UNIVERSIDAD NACIONAL DEL SUR

TESIS DE DOCTORA EN FILOSOFÍA

Mindreading y evidencia neurocientífica: en defensa de los modelos híbridos de teoría y simulación

Fernanda Velázquez Coccia

BAHÍA BLANCA ARGENTINA

2015

PREFACIO

Esta Tesis se presenta como parte de los requisitos para optar al grado Académico de Doctora en Filosofía, de la Universidad Nacional del Sur y no ha sido presentada previamente para la obtención de otro título en esta Universidad u otra. La misma contiene los resultados obtenidos en investigaciones llevadas a cabo en el ámbito del Departamento de Humanidades durante el período comprendido entre el 3 de mayo de 2007 y el 9 de marzo de 2015, bajo la dirección de Dr. Oscar Esquisabel, Universidad Nacional de La Plata, y Dr. Gustavo Bodanza.

UNIVERSIDAD NACIONAL DEL SUR
Secretaría General de Posgrado y Educación Continua

La presente tesis ha sido aprobada el…../…../….., mereciendo la calificación de.....(……)

AGRADECIMIENTOS

Quisiera a gradecer a Oscar Esquisabel, director, y a Gustavo Bodanza, co-director de esta tesis. Le debo un profundo agradecimiento a Liza Skidelsky, por su generosidad y dedicación. Gracias a Sergio Barberis, Sabrina Haimovici, Mariela Destéfano, Abel Wajnerman y Nicolás Serrano por el entusiasmo en todos los intercambios intelectuales y no tanto, y a todos los integrantes del grupo de investigación “Cognición, Lenguaje y Percepción” (coordinado por Liza Skidelsky) que, con el compromiso, el esfuerzo y la dedicación, hacen de la tarea de investigación algo menos solitario.

Agradezco también a Susana Koreck, Marcelo Auday & Fernando Tohme por el apoyo inicial, la confianza y las charlas inspiradoras. Quisiera mencionar también a Diego Lawler, Bruno Wicker, Jorge Roetti que confiaron en mi en distintas presentaciones de proyectos de investigación y becas.

Gracias a mis Papás, por todo el apoyo, a Peter por la paciencia y mucho más, a Delfi y Lauti por la inspiración, a Gise y Maxi por el apoyo y la complicidad en los años porteños, a Jul y Nagui por el cariño recibido, a Joaco y Azul por preguntar siempre cómo va la tesis y saber del esfuerzo, a Lu & Marian por las pilas, a Nuria Chauvié y La Jirafita Pinta que lo cuidan tanto a Lauti.

RESUMEN

Mindreading es la capacidad de atribuir estados mentales, y de predecir y explicar el comportamiento y el pensamiento propio y ajeno, a partir de tales atribuciones. Una de las cuestiones principales en el estudio de *mindreading* es la indagación por los procesos subyacentes a esta capacidad. En el debate teoría-simulación se ha planteado una dicotomía. O bien subyacen procesos de tipo ricos en información, o bien procesos de tipo simulacional. Recientemente, se han propuesto enfoques híbridos que postulan el involucramiento de ambos tipos de procesos en *mindreading*. El propósito general de esta investigación es elucidar algunas de las similitudes y diferencias entre los enfoques híbridos de teoría y simulación de *mindreading*, con el objetivo de proponer una hipótesis híbrida enriquecida con los hallazgos empíricos. Así, esta investigación se inscribe en el ámbito de la filosofía de la ciencia cognitiva siguiendo la idea de que podemos obtener respuestas satisfactorias a preguntas filosóficas en relación a la mente al vincularlas con el trabajo empírico en psicología y neurociencias. Los objetivos particulares son: (a) desarrollar una propuesta sobre cuáles serían los requisitos de mínima de un enfoque híbrido de teoría y simulación de *mindreading*, (b) analizar los alcances y límites explicativos, en función de los requisitos de mínima establecidos en (a), de algunas propuestas teóricas en las que se adoptan posturas híbridas de teoría-simulación sobre *mindreading*; y (c) desarrollar los lineamientos generales de un enfoque híbrido que cumpla con los requisitos de mínima y atienda a los hallazgos empíricos en neurociencias. La presente investigación aborda los enfoques híbridos que postulan la intervención de procesos de simulación y, a la vez, de bases de información para realizar atribuciones de estados mentales. Para ello, en primer lugar se analizan los enfoques puros de teoría y simulación con el objetivo de establecer las razones que motivan la adopción de enfoques híbridos de teoría-simulación. En segundo lugar, se establecen los requisitos de mínima para una propuesta híbrida sobre la capacidad de atribución mentalista. Tercero, se evalúan las ventajas y limitaciones explicativas de las propuestas híbridas de teoría-simulación en función de los requisitos de mínima. Y cuarto, se desarrollan los lineamientos generales de un enfoque híbrido que cumpla con los requisitos de mínima y atienda a los hallazgos empíricos en neurociencias. De esta manera, me propongo seguir una línea metafilosófica en ciencias cognitivas que pone en evidencia la importancia del estudio interdisciplinario de la capacidad de atribución mentalista.

ABSTRACT

Mindreading is the capacity to attribute mental states and to predict and explain the behavior and thinking of self and others, based on such attributions. One of the main issues in the study of mindreading is the inquiry about the underlying processes of this capacity. In the early days of the simulation/theory-theory debate, the issue was characterized as dichotomous. Either mindreading derives from a theory or from simulation. Recently, the hybrid approaches postulate the involvement of both types of processes in mindreading. The general purpose of this research is to elucidate the similarities and differences between the hybrids approaches of mindreading, with the aim of proposing a hybrid hypothesis enriched with empirical findings. This research falls within the scope of the philosophy of cognitive science following the idea that we can get satisfactory answers to philosophical questions about the mind, by linking them with empirical research in psychology and neuroscience. The specific objectives are: (a) to develop a proposal for what would be the *minimum* requirements that a hybrid theory of mindreading needs to accommodate, (b) to analyze, based on the *minimum* requirements set out in (a), the explanatory scope and limits of some approaches in which hybrid positions on the simulation/theory-theory debate in mindreading are adopted; and (c) to develop the broad outlines of a hybrid approach that meets the minimum requirements and attends the empirical findings in neuroscience. This research addresses the hybrid approaches that postulate the intervention of simulation processes and, at the same time, information-rich processes to make attributions of mental states. To do this, first the pure theory and the pure simulation approaches are analyzed in order to establish the reasons for the adoption of hybrid theory-simulation approaches. Second, minimum requirements for a hybrid theory of mindreading are established. Third, the scope and limits of some hybrid approaches are evaluated based on the minimum requirements. And fourth, the general outlines of a hybrid approach that meets the minimum requirements and attend the empirical findings in neuroscience are developed. Thus, I propose to follow a metaphilosophical line of inquiry in cognitive science that demonstrates the importance of the interdisciplinary study of mindreading.

ÍNDICE

INTRODUCCIÓN

1. *Mindreading*: del debate teoría-simulación a los enfoques híbridos

En los últimos 30 años, ha habido un intenso debate en filosofía de la mente, psicología, primatología y neurociencias cognitivas sobre la naturaleza de *mindreading*[1]. La controversia central se refiere a la cuestión de explicar cómo es que podemos entender y predecir, sin esfuerzo, el comportamiento de las otras personas en la mayoría de las situaciones cotidianas. Al parecer, en la comprensión cotidiana de las otras personas usualmente vamos más allá de su comportamiento físico, y les adscribimos estados y procesos mentales. Ahora bien, los estados y procesos mentales no se pueden observar directamente y tampoco existe una correlación simple entre los estados mentales y la conducta. No obstante, se asume que "leer" la mente de los otros es importante porque los estados mentales son los causantes de las acciones.

En el ámbito de la filosofía del siglo XX, se ha sostenido que una teoría que reúne el conjunto de nociones de sentido común, o perogrulladas, sobre el funcionamiento de la mente subyace a *mindreading*. En debate con el empirismo lógico (Ryle 1949), Sellars (1956) propone que al adscribir estados mentales a otros para explicar y predecir su comportamiento, lo que hacemos es postular entidades teóricas tal como los científicos introducen términos teóricos para explicar y predecir fenómenos observables[2]. Esta idea de una proto-teoría científica subyacente a *mindreading* fue presentada por Sellars bajo la forma de un mito, el mito de "nuestro ancestro Jones". Según este relato, hubo un tiempo en el que entre nuestros ancestros *ryleanos* no se utilizaban términos mentales para entender a los demás, ni para dar

[1] El término "*mindreading"* se ha acuñado recientemente para referir a la capacidad de atribuir estados mentales y de explicar y predecir el comportamiento, propio y ajeno, en términos mentalistas. En filosofía, esta capacidad se suele denominar "psicología de sentido común" y, en este sentido, el término está estrechamente relacionado con el enfoque que postula una teoría psicológica subyacente. "*Mindreading*", en cambio, no se asocia a ningún enfoque y resulta más neutro. Por esta razón, considero que es más apropiado para nombrar a la competencia mentalista y lo adoptaré en esta tesis.

[2] El planteo de Sellars (1956) surge como una alternativa al enfoque del empirismo lógico que sostiene que los términos mentales no refieren a entidades internas sino a disposiciones para actuar (Ryle 1949).

cuenta de las acciones propias. Sin embargo, cierto día Jones advirtió que el comportamiento verbal explícito es la culminación de un proceso que comienza con el habla interna y creó una teoría según la cual las palabras manifiestas no son más que sucesos que comienzan con episodios internos (Sellars 1956). Así, los estados internos no observables, y en este sentido, referidos por términos "teóricos", son causa del comportamiento.

