un mayor éxito individual para sobrevivir a la competencia social con otros individuos humanos.

El principal problema de la primera hipótesis es que no queda claro cómo tener un conocimiento más precioso de las otras mentes se traduce necesariamente en una mejor coordinación entre individuos. En principio, los defensores del constructivismo evolutivo aceptan que existen capacidades para la cognición social más básicos que la lectura de mentes, como, por ejemplo, la sub-mentalización, que nos permiten rastrear estados mentales con mecanismos generales de aprendizaje. (Heyes 2014)[5]. Estos mecanismos nos permiten coordinarnos de manera eficiente y automática, lo que haría de la atribución de estados mentales una facultad redundante o superflua en lo que respecta a esta función. Para que esto no fuera así, la facultad debería suponer una mejora más que notable de la coordinación. Sin embargo, parece dudoso que esto sea el caso, especialmente teniendo en cuenta las habilidades a las que apunta Moore. Retomemos el ejemplo de Ricardo y María. Aunque pudiéramos inferir con cierta certeza sus diferentes actitudes hacia la proposición de que el tren va a salir a tiempo, esa exactitud no se traduce necesariamente en una mejor coordinación debido al poco poder de predicción de la atribución de estados mentales. Como Morton (1996) y Zawidzki (2008, 2013) han argumentado, la capacidad de predicción de la atribución de estados mentales es reducida debido a que las actitudes proposicionales tienen una

[5] La aceptación de este punto se debe en parte a la aparición de los test de creencia falsa implícitos. En el experimento clásico, Onishi y Baillargeon (2005) midieron el tiempo de observación de bebés de 15 meses para probar sus reacciones en contextos similares a los de creencia falsa. los bebés miran más tiempo cuando una persona, que no estaba presente cuando un objeto era reubicado de un contenedor a otro, seleccionaba el objeto del correcto. Esto implicaba que la persona viola las expectativas del niño y que, por tanto, son sensibles a la creencia falsa.

conexión muy tenue con la acción. El holismo de lo mental hace que no haya un solo comportamiento asociado a un estado mental en un contexto específico por lo que la predicción a través de las atribuciones de actitud proposicional es altamente compleja. Por ejemplo, María podría no quedarse a comer a pesar de dudar de que le tren salga a tiempo, sencillamente porque es una persona precavida o porque sabe que tiende a perder la noción del tiempo y podría perder el tren a pesar del retraso[6].

El argumento de la mejora de la coordinación parece que se aplica mejor a la habilidad de toma de perspectiva. Uno podría imaginar casos en los que tener en cuenta la perspectiva de otros agentes podría suponer una ventaja para predecir su comportamiento y por tanto modificar la conducta propia para adaptarse al objetivo compartido en una acción conjunta (Pacherie 2011). Sin embargo, los casos donde tener una representación fidedigna de la perspectiva del otro suponga una mejora fundamental para la coordinación parecen ser limitados. En la mayoría de las acciones conjuntas no parece que nos enfrentemos a objetos o aspectos del entorno que cambien radicalmente de aspecto dependiendo de la perspectiva. En este sentido, no hay demasiados casos donde necesitemos tener en cuenta la perspectiva del otro, sobre todo porque ya poseemos otros mecanismos cognitivos que nos permiten coordinarnos y orientarnos teniendo en cuenta los objetos involucrados en la acción como, por ejemplo, mecanismos de atención o percepción conjunta (Knoblich *et al.* 2011, Vesper *et al.* 2017). Además, no parece que, desde una perspectiva evolutiva, la mejora que implica el poder representarnos la perspectiva del otro en comparación con poder compartir la atención o la percepción suponga una diferencia tan grande que guiara la evolución cultural en esa dirección. Es decir, la teoría de la mente sería un mecanismo redundante cuando atendemos a su capacidad de mejora de la coordinación.

De este modo, tanto la toma de perspectiva como la iteración de estados mentales parecen apuntar a la segunda hipótesis, estas nos dotaron de habilidades para el engaño. El problema principal de esta hipótesis reside en que parece ser incompatible con varios aspectos fundamentales del marco evolutivo que plantean las posiciones constructivistas. Por un lado, entender que la psicología popular es producto de la evolución cultural implica que los individuos sujetos a ella deben

[6] Uno podría pensar que la atribución más sofisticada de estados mentales no supone una mejor predicción en el momento de la atribución, pero si una mejor explicación a posteriori, lo que podría ayudar al atribuidor ayudar a generar mejores predicciones futuras. Sin embargo, una mejor explicación no siempre se traduce en mejores predicciones (Andrews 2012).

heredar la capacidad a través de mecanismos de aprendizaje social. Sin embargo, parece poco probable que el funcionamiento de estos mecanismos de aprendizaje pueda darse en contextos donde los individuos no sean ya altamente cooperativos, lo que supone un desafío para cualquier teoría que intenta dar cuenta de la aparición de teoría de la mente en términos de engaño (ver Sterelny 2012: 6-10, Zawidzki 2013: 104-111). Al fin y al cabo, los individuos que se arriesgaran a engañar a otros agentes corren el riesgo de ser sancionados, expulsados del grupo o no volver a ser incluidos en interacciones cooperativas beneficiosas. Por otro lado, los defensores del constructivismo evolutivo parecen compartir la idea de que evolución cultural viene asociada a la selección de grupo. Es decir, los mecanismos selectivos que impulsan la evolución cultural no operan a nivel de individuo sino a nivel de grupo (Heyes 2019, Zawidzki 2013). Esto hace que grupos más cohesionados, cuyos individuos son más cooperativos y coordinados tienen una ventaja sobre aquellos cuyos individuos son más competitivos entre ellos. En este contexto, que la teoría de la mente diera una ventaja para iterar estados mentales para propiciar el engaño a los individuos de un mismo grupo no parece ser una estrategia evolutivamente estable.