Posteriormente, en el marco de la discusión sobre cómo definir los términos teóricos, Lewis (1970, 1972) retoma este desarrollo y propone un método para definir funcionalmente los conceptos mentales. Según este enfoque, una teoría científica que introduce términos teóricos mentales subyace a *mindreading*, aunque inventada antes de que exista la ciencia. Esta teoría recoge las nociones de sentido común sobre cómo están relacionados causalmente los estados mentales con los estímulos sensoriales, las respuestas motoras y otros estados mentales. Así, los términos mentales cotidianos pueden ser definidos por el rol funcional/causal específico que ocupan en la teoría (Lewis 1972). En el capítulo 1 (sección 2), me ocuparé con detalle de la relación entre el funcionalismo, los conceptos mentales y la explicación de las capacidades mentalistas.

En el ámbito empírico, los primatólogos Premack & Woodruff (1978) propusieron que la capacidad de predecir el comportamiento en base a estados mentales (*mindreading*) supone una "teoría". Más precisamente, poseer una "teoría de la mente" significa que un individuo se adscribe a sí mismo y a los demás estados mentales mediante un sistema de inferencias que puede ser considerado una teoría, en tanto tales estados no son observables directamente y en tanto el sistema puede ser usado para realizar predicciones, específicamente, sobre el comportamiento de otros organismos (Premack & Woodruf 1978). A partir de este trabajo seminal, el estudio empírico de *mindreading* se ha regido por la idea de que la atribución de estados mentales depende de una teoría, y este ámbito específico de investigación pasó a denominarse "Teoría de la Mente". De este modo, la idea de un elemento teórico subyacente a *mindreading* surge no sólo en el ámbito filosófico sino también en el científico.

De acuerdo con la "Teoría de la Teoría" (en adelante TT), tal como adelanté, la atribución de estados mentales se lleva a cabo utilizando una teoría psicológica conformada por un conjunto de conceptos de entidades mentales (tales como los conceptos de creencia y deseo) y ciertos principios o reglas generales respecto de la interacción entre los mismos (por ejemplo, las personas actúan en base a sus creencias para satisfacer sus deseos)[3]. Particularmente, esta teoría psicológica rudimentaria consiste en los vínculos causales entre los *inputs* ambientales, los *outputs* comportamentales y otros estados internos. Haciendo uso de esta teoría y a partir de información apropiada sobre el agente que es blanco de la atribución, se subsumen los casos particulares a los principios generales por medio de algún tipo de inferencia. De este modo, se arriba a conclusiones (descripciones, predicciones o explicaciones) sobre el comportamiento (Fodor 1987; Stich & Nichols 1992, 1995, 1996a, 1996b; Botteril & Carruthers 2003).

Según la "Teoría de la Simulación" (en adelante TS), en cambio, la atribución de estados mentales a otras personas se lleva a cabo sin que sea necesario utilizar generalizaciones psicológicas, sino basta con la simulación mental (Gordon 1995a; Heal 1995a; Goldman 1995a). Ésta consiste en la replicación del pensamiento, la toma de decisiones, las respuestas emocionales u otros aspectos de la vida mental de otra persona que, usualmente, se produce en la imaginación. Particularmente, el tipo de simulación mental propuesto por la TS implica que los mismos recursos mentales que son utilizados en nuestro propio pensamiento, toma de decisiones o respuestas emocionales son re-utilizados en la imaginación para proveer el entendimiento del pensamiento, la toma de decisiones y las respuestas emocionales de las otras personas. Así, basta con ponerse uno mismo en la situación del otro y, de alguna

[3] Usualmente, los filósofos consideran a las creencias y a los deseos, que son estados de actitud proposicional, como los estados mentales paradigmáticos (aunque también se han ocupado extensamente del dolor que es un estado cualitativo). Ambos, las creencias y los deseos, representan estados posibles del mundo. Sin embargo, difieren en el modo de hacerlo. Siguiendo a Searle (1983), esta diferencia puede caracterizarse según la dirección de ajuste que cada estado establece con el mundo. Las creencias resultan verdaderas si el modo en que representan el mundo se "ajusta" al modo en que el mundo es. En cambio, los deseos están satisfechos si y sólo si el mundo se "ajusta" a ellos. Así, en el caso de las creencias se establece una dirección desde la mente hacia el mundo, mientras que en el caso de los deseos parece haber una dirección desde el mundo hacia la mente.

manera, uno sabe cómo actuar, sentir o pensar. Esto tiene lugar en la medida que es posible generar en uno mismo estados mentales similares a los ajenos, adoptando imaginativamente su perspectiva, para luego observar qué otros estados suceden. Estos últimos son atribuidos al blanco.

A los fines de esta tesis, me interesa el enfrentamiento entre los enfoques de la TT y la TS que intentan dar cuenta de los procesos que subyacen a *mindreading*[4]. Es preciso mencionar que, sin embargo, no existe un enfrentamiento simple entre dos enfoques. La TT no constituye un enfoque unívoco, más bien existen varias versiones del mismo. Entre las diferencias, es de importancia la concepción respecto del modo de adquisición de la teoría subyacente a *mindreading*. El enfoque usualmente denominado "del niño científico", también conocido como "enfoque empirista" o "de teorización", propone que la teoría psicológica subyacente es aprendida por un proceso análogo a la investigación científica, en el sentido de que el desarrollo de la comprensión mentalista en los niños se desenvuelve de manera análoga al cambio teórico en ciencia. En este marco, *mindreading* se constituye como una capacidad de dominio general para teorizar (Wellman 1990; Perner 1991; Gopnik & Wellman 1992; Gopnik & Meltzoff 1997). Contrariamente, según el "enfoque modular" la teoría psicológica subyacente no se aprende sino que forma parte de un equipamiento innato de módulos dedicados a mentalizar, que madura en el curso del desarrollo. En este marco, *mindreading* se concibe como una capacidad especializada de dominio específico (Leslie 1991, 1994; Baron-Cohen 1995; Carruthers 2006). Me ocuparé de las distintas versiones de la TT en el capítulo 1 (sección 3).

Por su parte, la noción de "simulación" varía ampliamente entre los proponentes del enfoque y hasta se ha sugerido en la literatura la necesidad de abandonar una noción tan difusa (Stich & Nichols 1997). No obstante, dado que la TS nace como una reacción a la TT, es posible caracterizar las diferentes versiones de TS en relación a aquellos ítems de la TT a los que se oponen. Esquemáticamente, la TT postula la posesión de un cuerpo de conocimiento sobre el funcionamiento de la

[4] Los artículos clásicos de la polémica pueden encontrarse en las siguientes compilaciones: Davies & Stone (1995a), Davies & Stone (1995b), Carruthers & Smith (1996).

mente que contiene (i) generalizaciones psicológicas que permiten (ii) realizar adscripciones de estados mentales usando conceptos mentales. En principio, todas las versiones de la TS niegan la necesidad de acudir a (i) generalizaciones psicológicas para entender y predecir el comportamiento y el pensamiento de otras personas, pero varían en su aceptación de (ii) la necesidad de utilizar conceptos mentales en las atribuciones mentalistas.

La versión más radical de la simulación (Gordon 1995a, 1995b, 1995c) niega (ii) y postula el procedimiento de las "rutinas de ascenso" (Evans 1982). Por medio de éste, se generan atribuciones sin utilizar conceptos de estados mentales en tanto este procedimiento permite responder a preguntas sobre estados mentales con respuestas basadas en los hechos sobre los que tratan los estados mentales. En la versión de Goldman (1993a, 1993b, 2006), en cambio, se acepta (ii), puesto que es necesario utilizar conceptos mentales en las atribuciones mentalistas. Por un lado, es preciso identificar y categorizar los estados mentales de los otros con el propósito de fijar los parámetros para una simulación, y para llevar a cabo estas tareas se requieren conceptos de estados mentales. Por otro lado, es preciso identificar y categorizar los estados mentales generados por los sistemas cognitivos propios, al finalizar la simulación y antes de que tenga lugar la proyección de los mismos al blanco. Me ocuparé con más detalle de este enfoque en el capítulo 2 (sección 2).

Ante la dificultad de confrontar dos enfoques entre los cuales no existe una simple oposición, desde una perspectiva teórica, no ha sido posible elegir entre TT y TS sopesando argumentos. No obstante, se ha intentado la comparación utilizando varios criterios. Algunos de estos criterios están relacionados con las virtudes generales de una teoría, como la simplicidad o el alcance empírico, entre otros. En línea con esto se ha señalado, por ejemplo, que la TT parece ser menos parsimoniosa que la TS en la medida en que postula la posesión de generalizaciones específicas sobre estados mentales, perdiendo de este modo simplicidad. A su vez, se ha señalado que la TS como propuesta teórica es difusa, ya que no existe una única noción de "simulación" entre los defensores del enfoque (Stich & Nichols 1997).

A su vez, desde una perspectiva empírica, los enfoques de TT y TS se han considerado excluyentes esperando que de ellos se desprendieran predicciones y explicaciones empíricas en competencia que, al ser testeadas, permitieran elegir entre los mismos. En este sentido, el trabajo empírico en relación a *mindreading* ha influido y estructurado el debate. Las disciplinas empíricas más influyentes han sido la psicología del desarrollo y cognitiva, y, más recientemente, la neurociencia cognitiva social. A fines de los '90 los psicólogos abandonaron el intento de elegir entre TT y TS, dado que los estudios no eran conclusivos para preferir alguna del las teorías, y empezaba a ser claro que ambos enfoques podían dar cuenta de los hallazgos empíricos. En aquel entonces las esperanzas se colocaron en las novedosas técnicas de neuroimagen (Stich & Nichols 1997; Apperly 2008). Sin embargo, hasta ahora estos estudios, de los que me ocuparé particularmente en el capítulo 6, tampoco han mostrado ser concluyentes.

Estas dificultades han conducido a que, este enfrentamiento entre teoría y simulación, que ha constituido el modo tradicional de abordar *mindreading*, en cierto sentido, se haya diluido. En la literatura sobre *mindreading* se puede observar la tendencia a abandonar posturas puras de teoría o de simulación, incluso entre quienes las sostuvieron alguna vez. A la base de este cambio subyace un amplio acuerdo respecto de que *mindreading* muestra ser un fenómeno complejo y multifacético. Por esto, es probable que ciertos aspectos de *mindreading* no puedan ser explicados por enfoques puros de teoría o de simulación. En este sentido, se asume que las cuestiones que no pueden ser abordadas por teorías puras, pueden serlo por enfoques mixtos de teoría y simulación (Stich & Nichols 1995; Nichols *et al.*1996; Perner 1996; Davies & Stone 1998; Nichols & Stich 2003). En términos más cognitivistas, la complejidad de la capacidad de atribución mentalista sugiere, especialmente, que ésta ya no puede concebirse como un sistema único basado en un único tipo de mecanismo (Goldman 2006; Apperly 2008; de Vignemont 2009).

Desde la filosofía, se han propuesto enfoques híbridos de teoría y simulación para dar cuenta de *mindreading* (Goldman 2006; Stich & Nichols 2003). En esta tesis, me interesa evaluar si los enfoques híbridos teoría y simulación logran una descripción

y explicación adecuadas de la arquitectura y los procesos (subpersonales) responsables de la capacidad de atribución mentalista[5]. Para ello revisaré en primer lugar las desventajas de la TT y la TS para que se logre una mejor comprensión de la necesidad de los enfoques híbridos. Luego, evaluaré las propuestas teóricas de los procesos subyacentes a *mindreading* que postulan enfoques híbridos de teoría y simulación. Así, voy a analizar los alcances y límites explicativos de algunos enfoques filosóficos en los que se adoptan posturas híbridas respecto del debate teoría-simulación sobre la atribución de estados mentales (Goldman 2006; Nichols & Stich 2003). A mi entender, la adecuación de las propuestas híbridas depende de que puedan brindar un criterio claro para poder distinguir el tipo de proceso subyacente a cierta instancia de *mindreading*. Sostendré que los enfoques híbridos que analizaré tienen dificultades para brindar un criterio que permita distinguir entre procesos subyacentes de teoría o simulación. Dada esta situación, ofreceré un criterio para distinguir entre teoría y simulación atendiendo a los hallazgos empíricos, particularmente en neurociencias.

[5] No todas las propuestas teóricas en relación a la atribución mentalista se ocupan de la pregunta por la arquitectura y los procesos subpersonales (por ejemplo, los enfoques fenomenológico y de cognición corporizada que mencionaré en la sección 1.3 de esta introducción). En este sentido, tampoco me ocuparé de la propuesta simulacional de Jane Heal. Este enfoque es una propuesta *a priori* y se ubica en el nivel personal de las capacidades cognitivas. Según Heal (1995a), poseemos conocimiento teórico sobre las personas, pero éste es general y no informa acerca de las creencias y propósitos de los individuos particulares. Por esta razón, este conocimiento no puede utilizarse para realizar predicciones sobre individuos particulares. En su lugar, para lidiar con el contenido de los estados mentales ajenos recurrimos a un conocimiento de tipo *know how* (intrínsecamente no traducible a conocimiento *know what*), que consiste en una estrategia imaginativa que utiliza a la capacidad intelectual de pensar en lo posible, y de explorar las consecuencias y ramificaciones de las situaciones consideradas meramente posibles para obtener un *insight* sobre las otras personas (estrategia de replicación). Así, cuando necesito saber qué pensará o decidirá cierta persona trato de imaginarme el mundo tal como se le aparece a ella misma, y de explorar algunos de los estados de cosas y requisitos, para la acción implicados en tal mundo. Si esto es llevado a cabo con éxito, habré recreado (en parte) el punto de vista de la persona en cuestión, su cadena de pensamientos y sus posibles decisiones. Este modo de acceder a la visión del mundo de las otras personas no requiere de una teoría acerca de la interacción entre estados mentales, ni de sus consecuencias. El uso de la imaginación para lograr un *insight* acerca de los pensamientos e intenciones ajenas, es posible en virtud del hecho de que las personas compartimos la capacidad intelectual de primer orden de pensar acerca de lo posible. En principio, esta estrategia implica que pensamos sobre el mundo y no sobre la mente del otro, establece un compromiso con ciertos conceptos mentales ("el otro es como yo") y es necesaria para pensar acerca de los contenidos y para dar cuenta de la racionalidad.

Al ocuparme de evaluar los enfoques híbridos, me propongo seguir una línea metafilosófica en ciencias cognitivas que pone en evidencia la importancia del estudio interdisciplinario de *mindreading*. El trabajo empírico en *mindreading* ha revelado que la suposición de que nuestra competencia psicológica cotidiana se explica postulando una psicología de sentido común compuesta de conceptos mentales y generalizaciones psicológicas es, al menos, problemática. Esta tesis se desarrolla siguiendo la idea de que podemos obtener respuestas satisfactorias a preguntas filosóficas en relación a la mente al vincularlas con el trabajo empírico en psicología y neurociencias. Esto es viable en el caso específico de *mindreading*, incluso si se toma como punto de partida las teorías empíricas sobre nuestra concepción ordinaria de la mente. Es preciso señalar que no creo que con esto se estén resolviendo todos los problemas vinculados a la mente sino, más bien, consideraré que la colaboración entre la filosofía y las ciencias cognitivas indica una dirección productiva de investigación. A continuación me referiré a los estudios empíricos más influyentes tanto de la psicología del desarrollo como de las neurociencias en relación al estudio de *mindreading*.

1.1. Los estudios clásicos en psicología

La TT ha encontrado apoyo, principalmente, en numerosos estudios de psicología del desarrollo. El impacto del artículo seminal de Premack & Woodruff (1978) se extendió del dominio de la primatología a los dominios de la psicología y la filosofía. En sus comentarios a este artículo, Dennett (1978), Bennett (1978) y Harman (1978) señalan cuáles deberían ser las condiciones con las que tiene que cumplir una conducta observable para poder inferir, a partir de la misma, que una criatura posee la noción de que los otros tienen mente. Primero, el organismo tiene que ser capaz de tener creencias acerca de las creencias de los otros y ser capaz de distinguirlas de las propias. Segundo, éste tiene que ser capaz de hacer o predecir algo en relación a las creencias atribuidas que distingue de las propias. En este sentido, una prueba adecuada debe basarse en la generación de una predicción sobre la conducta de otro

agente a partir de atribuirle a éste una creencia equivocada (o falsa). Si no fuera así, esto es, si la conducta estuviera relacionada sólo con el "estado de cosas en el mundo", ésta podría ser predecida en general en base al conocimiento de las regularidades que conectan la situación y la conducta, sin necesidad de apelar a estados internos de un organismo (*e.g.* creencias).

Los psicólogos Wimmer & Perner (1983) adaptaron la sugerencia de los filósofos y elaboraron un paradigma experimental para utilizar con niños, al que denominaron la "tarea de falsa creencia" (TFC). Originalmente, en esta tarea se narra una historia que involucra un personaje, Maxi, que coloca un objeto (un chocolate) en una ubicación particular de una habitación (alacena *x*) y se retira. Luego, en su ausencia, la mamá de Maxi mueve el chocolate de lugar (alacena *y*). Llegada esta instancia, se le pregunta al participante dónde buscará Maxi el chocolate cuando regrese. Para desempeñarse exitosamente en la tarea, el participante debe entender que Maxi cree que el chocolate está aún donde él lo colocó. Esto es, el participante debe entender que Maxi tiene una creencia errónea, es decir, una creencia falsa[6].

De este modo, la tarea se adecua a lo sugerido por los filósofos en tanto requiere que el participante sea capaz de realizar una predicción de la conducta del personaje (*viz.* dónde buscará Maxi el objeto) a partir de la creencia discrepante de que el objeto se encuentra ubicado donde el participante mismo sabe que, de hecho, no está. Contrariamente, si la tarea implicara una predicción basada en buscar un objeto dónde realmente está ubicado, esto no permitiría distinguir entre las creencias propias del participante y las del personaje porque no habría discrepancia. Como se mencionó, si este fuera el caso, bastaría con el conocimiento de los hechos para resolver la tarea y no habría necesidad de tener conocimiento de las creencias ajenas.