En resumen, aunque la facultad para atribuir estados mentales está relacionada con diferentes habilidades que mejoran sustancialmente nuestra capacidad para entender a los demás y comparar las mentes de otros individuos, no parece que esto sea suficiente para respaldar la idea de que la evolución cultural fue motivada por la selección de aquellos individuos que tuvieran un mejor modelo de otras mentes y una mejor predicción. A esto, además, se suma la evidencia empírica apuntada en la sección 3 que sugiere que nuestras representaciones de otras mentes están sistemáticamente sesgadas.

5. Atribución y Compromisos Prácticos

Fenici y Zawidzki (2020) proponen una historia diferente de la evolución cultural de la lectura de mentes en la que toman, como punto de partida, una de las intuiciones de Gordon (1986) sobre cómo normalmente la capacidad de atribuir intenciones a otra persona es un producto del mismo tipo de razonamiento práctico que usamos para decidir sobre nuestras propias intenciones. Siguiendo a Evans (1982), Gordon considera que las auto-atribuciones no están basadas en la introspección, sino que vienen determinadas por la evaluación de nuestra propia situación —es decir, respondemos a la pregunta sobre si creo que P observando el mundo y no nuestros propios estados mentales—. Bajo esta misma premisa,

Gordon considera que las atribuciones de estados mentales en tercera persona funcionan de una manera análoga, son el producto de razonar sobre el contexto en los que otros agentes están inmersos y las consecuencias conductuales y prácticas de este.

De este análisis, Fenici y Zawidzki concluyen que la idea Gordon refleja una intuición de tipo semántico, a saber, que cuando evaluamos la verdad de las atribuciones de estados mentales del tipo "S cree que P" no necesitamos caracterizar los estados internos del sujeto. El significado de verbos como "creer" es deflacionista en el sentido de que son simplemente dispositivos lingüísticos que reciben su significado de la capacidad del atribuidor para activar su proceso de toma decisiones desde una perspectiva desplazada hacia el otro agente. Por tanto, la atribución de estados mentales no sirve para escudriñar los estados mentales de los demás sino para caracterizar la relación de un agente con su entorno. Sin embargo, esta posición plantea una pregunta ¿Por qué debemos atribuir estados mentales en lugar de hablar directamente sobre el comportamiento del agente, y el entorno en el que se produce?

La respuesta se encuentra en la función pragmática de los verbos de actitud proposicional. Imaginemos dos amigas Andrew y Barbara que intentan determinar a qué sitio van a ir a comer. Andrew propone ir a un restaurante indio, sin embargo, el restaurante está cerrado y, aunque Andrew no lo sabe, Barbara sí. Dado que la intención de Andrew no puede ser satisfecha, un movimiento natural de Barbara sería decir "el restaurante indio está cerrado". Sin embargo, nos dicen Fenici y Zawidzki, este tipo de aseveraciones conllevan un grado de compromiso fuerte. Si después de modificar los planes, Andrew descubriera que el restaurante indio en realidad, estaba abierto, Barbara se expondría al enfado o la sanción de Andrew. Algo parecido sucede con otros verbos intencionales como por ejemplo "pretender" (*intend*). Si Andrew manifestara su intención de ir al restaurante con la aserción "yo voy al restaurante indio", esto podría entenderse por parte de Barbara como una imposición por parte de Andrew, mientras que decir "Yo pretendía que fuéramos al restaurante indio" implica una invitación a negociar el plan y atenúa la fuerza de la aserción.

Ahora bien, ¿qué nos muestra esta práctica comunicativa en relación a las raíces evolutivas de la teoría de la mente? De acuerdo con Fenici y Zawidzki, la práctica verbal de atribuir intenciones y creencias ayuda a explicitar los compromisos prácticos de un agente, y por lo tanto juega un papel importante en el contexto de crear, compartir y negociar planes y objetivos conjuntos. Siguiendo a Gordon, esto significa que la práctica verbal de atribuir estados mentales tiene

como objetivo caracterizar las relaciones entre un agente y su entorno. Caracterizar los compromisos que un agente hace explícitos al crear, compartir y negociar conjuntamente planes expresan, en efecto, la forma en que un agente se inclina a actuar, y por lo tanto se caracterizan por cómo el agente está orientado hacia su entorno. En otras palabras, la emergencia de las adscripciones de estados mentales está ligada a la práctica comunicativa de explicitar compromisos prácticos con la acción, lo que se traduce en una mejora de la capacidad para crear planes y acciones conjuntas y, por tanto, coordinarnos mejor con otros agentes. Es precisamente esta mejora a la hora de coordinarnos lo que supone una ventaja en términos adaptativos y que por tanto podría explicar la evolución cultural de la teoría de la mente.

La posición de Fenici y Zawidzki supone una ventaja con respecto a la posición de Moore en el sentido de que explica cómo la teoría de la mente podría suponer una ventaja adaptativa en la mejora de la coordinación sin necesidad de hacer descansar dicha mejora en la idea de que la psicología popular nos permite escudriñar la mente de otros de una manera más efectiva. Es decir, se podría defender que las adscripciones son adaptativas porque mejoran la coordinación evitando el problema de tener que hacer compatibles la tesis de que las adscripciones generan representaciones fidedignas de la mente de otros y la evidencia que señala que nuestras adscripciones son frecuentemente sesgadas e imprecisas. A pesar de las ventajas, sin embargo, esta posición no está exenta del problema de la redundancia.