El hallazgo de Wimmer & Perner (1983) consiste en que los niños menores de tres años tienden a fallar sistemáticamente en la tarea de falsa creencia, mientras que los niños de cuatro años la resuelven con éxito. Para explicar esto, se postula la

[6] Existen numerosas variaciones de la TFC original, por ejemplo, la tarea de Sally-Anne donde la narración se acompaña con una historieta gráfica (Baron-Cohen, Leslie & Frith 1985). Al conjunto de tareas de falsa creencia de esta clase se lo conoce como tareas de falsa creencia "de cambio de lugar".

existencia de un cambio conceptual en los niños, entre los tres y los cuatro años, que implica la emergencia de la habilidad de representar creencias erróneas y de entender que las personas actúan en base a sus creencias. Según esta explicación, existe previamente una noción precaria de "creencia" que no contempla la posibilidad de distorsión en la representación del mundo. En este sentido, es una noción incompleta. Desde entonces, los psicólogos han replicado y redefinido esta tarea, y han tratado de hacer descender la edad crítica de adquisición. Asimismo, se han desarrollado otras tareas que parecen apoyar este hallazgo. Particularmente, los resultados del "paradigma de la esponja" (Flavell, Flavell & Green 1987) y del "paradigma del recipiente engañoso" (Perner, Leekman & Wimmer 1987) sugieren que los niños menores de 4 años no tienen aún una concepción de la capacidad representacional de la mente y, por esto, no son capaces de concebir que algo puede ser de un modo y aparecer de otro[7]. En conjunto, estos hallazgos han brindado apoyo, principalmente, a la versión del "niño científico" del enfoque de TT, de la que me ocuparé en detalle en el capítulo 1 (sección 3.2).

Una de las objeciones principales a la tarea de falsa creencia estándar es que se vale del lenguaje para testear la capacidad de psicología de sentido común. Si bien es cierto que los seres humanos atribuimos explícitamente deseos y creencias, y que predecimos manifiestamente conductas por medio del lenguaje, se ha señalado que las tareas basadas en el lenguaje pueden subestimar las capacidades mentalistas de los niños pequeños y de los primates no humanos[8]. Además de objetar las exigencias

[7] En el "paradigma de la esponja" se muestra al niño una esponja pintada de modo tal que simula ser una piedra y se le pregunta qué es. El niño debe contestar que "es una piedra". A continuación, se le pide que la toque y diga qué es. El niño debe contestar "una esponja". Finalmente, se le pregunta qué parece. Para desempeñar con éxito la tarea, el niño debe contestar que "parece una piedra". Sin embargo, los niños menores de 4 años contestan que "parece una esponja". En la tarea del recipiente engañoso o "tubo de *smarties*" se muestra al niño un recipiente tubular que, usualmente, contiene confites de chocolate y se le pregunta por el contenido del mismo. El niño debe responder "confites". Luego, se le muestra que se ha cambiado el contenido del tubo y que, en lugar de confites, hay un lápiz. A continuación, se hace entrar a otro niño y se le pregunta al primero qué dirá el niño nuevo cuando se le pregunte por el contenido del tubo. El niño debe responder "confites". Sin embargo, los niños menores de 4 años contestan erróneamente "lápiz".

[8] Recientemente, utilizando como herramienta una tarea de falsa creencia no-lingüística, se ha sugerido que hay una comprensión temprana de las creencias a una edad de apenas 15 meses (Onishi & Baillargeon 2005). En esta tarea de falsa creencia simplificada (no-lingüística), los niños de 15 meses

lingüísticas de la tarea de falsa creencia estándar, se han cuestionado sus exigencias computacionales. Esta prueba requiere que, además de interpretar las preguntas de los adultos y dar respuestas apropiadas a tales preguntas, los niños recuerden detalles de la historia. Éstas se consideran exigencias computacionales altas para los niños menores de 4 años. Al mismo tiempo, al parecer muchas de las demandas de esta tarea no están relacionadas con *mindreading* sino, más bien, revelan el desempeño en funciones ejecutivas, *e.g.* inhibición de la respuesta dominante. En suma, la tarea de falsa creencia es una tarea compleja y el desempeño exitoso en la misma no parece depender solamente de la capacidad de *mindreading* sino, además, parece implicar demandas lingüísticas, de memoria y de funciones ejecutivas.

1.2. Los estudios en neurociencias

Como hemos visto sucintamente, los estudios en psicología del desarrollo que emplean tareas de falsa creencia sugieren que, dada la complejidad de la tarea, los niños deben utilizar habilidades cognitivas generales tales como memoria de trabajo, representaciones perceptuales y lingüísticas sobre el relato, y representaciones motoras de sus propias respuestas para reconocer y razonar sobre estados mentales ajenos. En relación a esto, se ha planteado en el ámbito empírico la cuestión de si para explicar el fenómeno de la atribución mentalista basta con postular habilidades

---

miran a un participante esconder un juguete en una caja que está adyacente a otra, una caja amarilla o una caja verde. Luego, el juguete es movido a la otra caja. Esto se lleva a cabo en dos condiciones. En una condición, el participante puede ver el movimiento del juguete. Bajo la otra condición, el participante no puede ver el movimiento porque su visión está ocluida. Además, en una mitad de los ensayos de la tarea, el juguete es movido de una caja a la otra, y devuelto a la caja original. En la mitad restante de los ensayos, el juguete es movido, simplemente, a la otra caja. Usando el método de violación de las expectativas, se halló que los niños de 15 meses miran por más tiempo en dos casos. Primero, cuando el participante que no ha visto que el juguete ha sido cambiado de lugar busca, de todos modos, en la caja correcta (ubicación actual del juguete). Segundo, cuando el participante que ha visto que el juguete ha sido cambiado de lugar busca, sin embargo, en la caja incorrecta. Según Onishi & Baillargeon (2005), estos resultados sugieren que los niños de 15 meses esperan que el participante busque el juguete en base a la creencia que posee acerca de la ubicación del mismo. Cuando el participante no busca el juguete en base a sus creencias, las expectativas de los niños son violadas y, por esto, miran prolongadamente tales eventos.

cognitivas de dominio general o si bien es preciso postular un mecanismo específico dedicado a razonar sobre las creencias y los deseos.

Los defensores del enfoque modular, del que me ocuparé en el capítulo 1 (sección 3.3), han sugerido la posibilidad de que exista un mecanismo especializado en la atribución de estados mentales (Leslie 1991, 1994; Baron-Cohen 2005). Se supone que este mecanismo especializado de procesamiento de la información tiene un correlato a nivel cerebral. En línea con esto, se han buscado los correlatos neuronales de *mindreading* aplicando técnicas de neuroimagen a sujetos voluntarios normales, mientras estos realizan tareas que tienen como requerimiento común tener que pensar sobre los estados mentales de otras personas. Así, se ha utilizado una amplia gama de tareas tales como leer historias, observar figuras moviéndose, ver películas de dibujos animados o jugar juegos interactivos (Frith & Frith 1999, 2003).

En general, los estudios de correlatos neuronales se guían por los siguientes presupuestos. Se asume que el mecanismo neuronal que subyace a *mindreading* consiste en una red o circuito neuronal y no en una única región cerebral, puesto que se trata de una habilidad compleja. Se espera que la base neuronal presente variaciones en relación a la edad del agente que realiza las atribuciones e inferencias. Esto es, por un lado, deberían observarse diferencias en el mecanismo de los niños respecto del de los adultos. Por el otro, deberían observarse diferencias en el mecanismo si las atribuciones e inferencias se realizan respecto de uno mismo o en relación a otras personas. En principio, las observaciones provenientes de los estudios de neuroimagen son compatibles con la existencia de mecanismos especializados, ya que se ha observado un patrón específico de activación durante tareas que requieren atribuir estados mentales, tal como veremos en el capítulo 6. A principios de la primera década del 2000, en numerosos laboratorios de diferentes países, utilizando diferentes tareas, estímulos y escáneres, los científicos se preguntaron por las regiones cerebrales implicadas en *mindreading*. Llamativamente, todos estos estudios arribaron, de manera independiente, a la misma respuesta. El conjunto de áreas cerebrales asociadas a los procesos que ocurren durante la atribución de estados mentales resultó

ser la porción anterior derecha del surco temporal superior junto al área adyacente denominada juntura temporo-parietal (izquierda y derecha), los polos temporales, el precúneo medial, la corteza cingulada posterior y la corteza prefrontal medial (Fletcher *et al*. 1995b; Goel, Grafmanm, Sadato & Hallett 1995; Castelli *et al*. 2000; Gallagher *et al.* 2000; Saxe & Kanwisher 2003; Vogeley *et al.* 2001; German, Niehaus, Roarty, Giesbrecht & Miller 2004). Desde entonces, nunca volvió a haber tanto consenso (Saxe 2010). Los investigadores no se pusieron de acuerdo en la interpretación de estos resultados, ni respecto de cuáles regiones estaban específicamente involucradas en *mindreading*, ni sobre cuántas regiones formaban parte del conjunto. Particularmente, se ha cuestionado que la corteza prefrontal medial tenga algún rol en las atribuciones mentalistas, tal como sus porciones ventral y dorsal parecen tener un rol específico en otras capacidades de la cognición social, la empatía emocional y la atención conjunta respectivamente.

Para algunos, la información proveniente de los estudios de neuroimagen combinada con los estudios de neuropsicología (estudios de pacientes con daño cerebral) provee un panorama consistente sobre lo que algunos psicólogos llaman la "base cognitiva" de *mindreading* (Apperly 2010). *Mindreading* parece reclutar una red compleja de procesos funcionales y neuronales, entre los que se incluyen regiones neuronales que parecen ser altamente selectivas para la atribución mentalista. En concordancia con los presupuestos mencionados, en la comunidad científica existe consenso respecto de que, en sí misma, la capacidad cognitiva de atribución mentalista es la función de una red más que de una única región cerebral especializada y que entender la red cerebral, y las funciones que ésta sustenta, proveerá evidencia valiosa para entender la capacidad cognitiva de *mindreading* (Apperly 2010).