Como hemos visto anteriormente, la coordinación entre humanos en acciones conjuntas involucra una gran diversidad de mecanismos como la atención y percepción conjuntas, pero también, mecanismos de simulación o *affordances* compartidas (Knoblich y otros 2011, Vesper y otros 2017). Muchos de estos mecanismos no sólo involucran procesos de nivel bajo que sirven para la coordinación situada, sino también otros de nivel alto como la coordinación planeada a través de representaciones compartidas o el razonamiento colectivo que nos permite tomar decisiones como grupo (Pacherie 2011). Además, como el mismo Zawidzki (2008, 2013; ver también Fernández Castro y Heras-Escribano 2019 y McGeer 2007) ha argumentado, una parte importante de nuestra coordinación se realiza a través de una miríada de normas sociales y patrones culturales que regulan nuestra mente y nos permiten coordinarnos y predecirnos sin necesidad de leer la mente a los demás. Por último, nuestras emisiones lingüísticas portan compromisos prácticos con la acción que nos permiten coordinarnos de manera efectiva sin necesidad de usar conceptos mentales. En resumen, parece difícil ver cómo agentes capaces

de desplegar la teoría de la mente podrían ser capaces de coordinarse de manera radicalmente mejor en comparación con agentes que ya poseen estas habilidades en un sentido que suponga una diferencia efectiva que permitiera a la "selección cultural" hacer su trabajo.

6. Estados Mentales como Respaldo Social

Parece que las tres propuestas presentadas hasta ahora tienen problemas para explicar la ventaja adaptativa de la teoría de la mente como producto de la evolución cultural. Curiosamente, y en contra de la visión estándar en ciencia cognitiva, esto sugiere que es poco probable que la teoría de la mente emergiera como una herramienta para la coordinación o predicción, especialmente si asumimos que los individuos que la usaron por primera vez ya estaban dotados de mecanismos sofisticados para ello, incluido el lenguaje. Entonces, ¿Para qué sirve la teoría de la mente.

Para encontrar una alternativa plausible debemos mirar a un aspecto importante de lo dicho por Fenici y Zawidzki (2020): la función de los verbos de actitud proposicional es pragmática, no propiamente semántica. Esto significa que, en las oraciones como "creo que el restaurante indio está cerrado", el verbo "creer" no está describiendo un estado mental propiamente sino simplemente indicado un grado bajo de compromiso con la proposición "el restaurante indio está cerrado". Este uso de los verbos mentales en primera persona es lo que Urmson (1952) denominó usos parentéticos y que, como varios lingüistas defienden (Goddard 2003, Wierzbicka 2006), suponen un uso mayoritario de verbos como creer (*believe, think*) o conocer (*know*). Los verbos en uso parentético sirven para modular la interpretación de la proposición que cae bajo el alcance del verbo. Como indica Wierzbicka (2006), el verbo "creer" en su uso parentético sirve para negar nuestro conocimiento de algo, pero no diciendo "no lo se", sino diciendo "No digo: lo sé". Es decir, los usos parentéticos de los verbos "creer" o "conocer" expresan nuestro grado (in)certidumbre.

Este tipo de usos de los verbos de actitud proposicional parecen tener una función social interesante como instrumentos para el resguardo social. Decir "creo que el restaurante indio está cerrado" frente a "el restaurante indio está cerrado", nos permite hacer la misma contribución conversacional, pero evitando posibles sanciones derivadas de los compromisos prácticos de los que nos hacemos responsables cuando emitimos la proposición en cuestión.

En Almagro-Holgado y Fernández Castro (2020; ver también Fernández Castro 2017, 2019), he argumentado que esos *usos protectores* no son exclusivos de la

primera persona, sino que existen usos en tercera persona que apuntan al resguardo social como función principal de la teoría de la mente. En dicho artículo presentamos dos usos. En primer lugar, usamos adscripciones de creencia para proporcionar evidencia indirecta de nuestras proposiciones (e.g. Hazlett 2010, Simons 2007). En estas atribuciones, el hablante apoya la proposición expresada en la frase de la cláusula afirmando que una tercera persona también la apoya. Simons (2007) sostiene que esta interpretación de las atribuciones de creencia es la más plausible en contextos como el siguiente.

A: ¿Está abierta la biblioteca?

B: No lo sé con seguridad. Kautar cree que si está abierta.

Dar evidencia indirecta sirve para reforzar cierto patrón de acción, pero también para evitar asumir algunos de las responsabilidades involucradas en la aseveración "la biblioteca está abierta" y, por tanto, las posibles sanciones o consecuencias negativas asociadas a que el interlocutor descubriera a posteriori la falsedad de la proposición. Así mismo, las atribuciones de creencia sirven para llevar a cabo aseveraciones vicarias (Gauker y Van Cleave 2010, Tooming 2014, Fernández Castro 2019) o citar a otros (Geurts en prensa) lo que también sirve para aprovechar la autoridad de la tercera persona para justificar la acción de uno mismo (*El jefe cree que deberíamos salir ya*). Lo mismo se aplica al caso de la adscripción de deseos, que podemos usar para llevar a cabo órdenes por boca de otros (*Mamá quiere que recojamos la habitación*).