No obstante, en el ámbito de las neurociencias, se ha propuesto la hipótesis de que existe una región cerebral especializada en la representación de los pensamientos: la juntura temporo-parietal derecha. Según esta propuesta, esta región está especializada en pensar sobre las creencias y los deseos de otras personas pero no sobre otros estados internos ajenos como las sensaciones físicas, y no tiene ningún rol

en otras capacidades cognitivas (Saxe & Kanwisher 2003; Saxe & Wexler 2005; Saxe 2006). Esta hipótesis es interesante en tanto los estudios de neuroimagen interpretados de esta forma refuerzan y elaboran un enfoque aceptado ampliamente en relación al desarrollo de la capacidad mentalista. Según éste, el entendimiento de las otras mentes constituye un dominio especial de la cognición y está compuesto de, al menos, dos componentes: un sistema temprano para razonar sobre metas, percepciones y emociones, y un sistema tardío para razonar sobre creencias[9]. Este sistema especial para la atribución mentalista parece diferenciarse de los sistemas del lenguaje y del control inhibitorio. Así, el hallazgo de un área cerebral especializada en la atribución de creencias brinda apoyo a la hipótesis de la existencia de un sistema tardío especializado en creencias (Saxe & Kanwisher 2003; Saxe & Wexler 2005; Saxe 2006). Sin embargo, la hipótesis sobre la existencia de una única región especializada en la representación del pensamiento ajeno es aún controversial (Apperly 2011).

Por su parte, el apoyo empírico del enfoque de TS, en el ámbito neurofisiológico, se relaciona especialmente con el hallazgo de las neuronas espejo (NE). Las NE fueron descubiertas en la corteza premotora de los monos macacos (Gallese *et al.* 1996; Rizzolatti *et al*. 1996) y, posteriormente, neuronas con propiedades similares fueron descubiertas en un sector de la corteza parietal posterior (Rizzolatti, Fogassi & Gallese 2002; Gallese *et al.* 2002). El hallazgo en relación a las NE consiste en la observación de que éstas se activan característicamente cuando el mono lleva a cabo una acción motora (un movimiento dirigido a una meta) del tipo de una interacción mano-objeto (agarrar, sostener, romper, manipular) y llamativamente, también, cuando el mono observa a un experimentador, o a otro mono, llevar a cabo

[9] Distintos enfoques asumen esta trayectoria en el desarrollo que va de la posesión de una "psicología de deseos" a la posesión de una "psicología de deseos y creencias" (Bartsch & Wellman 1995). Así, se observa tempranamente la capacidad de atribuir estados mentales de tipo deseos, percepciones y emociones, y posteriormente se agrega la capacidad de atribuir creencias, que implica poseer conceptos de estados mentales (Wellman 1990; Bartsch & Wellman 1995; Baron-Cohen 1995, 1997; Gopnik & Meltzoff 1997; Leslie 1995, 2000; Tager-Flusberg & Sullivan 2000). Me referiré a esto con más detalle en el capítulo 1 (sección 3).

acciones similares[10]. Este patrón de activación es tan selectivo que las NE no se activan ante la sola visión del objeto, ni del agente. A su vez, su actividad tampoco es provocada por otras acciones como la imitación de la acción sin el objeto o la ejecución de la misma acción pero con herramientas[11].

Numerosos estudios, principalmente de neuroimagen, han mostrado la presencia de un sistema similar en los seres humanos. Las áreas premotoras y parietales humanas más homólogas a las áreas donde se hallaron NE en macacos, han mostrado una fuerte activación durante la observación de acciones (Rizzolati, Fogassi & Gallese 2001; Rizzolatti & Craighero 2004). De este modo, se asume que este sistema empareja la ejecución y la observación de las acciones y, usualmente, se denomina "sistema espejo" (Fadiga *et al*. 1995; Rizzolatti *et al*. 1996; Grafton *et al*. 1996). Además, en los humanos, el sistema espejo parece estar organizado somatotópicamente, es decir, con regiones corticales premotoras y parietales distinguibles según las acciones estén relacionadas con la boca, las manos o los pies (Buccino *et al*. 2001).

En base a estos hallazgos, Gallese & Goldman (1998) proponen la hipótesis de las NE como una parte componente, al menos, rudimentaria, quizás al estilo de un precursor, de la capacidad de *mindreading*. El punto de partida de esta propuesta es la tesis clave de la TS, según la cual, dada la similitud interna entre el sujeto y el blanco, el proceso que subyace a la atribución de estados mentales consiste en utilizar los mecanismos mentales propios como modelo del otro. De este modo, "simular" puede

[10] Este hallazgo resultó llamativo porque la corteza premotora ventral, donde se hallaron las NE, era conocida por su rol en la planificación de la acción pero no se le conocía rol visual alguno. Estas neuronas parecen constituir un sistema de emparejamiento entre la ejecución y la observación de la acción, conocido como "sistema de resonancia". En el sentido de que cierta actividad neuronal en el observador "resuena" con la actividad observada en el agente. Se asume que las NE constituyen el sustrato de la representación de un plan motor para una acción que, cuando se activa en el observador, no se ejecuta.

[11] Además de cumplir con estas características, las NE del área F5 de la corteza premotora de los macacos muestran otros rasgos particulares en su actividad. Se ha observado que estas neuronas también se activan cuando la parte final y crítica de una acción observada está ocluida para la visión (Umilitá *et al*. 2001). Algunas de estas NE son audiovisuales, esto es, no sólo se activan ante la ejecución/observación de acciones motoras sino que, también, responden ante el sonido producido por la acción (Kohler *et al.* 2002). Asimismo, se ha observado que las NE se activan ante la ejecución/observación de acciones llevadas a cabo con la boca como agarrar, morder o lamer (Ferrari *et al*. 2003).

entenderse como la capacidad de replicar la vida mental del blanco. En este sentido, tal como se verá en el capítulo 2, el enfoque simulacional postula esencialmente una correspondencia entre la actividad mental del sujeto y la del blanco. Una correspondencia similar, aunque en el aspecto cerebral, parece observarse en el patrón de activación de las NE, dado que éstas se activan cuando un agente lleva a cabo una acción motora y, también, cuando observa a otro individuo llevar a cabo una acción similar. La propuesta de que las NE brindan apoyo empírico a la TS es controversial. Me referiré con más detalle a esta controversia en el capítulo 5 (sección 1.1).

1.3. Hacia los enfoques híbridos

Recapitulando, los esfuerzos para tratar de elegir entre TT y TS sopesando argumentos conceptuales y evidencia empírica, tal como he mencionado, no han sido fructíferos y, por lo tanto, hasta el momento no existe un modo claro de elegir entre ambos enfoques. Una primera dificultad, también ya mencionada, reside en que hay distintas versiones de cada enfoque. En segundo lugar, ambos enfoques (o, más bien, todas las variantes de los mismos) tienen ciertas virtudes y desventajas conceptuales. En tercer lugar, no ha sido posible determinar empíricamente qué teoría es más adecuada. En este sentido, ambos enfoques pueden dar cuenta del hallazgo en el desarrollo relacionado con el desempeño infantil en tareas de falsa creencia, encuentran apoyo en la evidencia proveniente de las neurociencias y tienen casos a su favor.

Esta situación ha llevado a algunos teóricos o bien al rechazo de ambos tipos de teorías y, en especial, al rechazo de la psicología de actitudes proposicionales a la que ambos enfoques adhieren, o bien a intentar articular ambas teorías en enfoques híbridos. En el ámbito de la filosofía, desde una perspectiva fenomenológica se considera que la interpretación de la conducta no está mediada por conceptos mentales sino, más bien, por una interpretación no-mentalista, basada en

percepciones de las intenciones y las disposiciones de las otras personas (Gallagher 2001, 2006; Zahavi 2008). La interpretación aludida se puede concebir como una "lectura del cuerpo", según la cual con sólo presenciar las acciones y los movimientos de una persona basta para acceder a su significado, sin que sea necesaria inferencia alguna a estados mentales; en esto consiste la intersubjetividad primaria (Trevarthen 1979). Desde una perspectiva de cognición corporizada, Daniel Hutto (2008) propone la hipótesis de la práctica narrativa. Según ésta, la fuente de nuestra capacidad mentalista sofisticada son los encuentros directos con narrativas de psicología de sentido común. Esta práctica narrativa no depende de ni consiste en una psicología de actitudes proposicionales, sino en las razones que los otros tienen para sus acciones. Así, el desarrollo de la competencia mentalista consiste en aprender sobre la forma y las normas para las explicaciones basadas en razones: las narrativas de psicología de sentido común (Gallagher & Hutto 2007). En el ámbito empírico también se ha sugerido que los desarrollos de la TT y la TS quizás no provean el mejor marco teórico para generar predicciones e interpretar los datos de las investigaciones neurocientíficas. Se ha propuesto la necesidad de abandonar ambos enfoques y sus conceptos para utilizar, en su lugar, las herramientas conceptuales del propio campo de la neurociencia cognitiva social (Apperly 2008).

Una segunda actitud respecto a esta situación, y la que analizaré en la presente investigación, es tratar de aprovechar las ventajas de TT y TS, y articularlas en un enfoque híbrido con elementos de teoría y simulación. Los enfoques híbridos postulan la intervención de procesos de simulación y, a la vez, de ciertas bases de información para realizar las atribuciones. De este modo, simulación y teoría no entran en competencia. Las propuestas híbridas varían y un modo de distinguirlas es en función del proceso en que se pone énfasis. En este sentido, se puede establecer una clasificación de enfoques híbridos según se asuma que el proceso subyacente a la atribución mentalista es por *default* la capacidad de realizar inferencias a partir de un cuerpo de información (teoría) o bien, la capacidad de proyectarse imaginativamente en la situación del otro (simulación).