En segundo lugar, usamos las atribuciones de estados mentales para colocar a otra persona o su comportamiento bajo una luz positiva, es decir, para expresar una cierta actitud positiva o exculpatoria hacia la conducta de alguien. Esto sucede, por ejemplo, en casos donde intentamos justificar una conducta extraña o que viola una norma (*Antonio se saltó el semáforo porque pensaba que llegábamos tarde; Ana corrió hacia la cafetería porque quería ir al baño).* Nótese que, en estos casos, las atribuciones de estados mentales no son una mera explicación neutral del comportamiento en cuestión, sino que normalmente involucran una actitud positiva, o al menos, una asunción de que la conducta es cuestión es racional o está aceptada, aunque esta sea errada, incorrecta o extraña. En este sentido, las atribuciones no funcionan como meras explicaciones sino como justificaciones o disculpas de cursos de acción. De nuevo, las atribuciones de estados mentales tienen una función de resguardo social. La existencia de esta función de las atribuciones de tercera persona viene respaldada por evidencia empírica que muestra que usamos más explicaciones en términos mentales —en contraste con explicaciones de tipo

histórico causal, por ejemplo— cuando se nos pide que coloquemos al sujeto de la acción bajo una luz positiva (Malle y otros 2007) o cuando el comportamiento que tenemos que explicar es contra-normativo (Korman y Malle 2016). Por tanto, no sólo los usos en primera persona apuntados por Fenici y Zawidzki exhiben la función de resguardo social.

Estos tres usos de adscripciones mentales nos ponen en la pista de una de las posibles razones por las cuales la evolución cultural pudo promover la aparición de la teoría de la mente: sus usuarios podrían ser capaces de defender y justificar mejor sus cursos de acción de posibles sanciones e interpelaciones de otros agentes lo que, en general, podría mejorar el manejo de las impresiones que los otros agentes se llevan de ellos y, por tanto, mejorar su reputación. De acuerdo con el modelo de del resguardo social, entonces, las atribuciones de estados mentales protegen nuestro estatus social y reputación o los de otros agentes, evitando la responsabilidad, utilizando a un tercero como fuente de autoridad cuando se trata de apoyar una razón particular, o racionalizando una acción particular.

Ahora bien, ¿en qué sentido este resguardo social supone una ventaja suficiente para motivar la evolución cultural de la teoría de la mente? El resguardo social sirve para promover una imagen positiva de quien la usa. La reputación es un mecanismo conocido dentro de las explicaciones evolutivas de la cooperación —por ejemplo, en las teorías de la reciprocidad indirecta (Nowak y Sigmund 2005) o el altruismo competitivo (Barclay y Willer 2007)—. Estas teorías sugieren que las personas cooperen para mantener una buena reputación en su entorno social, donde esta reputación, a su vez, atrae a valiosos socios y aliados, lo que afecta positivamente a sus beneficios futuros. Esas opiniones explican por qué la proyección de una imagen positiva por parte de las personas sería esencial en las interacciones sociales. En ese contexto, poseer diferentes estrategias para rehabilitar nuestra condición social, evitar posibles sanciones o expresar nuestro cumplimiento de las normas parece ser una capacidad indispensable para incluir en nuestro repertorio social. En este sentido, la función de la teoría de la mente es proporcionar diferentes explicaciones, racionalizaciones, exculpaciones, anticipaciones o justificaciones para evitar responsabilidades o para cubrir la condición social del intérprete o del interpretado.

Esta posición tiene dos ventajas en comparación con las posiciones anteriores. En primer lugar, la teoría puede amoldar la evidencia sobre los sesgos que puede influenciar nuestras adscripciones. De hecho, la teoría conecta bien con teorías más generales que asocian nuestra capacidad de razonar con la justificación social más que con la búsqueda de la verdad. Mercier y Sperber (2017, ver también

Norman 2016) han argumentado que muchos de nuestros sesgos —por ejemplo, la tendencia a proporcionar razones que refuerzan nuestras creencias previas (Nickerson 1988)— sugieren que la capacidad humana para proporcionar razones no evolucionó para derivar información fiable o ayudarnos a tomar mejores decisiones. En su lugar, "el razonamiento puede conducir a resultados deficientes no porque los humanos sean malos en ello, sino porque buscan sistemáticamente argumentos para justificar sus creencias o sus acciones" (Mercier y Sperber 2011, p. 72). Desde este punto de vista, evaluamos y proporcionamos razones con la intención de persuadir a otros y, por consiguiente, nuestra competencia para dar y evaluar razones está también relacionada con la justificación de creencias y comportamientos en contextos de grupo (Mercier y Sperber 2011, pp. 62-63).

En segundo lugar, esta posición evita la acusación de redundancia en tanto que no asocia la emergencia de la teoría de la mente a funciones que podrían estar desempeñadas por otros mecanismos. Aunque, como argumentan Fenici y Zawidzki, las atribuciones de estados mentales nos ayudan a hacer explícitos nuestros compromisos prácticos y, por tanto, a mejorar nuestra coordinación, esta no es la función principal de las atribuciones. Además, el modelo del resguardo social nos permite evitar ciertos compromisos con el adaptacionismo de Heyes. Como algunos de sus críticos han argumentado, los mecanismos de evolución cultural, al contrario que la evolución biológica, no necesariamente promueven aquellos mecanismos que suponen una ventaja adaptativa o son seleccionados de manera ciega, sino que, en muchas ocasiones, un mecanismo psicológico o un comportamiento se disemina o selecciona dentro de una población simplemente porque es más atractivo desde un punto psicológico. En este sentido, utilizar las atribuciones mentales como resguardo social podría ser una estrategia atractiva psicológicamente, lo que la hace la teoría plausible tanto desde del punto de vista adaptacionista como desde el punto de vista de sus críticos.

En resumen, la propuesta de la emergencia de la teoría de la mente como resguardo social parece una alternativa plausible dentro del constructivismo evolutivo. En primer lugar, la posición presenta una estrategia evolutivamente estable razonable sobre cómo la teoría de la mente podría haber evolucionado como herramienta lingüística. En segundo lugar, frente a otras posiciones, la propuesta parece poder amoldar tanto la evidencia a favor del constructivismo como la evidencia problemática con las otras propuestas: los sesgos involucrados en nuestras adscripciones.

7. Reputación, resguardo y la metafísica de lo mental

En esta sección, intentaremos avanzar dos posibles objeciones al modelo del resguardo social. La primera apunta a ciertas incoherencias del modelo como estrategia evolutivamente estable para la evolución cultural. La segunda apunta a que la teoría conlleva ciertos compromisos metafísicos indeseables como algún tipo de ficcionalismo con respecto a lo mental.

7.1. Reputación y Selección de Grupo

Uno podría pensar que dado que no todos los usos de atribuciones de estado mental en tercera persona promueven una imagen positiva del individuo que los produce, no hay razones por las que esta estrategia sea adaptativa para dicho individuo. Aunque la reputación parece ser un mecanismo efectivo para promover la colaboración y cooperación entre individuos es una estrategia de tipo individual. Esto significa que podría parecer extraño que las atribuciones de tercera persona estuvieran orientadas a justificar la conducta de otros¿qué tiene de adaptativo para la persona que adscribe el promover una imagen positiva de otros agentes? La respuesta tampoco parece estar en la selección de grupo, ya que no queda claro cómo promover la justificación de la conducta de otra fomenta la adaptación del grupo a menos que se dieran ocasiones donde hubiera interacciones sociales entre miembros de diferentes grupos de manera que proteger al miembro del grupo propio supusiera una diferencia[7].

Existen varias modo de responder a esta objeción. En primer lugar, el mismo manejo de la reputación podría servir como motivación para prestar estrategias de protección social a terceras personas. Intentar justificar la conducta de terceras personas y, por tanto, promover una imagen positiva de ellos, también presenta al que atribuye como alguien deseable como compañero. Que los otros te perciban

[7] La selección de grupo necesita de condiciones bastante particulares. En un primer paso, se necesita que los individuos tengan motivaciones individuales para cooperar. En un segundo paso, una vez que cooperar fuese adaptativo, se necesita un periodo de balcanización, donde varios grupos se separaran y compitieran entre ellos (Zawidzki 2013, pp. 99-136). Uno podría pensar que fue después de este periodo, una vez los grupos empezaron de nuevo a tener relaciones no competitivas entre ellas (p.ej. transacciones comerciales), cuando la capacidad de justificar las conductas de los miembros del grupo propio comenzó a ser útil. Sin embargo, parece extraño que este tipo de contextos, al principio esporádicos en la historia evolutiva humana y donde los grupos no tendrían por que hablar el mismo idioma, se pudiera promover la utilización de teoría de la mente como justificación de otros.

como alguien que presta cobertura social a su compañero, te hace al mismo tiempo promover una imagen positiva tuya como compañero fiel y leal. Esta estrategia, además, se ve reforzada en individuos que buscan activamente las interacciones con otros o tienen una necesidad intrínseca de buscar relaciones estables y equilibradas (Baumeister y Leary 1995, Fernández Castro y Pacherie 2020).

En segundo lugar, justificar la conducta de otros tiene otro efecto beneficioso a nivel individual pero también de grupo. Cuando justificamos la conducta de los demás usando las mismas justificaciones que usaríamos para exculpar nuestra conducta, estamos promoviendo, dentro de la población, el tipo de comportamientos que se alinean con nuestros propios compromisos prácticos, lo que podría beneficiarnos a largo plazo. Si entendemos que la violación de una norma social está justificada en ciertos contextos y bajo ciertos compromisos prácticos, al defender a otros agentes bajo las mismas premisas, estamos evitando posibles sanciones a nosotros mismos en esos contextos. Pero, además, este tipo de estrategia tiene un impacto importante a nivel de selección de grupo: la homogeneización de la conducta dentro del grupo. Como el mismo Zawidzki (2013, ver también Peters 2019) ha argumentado, justificar ciertas conductas normativas y contra-normativas ayuda a que los individuos de un grupo se comporten de la misma manera en los mismos contextos. Esto tiene dos ventajas fundamentales para el grupo. Por un lado, la homogenización del comportamiento conlleva una facilitación de la predicción entre individuos, no porque puedan tener una imagen más fidedigna de los estados mentales de los demás, sino porque se comportan de a cuerdo a las mismas normas y pautas. Por otro lado, los grupos más homogéneos están más cohesionados lo que a su vez supone un mayor sentimiento de pertenencia y una mayor cooperación. Esto hace que los individuos de grupos dónde se promueven prácticas justificativas estén mejor respaldados que los de los individuos de grupos dónde no existen tales prácticas.