2. La cuestión de los enfoques híbridos

En la literatura sobre el tema, se han propuesto fenómenos que toda teoría de *mindreading* debe poder explicar, entre los cuales figuran poder dar cuenta de: (a) cómo se atribuyen estados mentales a otros, (b) cómo se atribuyen estados mentales a uno mismo, (c) cómo se atribuye agencia a otros y a uno mismo, (d) cómo se predicen inferencias de otras personas exitosamente aunque con ciertas limitaciones, (e) el significado de los conceptos mentales, (f) cuál es el patrón de desarrollo o adquisición de esta capacidad, (g) ciertos déficits en la habilidad de mentalizar, (h) los contenidos de los estados mentales, (i) el lugar que ocupa *mindreading* en la historia evolutiva de la cognición humana, (j) su relación con otras formas de cognición social (*e.g.* empatía, compasión), (k) cómo está relacionada con la arquitectura de otros dominios cognitivos (Nichols & Stich 2003; Goldman 2006).

La idea a la base de las propuestas híbridas es que un enfoque híbrido presenta ventajas respecto de las posturas teóricas que adhieren sólo a las tesis puras de la teoría o de la simulación, puesto que heredan las ventajas explicativas de estas teorías. Sin embargo, se debe mostrar que no heredan sus dificultades. En este sentido, a la par de evaluar los enfoques híbridos mencionados respecto de si ofrecen un criterio satisfactorio para distinguir entre procesos subyacentes, basándome en los fenómenos explícitamente mencionados, (a)-(k), propongo que, mínimamente, un enfoque híbrido de *mindreading* debe, además de (i) proponer un criterio para distinguir entre los procesos subyacentes, poder dar cuenta, al menos, de: (ii) cómo se atribuyen estados mentales a otros (el requisito (a) anterior), (iii) cómo se atribuyen estados mentales a uno mismo (el requisito (b) anterior). Básicamente, porque las atribuciones mentalistas de primera y tercera persona son las dos caras de una misma capacidad, la atribución de estados mentales a los seres con mente. Así, las propuestas híbridas de Nichols & Stich (2003) y Goldman (2006) serán evaluadas en relación a los requisitos de mínima (i)-(iii) propuestos. En lo que sigue caracterizaré a grandes rasgos, las propuestas

híbridas de atribución mentalista que evaluaré en función de los requisitos de mínima en los capítulos 4 y 5 de la tesis.

Nichols & Stich (2003) proponen un enfoque híbrido de *mindreading* de tercera persona, que postula la existencia de dos sistemas de *mindreading*, filogenética y ontogenéticamente diferenciados, el temprano y el tardío. El sistema temprano permite predecir el comportamiento mediante la atribución de metas y la definición de la mejor estrategia para alcanzarlas. El sistema tardío permite realizar atribuciones mentalistas a partir de un modelo del agente blanco, que incluye creencias discrepantes respecto de las del propio agente[12]. Específicamente, el sistema tardío permite construir un modelo (de las creencias) del blanco a partir de las creencias del propio agente, de modo que el sistema inferencial para *mindreading* se vuelve sensible a la posibilidad de que existan creencias, en el blanco, que sean incompatibles con las creencias del agente. En este sentido, aumenta la capacidad predictiva en relación con la del sistema temprano, que no contempla este aspecto.

Según los autores, la cuestión fundamental en relación a *mindreading* reside en tratar de determinar cuáles son los procesos cognitivos que subyacen a estos dos sistemas. En principio, el sistema temprano estaría constituido por tres tipos de sistemas: los mecanismos detectores de deseos, el mecanismo planeador y el mecanismo coordinador. El sistema tardío estaría compuesto por dos sistemas: el sistema temprano y la caja de mundos posibles (CMP). En este enfoque híbrido, se considera que, *por default*, una teoría subyace a la atribución mentalista. Sin embargo, ciertos procesos de simulación forman parte de los componentes de los sistemas mencionados. Específicamente, los procesos de simulación tienen lugar en el mecanismo de planeamiento, que forma parte del sistema temprano, y en el mecanismo de CMP, que forma parte del sistema tardío. Ambos mecanismos intervienen en la predicción inferencial. El planeador se encarga de las predicciones

[12] En principio, la afirmación de la existencia de dos mecanismos ontogenéticamente diferenciados no es controversial. En el ámbito de la psicología del desarrollo, existe un acuerdo amplio acerca de la existencia de, al menos, dos sistemas para *mindreading* (Wellman 1990; Baron-cohen 1995; Leslie 1995; Gopnik & Meltzoff 1997).

sobre cómo el blanco de la atribución alcanzará sus metas, mientras que la CMP funciona como un generador de creencias alimentado por creencias ficticias.

Es preciso mencionar que a la base de esta propuesta de *mindreading*, se asume, primero, una arquitectura cognitiva básica que contiene dos tipos diferentes de estados representacionales: creencias y deseos. Estos difieren funcionalmente en tanto son causados de diferentes maneras y tienen diferentes patrones de interacción con los otros componentes de la mente. Segundo, se asume un enfoque representacional de la cognición según el cual tener una creencia o un deseo con un contenido particular es tener una instancia de representación con un contenido almacenado funcionalmente de modo apropiado en la mente, esto es, almacenado en la "caja de creencias" o en la "caja de deseos". En línea con esto, el proceso subyacente a la autoatribución consiste en un proceso o mecanismo de monitoreo. Puesto que las creencias del organismo están alojadas en la "caja de creencias", el mecanismo de monitoreo se encarga de copiar una representación de la "caja de creencias" y agregarle el prefijo de la adscripción de la actitud "yo creo que" y la representación de segundo orden resultante se coloca en la "caja de creencias". Esta propuesta asume el acceso especial a los propios estados mentales, y la existencia de la introspección entendida como un mecanismo de monitoreo de los estados mentales propios. En principio este enfoque se distancia de la postura que, usualmente, defienden los psicólogos del desarrollo, según la cual se utilizan los mismos recursos para atribuir estados mentales a uno mismo y a los otros, a saber, la teoría psicológica de sentido común y cierta evidencia. En este sentido, el autoconocimiento resulta interpretativo y no directo (Gopnik 1993). Me ocuparé con más detalle de estas cuestiones en el capítulo 3 (sección 3.2).

Particularmente, se ha objetado a la propuesta de introspección de Nichols & Stich (2003) que si bien resulta plausible para dar cuenta de las creencias, no parece ser adecuada para otro tipo de actitudes proposicionales. Esta propuesta requiere que el mecanismo de monitoreo recupere las instancias de actitudes proposicionales de "cajas" diferentes para cada actitud y, en consecuencia, esto genera la predicción de

que cada caja actitudinal puede estar dañada independientemente del resto, de modo que son esperables múltiples perfiles de daño disociados. Sin embargo, la evidencia no provee evidencia de tales disociaciones (Engelbert & Carruthers 2010).

En mi opinión, esta propuesta no es satisfactoria respecto de su carácter híbrido, en tanto no proporciona un criterio satisfactorio para distinguir entre teoría y simulación que permita establecer si a una instancia de *mindreading* subyace un proceso de teoría o de simulación. Como he señalado, postular un enfoque híbrido de teoría y simulación implica brindar algún criterio para determinar el tipo de proceso subyacente. En este sentido, Stich & Nichols (2003) han ofrecido el siguiente argumento de corrección: si *mindreading* resulta correcto, es probable que el proceso subyacente sea de tipo simulacional, y si *mindreading* resulta incorrecto es probable que el proceso subyacente sea de bases de información. Discutiré este argumento propuesto dividiéndolo en dos aspectos. En relación al primer aspecto, *i.e.* si somos buenos en *mindreading* es probable que el proceso subyacente sea de tipo simulacional, sostendré que no resulta satisfactorio. Considero que los argumentos sobre la predicción de inferencias, en que se apoya, no permiten descartar explicaciones de procesos ricos en información subyacentes a casos de *mindreading* exitoso. Para que la explicación que propone la TS sobre la corrección de *mindreading* sea preferible, es condición necesaria que, a su vez, se descarte la explicación de la corrección de *mindreading* basada en procesos ricos en información. Según ésta, *mindreading* correcto resulta de la posesión de una buena teoría. En contra del segundo aspecto, *i.e.* si sistemáticamente nos equivocamos en *mindreading* es probable que el proceso subyacente sea de tipo rico en información, sostendré que es posible que un proceso de tipo simulacional subyazca a *mindreading* incorrecto. Considero que si el sistema de toma de decisiones funciona de manera no estándar en la rutina de simulación (*i.e.* alimentado por *inputs* ficticios y desacoplado de los sistema controladores de la acción) puede ser el caso que este modo de funcionamiento no reclute, para la simulación, los sesgos que usualmente operan sobre el sistema de

toma de decisiones cuando funciona normalmente[13]. Si los sesgos no se reclutan en la simulación, el *output* de *mindreading* (la predicción) no coincidirá con el comportamiento del blanco originado en una decisión afectada por los sesgos. Finalmente, concluiré que el criterio de corrección propuesto no permite distinguir entre tipos de procesos subyacentes.

Por su parte, Goldman (2006) propone un enfoque híbrido con énfasis en la simulación para la atribución de tercera persona, que asigna un rol a la teoría en tanto se asume que las rutinas de *mindreading* pueden combinar elementos de teoría y de simulación. Por ejemplo, la teoría puede ayudar a la simulación a seleccionar los *inputs* que ponen en marcha la simulación y así, establecer los parámetros de la misma. A mi entender, el aspecto más problemático de este enfoque se relaciona con su carácter híbrido. En principio, el papel explicativo de la teoría no queda muy claro en la medida en que este enfoque apenas enumera las formas en que se pueden relacionar estos elementos en un enfoque híbrido (independencia y cooperación) y, a mi entender, no proporciona criterio alguno para distinguir entre teoría y simulación.

En esta propuesta, tal como en la propuesta pura de simulación (Goldman 1995a, 1995b), la autoadscripción de estados mentales también tiene una relevancia particular dado que la atribución de tercera persona implica la atribución de primera persona. La relación entre estas atribuciones reside en la necesidad de clasificar los estados mentales propios para realizar atribuciones mentalistas de tercera persona en tanto que la simulación mental implica generar estados mentales en la mente del simulador con el propósito de coincidir con el estado mental del blanco. Según Goldman, para que esto sea posible es preciso clasificar el estado mental según su tipo (*e.g.* sensación, emoción o actitud proposicional) y establecer su contenido, si es que lo tiene. De este modo, el simulador no puede realizar una atribución a otro agente a menos que haya podido resolver estas cuestiones clasificatorias en relación a sus estados mentales. Así, una tarea de *mindreading* de primera persona está siempre implicada como una sub-tarea de *mindreading* de tercera persona.

[13] Por "ficticio" entiendo lo que en inglés es "*pretend*".

La propuesta de autoatribución postula dos procesos subyacentes, introspección e interpretación. Por un lado, un mecanismo de monitoreo que permite identificar y clasificar los estados mentales propios. Este es complementado por un mecanismo de reorientación, que permite atribuir contenidos a los estados, si es que lo tienen. En principio, falta detalle en la descripción de la propuesta en la medida en que no queda claro cómo los resultados de los dos procesos componentes postulados, monitoreo y reorientación, se integran para que los *outputs* respectivos den lugar a la autoadscripción de estados mentales. Asimismo, esta propuesta hereda las dificultades asociadas a la adopción de un enfoque introspeccionista. En particular, se adhiere a una concepción cuasi-perceptual de la aprehensión de estados mentales. En principio, esto es problemático en tanto que la introspección carece de un órgano receptor y esto dificulta establecer una analogía entre la percepción y la introspección. Por otro lado, en algunas ocasiones, en particular en los casos de confabulación, las autoatribuciones tienen lugar mediante un proceso de interpretación. Sin embargo, este aspecto de *mindreading* de primera persona no está desarrollado en la propuesta.

A su vez, este enfoque propone dos tipos de procesos de simulación subyacentes a *mindreading* de tercera persona. Por un lado, una serie de sistemas de simulación de nivel inferior, primitivos, automáticos, e inconscientes se ocupan del reconocimiento de las acciones, ciertas emociones y ciertas sensaciones mediando, así, la atribución de intenciones motoras, emociones y sensaciones. Por otro lado, un sistema simulacional de nivel superior, la “imaginación enactiva”, que en ocasiones puede ser un proceso voluntario o controlado y consciente, subyace a la atribución de estados mentales complejos tales como las actitudes proposicionales.

En mi opinión, la propuesta de niveles de simulación presenta varias dificultades. Básicamente, el apoyo empírico es cuestionable. Respecto a la simulación de nivel inferior, si bien los procesos espejo pueden entenderse como simulaciones, resulta controversial afirmar que los mismos subyacen a *mindreading*. La evidencia sólo muestra correlación y no permite afirmar que los mecanismos espejo causan *mindreading*, esto es sólo una inferencia. En particular, parece existir una brecha entre

la cognición primitiva que suponen los procesos espejo y la cognición superior que suponen la atribuciones explícitas. Por su parte, la propuesta de simulación de nivel superior, cuyo prototipo es la "imaginación enactiva", está basada en evidencia que no proporciona apoyo directo. Por ejemplo, si bien cierta evidencia proveniente del estudio de la imaginación visual y de la imaginación motora sugiere la generación de imágenes visuales mentales y de representaciones de la ejecución de un movimiento que guardan similitud con sus contrapartes (la percepción visual y los planes motores), hasta el momento no hay evidencia sobre la simulación de creencias en la imaginación. A su vez, se ha cuestionado la distinción conceptual entre los niveles superior e inferior de simulación. No parece haber una distinción esencial entre los mismos sino, más bien, una distinción de grado (de Vignemont 2009; Goldman 2009).

Como ya he mencionado, en mi opinión, es esencial que los enfoques híbridos de teoría y simulación puedan brindar un criterio que permita distinguir entre procesos subyacentes a instancias de *mindreading*, de modo que se pueda distinguir si interviene un proceso de teoría o de simulación. Retomando evidencia proveniente del campo de las neurociencias, en el capítulo 6, intentaré echar luz acera de la importancia de que los enfoques híbridos brinden un criterio para distinguir entre procesos subyacentes ya que, a mi entender, esto tiene impacto en el estudio empírico de *mindreading*. Particularmente, considero que es crítico para el estudio de *mindreading* con métodos de neuroimagen. A su vez, a partir de la evidencia reseñada intentaré proveer un criterio para los enfoques híbridos de teoría y simulación, aunque esto probablemente implicará algunas concesiones en relación al alcance explicativo de cada enfoque. A mi entender, la evidencia proveniente de las neurociencias sugiere un criterio que implica una división de tareas entre teoría y simulación.

Tal como intentaré mostrar en el capítulo 6, los estudios de *mindreading* en el ámbito de las neurociencias se pueden agrupar en tres grandes cuestiones. Un grupo mayoritario de estudios se dedica a buscar la base neuronal de *mindreading* de otras personas (Frith & Frith 1999, 2003, 2006; Saxe & Kanwisher 2003, Saxe & Wexler 2005). Estos estudios se complementan, a su vez, con los estudios sobre la base

neuronal de las autoatribuciones o del yo, lo que en inglés se entiende por "*self*" (Saxe *et al.* 2006). Un segundo grupo de estudios se dedican a evaluar predicciones que se asume se desprenden de los enfoques de teoría y simulación (Apperly 2008). Se trata de algunos pocos estudios con técnicas de neuroimagen que han puesto a prueba estas predicciones para tratar de determinar el tipo de proceso subyacente a *mindreading*, aunque con resultados dispares (Vogeley *et al.* 2001, Ramnani & Miall 2004, Grezes, Frith & Passingham 2004). Del análisis particular de estos estudios y los supuestos que asumen concluiré que resulta crítico que los enfoques híbridos de teoría y simulación puedan proveer un criterio para distinguir entre procesos subyacentes.

El tercer grupo que he detectado, brinda propuestas híbridas de teoría-simulación que surgen en el ámbito mismo de las neurociencias (Tager-Flusberg & Sullivan 2000; Pineda & Hetch 2008; Keysers & Gazzola 2006; 2007). Estas teorías pretenden conciliar los hallazgos de correlatos neuronales recurriendo a lo que considero cierta división de tareas en relación al procesamiento de los estados mentales. Según la evidencia, los procesos espejo intervendrían en el procesamiento de estados mentales tales como la intención motora, ciertas sensaciones (tacto y dolor) y ciertas emociones (miedo, asco), mientras que las regiones frontales y temporo-parietales serían reclutadas en el procesamiento de estados mentales más complejos como las creencias. En principio, estas propuestas de enfoques híbridos no resultan satisfactorias para dar cuenta de *mindreading* en tanto no dan cuenta de la autoatribución. No obstante, a mi entender, sugieren un criterio para determinar el tipo de proceso subyacente a *mindreading* de otras personas.

Finalmente, presentaré las Conclusiones retomando aquellos aspectos cruciales señalados en cada capítulo que me llevaron a proponer que resulta esencial que un enfoque híbrido brinde un criterio para distinguir entre procesos de teoría o simulación subyacentes a *mindreading*, así como señalaré el criterio, que a mi entender, se desprende de los estudios empíricos de *mindreading* en el ámbito de las neurociencias. A su vez, quedarán expuestas las dificultades que presentan los enfoques híbridos de teoría y simulación analizados para dar cuenta de *mindreading*, algunas de las cuales se

heredan de las propuestas puras mientras que otras surgen propiamente de las propuestas híbridas.

CAPÍTULO 1. EL ENFOQUE DE LA "TEORÍA DE LA TEORÍA"

Abordaré las cuestiones centrales en relación al debate teoría-simulación en este capítulo y en el siguiente. Una vez presentadas las propuestas de TT y de la simulación mental, estaré en condiciones de abordar la cuestión de los modelos híbridos. Esto tendrá lugar en el capítulo 3. Allí mostraré las razones en las que se basa el vuelco reciente, más o menos masivo, hacia posiciones híbridas de teoría-simulación en el estudio de la competencia mentalista. Una de las motivaciones se relaciona con cierta necesidad del complemento de simulación por parte de la TT y de teoría por parte de la TS, que conduce a los enfoques mixtos de teoría y simulación. Otra motivación reside en que ambos enfoques parecen dar cuenta convincentemente de ciertos casos de *mindreading*.