7.2. La Metafísica de la Mente

Uno de las principales lecciones que se recogen de la teoría del resguardo, pero también de la posición de Fenici y Zawidzki, es la idea de que la teoría de la mente tiene que ver con la atribución de compromisos prácticos y obligaciones más que con la atribución de estados psicológicos. Esto implica que atribuir estados mentales tiene que ver con monitorizar la responsabilidad de los agentes. Además, ambas posiciones están fuertemente comprometidas con la idea de que las atribuciones de estados mentales tienen un componente regulativo. Es decir, cuando atribuimos estados mentales a los demás estamos proyectando cierto tipo

de expectativas sobre su conducta de manera que cuando estos no se comportan como esperaríamos intentamos sancionarla o pedir explicaciones. Al mismo tiempo, comportarse de acuerdo con una adscripción significa regular la conducta de acuerdo a los compromisos prácticos asociados a ella. De ese modo, poseer una creencia o regular la conducta de uno de acuerdo con ella viene dictaminado por las interacciones sociales y los aspectos normativos dependientes de la comunidad (Zawidzki 2013, Fernández Castro 2020). Frente a la posición de Moore o Heyes, la teoría del resguardo y la posición de Fenici y Zawidzki implican que el significado de las adscripciones mentales no viene determinado por la descripción de los estados psicológicos del sujeto al que se le adscribe sino por las expectativas asociadas a los compromisos prácticos que se le presuponen y que vienen parcialmente determinada por la práctica social dentro de la comunidad.

Del hecho de que el significado de nuestras atribuciones de estados mentales no venga fijado por estados psicológicos independientes parece que nos compromete con la idea de que las atribuciones de estados mentales no son más que ficciones útiles que no tienen ninguna contrapartida real en el mundo (McCulloch 1990). Esta línea de argumento pondría a las posiciones constructivistas ante un problema importante, el de asumir que nuestra explicación presupone la existencia de ficciones útiles lo que conlleva una carga ontológica importante. Además, la aceptación de que las atribuciones de estados mentales son ficciones nos conduce a una aceptación de arbitrariedad insostenible. Las atribuciones de estados mentales y su relación con el comportamiento serían arbitrarias y convencionales de manera que la variación individual entre atribución y comportamiento sería masiva, convirtiendo cualquier función de justificación o explicación intencional en algo intratable.

Sin embargo, como Fenici y Zawidzki argumentan, que la relación entre la atribución de estados mentales y el comportamiento venga determinada por la práctica social no significa que no pueda ser objetiva y racional. De hecho, siguiendo a Brandom (1994), las atribuciones dependen de las obligaciones y compromisos prácticos y objetivos que guían nuestra práctica discursiva. En este sentido, los compromisos y obligaciones que atribuimos son racionales y objetivos en tanto que no dependen de la psicología del adscriptor si no de las normas que guían nuestra práctica discursiva y el comportamiento racional que se deriva de los compromisos y obligaciones que asumimos con esas prácticas. Cuando atribuimos a alguien la creencia de que P, lo estamos responsabilizando y asignando las obligaciones que se siguen racionalmente de que P sea el caso, es decir, actuamos como si el atribuido hubiera aseverado P.

Esta concepción no sólo nos permite escapar de la acusación de arbitrariedad, sino que nos permite escapar de toda especulación metafísica que se pudiera seguir de nuestras adscripciones. Una vez que entendemos que las adscripciones involucran el mismo tipo de compromisos y obligaciones que asociamos al contenido discursivo de la proposición incrustada; y entendemos el vocabulario mental en términos que involucran la expresión de actitudes positivas o de respaldo hacia esos compromisos y obligaciones, podemos abrazar un cierto tipo de *anti-descriptivismo* o *expresivismo* (Frápolli 2019, Frápolli y Villanueva 2012, Pinedo 2014) que evita compromisos metafísicos, bien sean estos realistas, ficcionalistas o eliminativistas. Dicho de otro modo, la tesis de que el significado del vocabulario mental debe ser entendido en términos de actitudes de respaldo o desapego de un contenido, nos libra de entenderlo en términos descriptivos, esto es, en términos de los objetos o estados del mundo que esos conceptos pudieran representar. Por tanto, si la función principal del vocabulario psicológico no es describir, no tenemos que presuponer ningún tipo de objeto o estado de cosas que ese vocabulario describa.

8. Conclusiones

La idea de que la teoría de la mente no es un producto de la selección natural mediada por mecanismos de herencia genética, sino fruto de la evolución cultural viene empíricamente respaldada por distintas fuentes de evidencia que van desde la variabilidad cultural hasta diversos tipos de experimentos que apuntan a que adscribir estados mentales requiere de un vehículo lingüístico. Esta idea abre un amplio abanico de preguntas y posibilidades teóricas relacionadas tanto con aspectos fundamentales de la evolución cultural en general como con los origines filogenéticos de la teoría de la mente.

Recientemente, podemos encontrar en filosofía y ciencia cognitiva diversas posiciones sobre cómo la evolución cultural podría haber motivado la emergencia de la psicología popular. En este ensayo, he intentado evaluar críticamente tres propuestas teóricas al mismo tiempo que he defendido una posición propia. De acuerdo con esa posición, que he denominado *el Modelo del Resguardo Social*, la evolución cultural de la teoría de la mente está asociada a ciertas prácticas lingüísticas que involucran la explicitación de compromisos prácticos y el respaldo social mediante estrategias como la justificación o la evasión de la responsabilidad. Este tipo de prácticas nos permiten evitar alguno de los problemas potenciales de las otras posiciones mientras que nos permiten dar una explicación constructivista

interesante de cómo la teoría de la mente se pudo dispersar siendo ventajosa tanto a nivel individual como de grupo.

Víctor Fernández Castro
Universidad de Granada
vfernandezcastro@gmail.com

Bibliografía

Almagro-Holgado, M. y Fernandez Castro, V. (2019): "The Social Cover View: a Non-epistemic Approach to Mindreading", *Philosophia*. Online First.

Andrews, K. (2012): *Do Apes Read Minds? Toward a New Folk Psychology*, Cambridge: MIT Press.

Astington, J. W. y Jenkins, J.M. (1999): "A longitudinal study of the relation between language and theory-of-mind development", *Developmental psychology*, nº 35(5), pp. 1311-1320.

Barclay, P. y Willer, R. (2007): "Partner choice creates competitive altruism in humans". *Proceedings of the Royal Society B: Biological Sciences*, nº 274, pp. 749-753.

Baumeister, R. F. y Leary, M. R. (1995): "The need to belong: desire for interpersonal attachments as a fundamental human motivation", *Psychological Bulletin,* nº 117, pp. 497-529.

Brandom, R. B. (1994): *Making it explicit: Reasoning, representing, and discursive commitment*, Cambridge: Harvard University Press.

Byrne, R. W. y Rapaport, L. G. (2011): "What are we learning from teaching?" *Animal Behaviour,* nº 82(5), pp. 1207-1211.

Claidière, N., Scott-Phillips, T.C. y Sperber, D. (2014): "How Darwinian is cultural evolution?" *Philos. Trans. Royal Soc. B*, nº 369 (1642): ID 20130368.

De Villiers, J.G. y De Villiers, P.A. (2000): "Linguistic determinism and the understanding of false" K. Riggs y P. Mitcheell (Eds.), *Children's reasoning and the mind*, New York: Psychology Press, pp. 191-228.

De Villiers, J.G. y Pyers, J. E. (2002) "Complements to cognition: A longitudinal study of the relationship between complex syntax and false-belief-understanding", *Cognitive Development*, nº 17(1), pp. 1037-1060.

Dunning, D. (1999): "A newer look: Motivated social cognition and the schematic representation of social concepts", *Psychological Inquiry*, nº 10(1), pp. 1-1.

Evans, G. (1982): *The varieties of reference*, New York: Oxford University Press.

Fernández Castro, V. (2017): "The expressive function of folk psychology" Unisinos, nº 18 (1), pp. 36-46.

Fernández Castro, V. (2019): "Justification, conversation and folk psychology". *Theoria,* nº *34*(1), pp. 75-91.

Fernández Castro, V. (2020): "Regulation, normativity and folk psychology". *Topoi,* nº 39(1), pp. 57-67.

Fernández Castro, V. y Heras-Escribano, M. (2020): "Social Cognition: a normative approach. *Acta Analytica"*, nº 35(1), pp. 75-100.

Fernández Castro, V. y Pacherie, E. (2020): "Joint actions, commitments and the need to belong", *Synthese*. Online first.

Frápolli, M.J. (2019): "The Pragmatic Gettier: Brandom on Knowledge and Belief", *Disputatio. Philosophical Research Bulletin*, Vol. 8, nº. 9, pp. 00-00.

Frápolli, M.J. y Villanueva Fernández, N. (2012): Minimal expressivism. *Dialectica*, nº 66(4), pp. 471-487.

Geurts, B. (en prensa): "First saying, then believing: the pragmatic roots of folk psychology".

Gopnik, G. y Meltzoff, A. (1997): *Words, thoughts, and theories*, Cambridge, MA: MIT Press.

Gordon, R. M. (1986): "Folk psychology as simulation", *Mind and Language,* nº 1 (2), pp. 158-171.

Hale, C.M. y Tager-Flusberg, H. (2003): "The influence of language on theory of mind: A training study", *Developmental science*, nº 6(3), pp. 346-359.

Happé, F.G.E. (1995): "The role of age and verbal ability in the theory of mind task performance of subjects with autism", *Child Development*, nº 66(3), pp. 843-855.

Hazlett, A. (2010): "The myth of Factive verbs", *Philosophy and Phenomenological Research,* nº 80, pp. 497-522.

Heyes, C. M. (2014): "Submentalizing: I am not really reading your mind", *Perspectives on Psychological Science,* nº 9(2), pp. 131-143.

Heyes, C. M. (2019): *Cognitive gadgets: The cultural evolution of thinking*. Cambridge, MA: Belknap Press.

Heyes, C. M. y Frith, C. D. (2014): "The cultural evolution of mind reading", *Science,* nº 344(6190), ID: 1243091.

Knoblich, G., Butterfill, S. y Sebanz, N. (2011): "Psychological Research on Joint Action", *Psychology of Learning and Motivation,* Vol. 54, pp. 59-101.

Korman, J. y Malle, B.F. (2016): "Grasping for traits or reasons? how people grapple with puzzling social behaviors", *Personality and Social Psychology Bulletin*, nº 42 (11), pp. 1451-1465.

Lavelle, J.S. (2019): "The impact of culture on mindreading", *Synthese*, on line first.

Lebra, T.S. (1993): "Culture, self, and communication in japan and the united states" en W.B. Gudykunst (Ed.), *Communication in Japan and the United States*, Albany, NY: SUNY Press, pp. 51-86.

Malle, B. M., Knobe, J. y Nelson, S. (2007): "Actor-observer asymmetries in explanations of behavior: New answers to an old question", *Journal of Personality and Social Psychology*, nº 93, pp. 491-514.

MAYNARD-SMITH, J. M. (1982): *Evolution and the theory of games*, Oxford: Cambridge University Press.

MCCULLOCH, G. (1990): "Dennett's Little Grains of Salt", *The Philosophical Quarterly*, nº40 (158), pp. 1-12.

MCGEER, V. (2007): "The regulative dimension of folk psychology", en D. D. HUTTO y M. RATCLIFFE (eds.), *Folk Psychology Re-Assessed*, Amsterdam: Kluwer/Springer Press, pp. 137-156.

MERCIER, H. y SPERBER, D. (2017): *The enigma of reason*, Cambridge, MA: Harvard University Press.

MOLL, H. y Meltzoff, A. (2011). "How does it look? Level 2 perspective-taking at 36 months of age", *Child Development,* nº 82(2), pp. 661-673.

MORIN, O. (2019): "Did social cognition evolve by cultural group selection?" *Mind and Language*, nº 34 (4), pp. 530-539.

MORTON, A. (1996): "Folk psychology is not a predictive device", *Mind,* nº 105(417), pp. 119-137.

MOORE, R. (2020): "The cultural evolution of mind-modelling", *Synthese*. On line first.

NICKERSON, R. S. (1988): "Confirmation bias: A ubiquitous phenomenon in many guises", *Review of General Psychology*, nº2(2), pp. 175-220.

NORMAN, A. (2016): "Why we reason: Intention-alignment and the genesis of human rationality", Biology and Philosophy, nº 31(5), pp. 685-704.

NOWAK, M. A. y SIGMUND, K. (2005): "Evolution of indirect reciprocity", *Nature*, nº 437, pp. 1291-1298.

OLIVOLA, C. Y. y TODOROV, A. (2010): "Fooled by first impressions? Reexamining the diagnostic value of appearance-based inferences", *Journal of Experimental Social Psychology*, nº 46(2), pp. 315-324.

ONISHI, K.H. y BAILLARGEON, R. (2005): "Do 15-month-old infants understand false beliefs?", *Science*, nº 308(5719), pp. 255-258.

PACHERIE, E. (2011): "Framing Joint Action" *Review of Philosophy and Psychology*, nº*2*(2), pp. 173-192.

PERNER, J. LEEKAM, S. R. y WIMMER, H. (1987): "Three-year-old's' difficulty with false belief: The case for a conceptual deficit", *British Journal of Developmental Psychology*, nº 5 (2), pp. 125-137.

PETERS, U. (2019): "The complementarity of mindshaping and mindreading" *Phenomenology and the Cognitive Sciences,* nº 18(3), pp. 533-549.

PINEDO, M. (2014): "¡No es un algo, pero tampoco es una nada! Mente y normatividad" *Análisis. Revista de investigación filosófica*, nº 1(1), pp. 121-160.

PROGOVAC, L. (2015): *Evolutionary syntax*, Oxford: Oxford UP.

RULE, N. O., AMBADY, N. y ADAMS, R. B., Jr. (2009): "Personality in perspective: Judgmental consistency across orientations of the face", *Perception*, nº 38, pp. 1688-1699.

SHAHAEIAN, A. C., PETERSON, C., SLAUGHTER, V. y WELLMAN, H.M. (2011): "Culture and the sequence of steps in theory of mind development", Developmental *Psychology,* nº 47, 1239-1247.

Spaulding, S. (2018): *How we understand others: Philosophy and social cognition*, London and New York: Routledge Focus.

Sperber, D. (1996): *Explaining culture. A naturalistic approach*, Oxford, UK: Blackwell.

Sperber, D. (2000): "Metarepresentations in an evolutionary perspective" en D. Sperber (Ed.), *Metarepresentations: A Multidisciplinary Perspective.* Oxford: OUP, pp. 117-137.

Simons, M. (2007): "Observatio≠ns on embedding verbs, evidentiality, and presupposition", *Lingua, nº 117*(6), pp. 1034-1056.

Taumoepeau, M. y Ruffman, T. (2008): "Stepping stones to others' minds: Maternal talk relates to child mental state language and emotion understanding at 15, 24, and 33 months" *Child Development*, nº 79, pp. 284-302.

Tooming, U. (2016): "Mental State Attribution for Interactionism", *Studia Philosophica Stonica,* nº 9(1), pp. 184-207.

Urmson, J. O. (1952): "Parenthetical verbs", *Mind*, nº 61(244), pp. 480-496.

Van Cleave, M. y Gauker, C. (2010): "Linguistic practice and false-belief tasks", *Mind and Language, nº 25*(3), pp. 298-328.

Vesper, C., Abramova, E., Bütepage, J., Ciardo, F., Crossey, B., Effenberg, A., Hristova, D., Karlinsky, A., McEllin, L., Nijssen, S. R. R., Schmitz, L. y Wahn, B. (2017): "Joint Action: Mental Representations, Shared Information and General Mechanisms for Coordinating with Others", *Frontiers in Psychology*, vol. *07*, 2039.

Wellman, H. R. (2014): *Making minds: How theory of mind develops*. NY: Oxford University Press.

Westra, E. (2017): "Stereotypes, theory of mind, and the action-prediction hierarchy", *Synthese*. On line first.

Wierzbicka, A. (2006): *English: Meaning and Culture*, Oxford University Press.

Wimmer, H. y Perner, J. (1983): "Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception", *Cognition*, nº13 (1), pp. 103-128.

Zawidzki, T.W. (2008): "The function of folk psychology: Mind reading or mind shaping?", *Philosophical Explorations*, nº 11(3), pp. 193-210.

Zawidzki, T.W. (2013): *Mindshaping: A New Framework for Understanding Human Social Cognition*, Cambridge: MIT Press, A Bradford Book.