En este capítulo me ocuparé de analizar el enfoque de la "teoría de la teoría" que sostiene básicamente que poseemos una teoría psicológica rudimentaria que utilizamos cotidianamente tanto en la atribución de estados mentales, como en la explicación y la predicción de la acción y el pensamiento en términos mentalistas (*mindreading*). En la sección 1, señalaré algunos presupuestos del debate teoría-simulación. En la sección 2, abordaré la relación entre la "teoría de la teoría", el funcionalismo y los conceptos mentales. Asimismo, presentaré la primera versión de la TT proveniente del funcionalismo filosófico. En la sección 3, presentaré las diversas versiones de la "teoría de la teoría": la TT genérica o inclusiva, proveniente de la filosofía, y dos enfoques de la psicología cognitiva: el enfoque del niño científico y el nativista/modular. En la sección 4, me ocuparé de las distintas analogías a las que han recurrido los "teóricos de la teoría" para dar sentido a la afirmación de que *mindreading* depende de la posesión de un cuerpo de conocimiento. Finalmente, en la sección 5, presentaré las conclusiones en términos de las diferencias entre las cuatro versiones de la TT que pueden ordenarse alrededor de tres ejes. Primero, según exista un compromiso estrecho o no con la presencia de generalizaciones legaliformes. Segundo, según que el modo de adquisición de la teoría sea aprendido o innato.

Tercero, según que los recursos cognitivos que la teoría utilice sean generales o específicos. Además, señalaré las tesis básicas de la TT que luego serán retomadas para el análisis de los modelos híbridos.

1. El debate teoría-simulación

Antes de abordar el enfoque de la "teoría de la teoría" voy a señalar algunos presupuestos en torno al debate teoría-simulación. Por un lado, se suele asumir que la cuestión de *mindreading* es una sola. Sin embargo, esta cuestión abarca preguntas diferentes que, a su vez, pueden contestarse, desde un enfoque de "teoría de la teoría" o de "teoría de la simulación"[14]. Algunas de estas preguntas son empíricas, otras conceptuales. Por esto, una posible respuesta planteada a alguna de las cuestiones no puede meramente extenderse a las demás. No obstante, el debate ha discurrido mayormente en relación a la pregunta empírica sobre cuáles son los recursos o procesos cognitivos que subyacen a *mindreading* en los seres humanos adultos normales (Davies & Stone 1995a)[15]. Particularmente, se ha discutido sobre los recursos subyacentes a la predicción y la explicación de las acciones. No obstante, tal como señalan Davies & Stone (1998), toda respuesta en relación a los recursos implica supuestos respecto de otras cuestiones, tales como la naturaleza de los conceptos mentales y la de cuáles son las características clave de *mindreading*.

Por otro lado, se suele asumir que los enfoques de TT y TS son los únicos enfoques viables para dar cuenta de todas las preguntas posibles sobre *mindreading*. Sin embargo, esto no es así. Los enfoques híbridos de teoría y simulación, de los que me ocuparé en los capítulos 4 y 5, ya pueden considerarse como una tercera posición. Asimismo, se han formulado enfoques alternativos por fuera del marco del debate teoría-simulación. Como mencioné en la introducción, estos no serán evaluados en

[14] Stone & Davies (1996: 119-120) han identificado, al menos, 9 preguntas relacionadas con el debate acerca de *mindreading*. Una lista de preguntas empíricas acerca de TT puede encontrarse en Ravenscroft (2010) y una lista de preguntas filosóficas en Nichols (en prensa).

[15] Por "seres humanos adultos normales" me refiero a la población de los seres humanos adultos, enculturados, no autistas y sin daño cerebral.

esta tesis porque no se ocupan de los procesos cognitivos que subyacen a *mindreading* sino, más bien, proponen dar cuenta del modo en que nos entendemos a nosotros mismos y a las otras personas apelando a la racionalidad, la "intersubjetividad", o el entendimiento en términos de razones. Finalmente, es preciso señalar que no resulta una cuestión simple sistematizar este debate. Una primera dificultad se relaciona con el uso de los términos "teoría" y "simulación". Estos han sido utilizados con diferentes sentidos por diferentes autores. A su vez, dado que no hay una sola TT, ni una TS, sino distintas versiones de las mismas, los términos se han usado para caracterizar un amplio rango de enfoques teóricos. A continuación comenzaré con el análisis de la TT. En la sección siguiente, me ocuparé del funcionalismo y los conceptos mentales con el propósito de aclarar la relación entre el enfoque de TT y la cuestión de la naturaleza de los términos mentales.

2. La "Teoría de la Teoría", el funcionalismo y los conceptos mentales

Las personas suelen tener una actitud mentalista. Se representan a sí mismos y a los demás como teniendo estados mentales. En este sentido, muchos consideran que cotidianamente utilizamos una teoría psicológica ordinaria para conducirnos en el mundo social. Ésta nos permite calcular qué harán las personas en determinadas circunstancias, basados en lo que conocemos sobre sus preferencias y sobre la información de que disponen. A su vez, aplicamos esta teoría para determinar las creencias y los deseos de las personas a partir de sus comportamientos, así como para predecir el comportamiento de las personas a partir de sus deseos y creencias. Asimismo, utilizamos la teoría para calcular lo que otros piensan de manera de coordinar nuestro comportamiento con el de los otros.

Según la TT, la teoría psicológica ordinaria que subyace a la habilidad de *mindreading* es una psicología de actitudes proposicionales. Ésta nos permite entendernos a nosotros mismos y a los demás adscribiendo estados mentales tales como creencias y deseos. Particularmente, la teoría psicológica ordinaria se constituye

como un conjunto de generalizaciones aplicables a los otros así como a uno mismo. En base a este cuerpo de generalizaciones, y a partir de premisas sobre lo que el otro cree o conoce, es posible obtener conclusiones sobre las acciones de los individuos, bajo la forma de predicciones y explicaciones. No obstante, la TT afirma algo más. La teoría psicológica ordinaria no sólo permite *mindreading* (la atribución, la explicación y la predicción mentalistas) sino que otorga significado a los conceptos de estados mentales, que son introducidos y definidos por la misma.

En este respecto, la TT se presenta asociada al funcionalismo, que constituye una respuesta posible a la cuestión de la naturaleza de los términos mentales. La primera versión de TT fue desarrollada por filósofos interesados en el análisis de los conceptos de estados mentales (Sellars 1956; Lewis 1972). Estos filósofos extendieron una idea heredada del empirismo en filosofía de la ciencia al ámbito de los conceptos mentales. A saber, los conceptos teóricos, no observables, son definibles por medio de sus relaciones legaliformes con los observables. En consecuencia, el repertorio conceptual utilizado en la atribución mentalista, que se caracteriza como inobservable, puede concebirse como articulado en una teoría, en este caso, de sentido común.

Particularmente, Sellars (1956) propone que, al adscribir estados mentales a otros para explicar y predecir su comportamiento, se postulan entidades teóricas para explicar y predecir fenómenos observables. Según este enfoque, el conocimiento de sentido común sobre la mente es una "teoría" psicológica, y esta teoría postula estados mentales, como las creencias, que no son observables. De este modo, esta propuesta permite hacer referencia a los estados mentales internos sin necesidad de recurrir a la introspección. Esto es importante porque, en el siglo XX, desde los enfoques del conductismo psicológico y lógico se cuestiona la idea cartesiana de la introspección como una fuente confiable de conocimiento (científico) sobre la mente. En las *Meditaciones Metafísicas* se puede dudar de todo excepto de que se está dudando, esto es, de que se está pensando (Descartes 1641). No se puede dudar de lo que pasa en la propia mente y así, el acceso a nuestros estados mentales es indubitable. En el ámbito de la psicología, el cuestionamiento de la introspección como fuente confiable

motivó el rechazo al uso de términos mentales, este enfoque es conocido como conductismo metodológico (Watson 1913), mientras que, en el ámbito de la filosofía, el conductismo lógico (Ryle 1949) aceptó la utilización de términos mentales con la salvedad de considerar que estos no refieren a entidades internas sino a fenómenos observables, en particular, a disposiciones para actuar de cierta manera en determinadas circunstancias.

En este sentido, la propuesta de Sellars (1956) se constituye como una alternativa al cartesianismo, que considera a la introspección como una fuente confiable de conocimiento mental, y, al mismo tiempo, se constituye como una alternativa al conductismo en tanto postula estados mentales internos para dar cuenta de la conducta[16]. Esta propuesta fue presentada bajo la forma del mito de "nuestro ancestro Jones". Según este relato, hubo un tiempo en el que entre nuestros ancestros *ryleanos* no se utilizaban términos mentales para entender a los demás, ni para dar cuenta de las propias acciones. Sin embargo, cierto día Jones advirtió que el comportamiento verbal explícito es la culminación de un proceso que empieza con el habla interna, y creó una teoría según la cual las palabras manifiestas no son más que sucesos que comienzan con eventos internos (Sellars 1956: 56).

Posteriormente, en el marco de la discusión sobre cómo definir los términos teóricos, Lewis (1970; 1972) retoma este desarrollo y propone un método para definir funcionalmente los conceptos mentales. Aquí, la teoría psicológica ordinaria se concibe como una teoría, aunque inventada antes de que exista la ciencia, que introduce términos teóricos mentales del mismo modo en que las teorías científicas introducen términos teóricos. Esta teoría recoge las nociones de sentido común sobre cómo los estados mentales están relacionados causalmente con los estímulos sensoriales, las respuestas motoras y otros estados mentales. Así, los términos mentales cotidianos pueden definirse por el rol funcional/causal específico que poseen (Lewis 1972). Dado

[16] La idea de que la psicología de sentido común es una teoría se conoce como el enfoque de la "teoría-teoría" (Morton 1980). La TT no sólo proporciona una nueva manera de construir la psicología de sentido común una vez que la introspección ha sido desplazada, sino que también implica un nuevo modo de concebir la introspección misma. El enfoque de Sellars sugiere que la "teoría psicológica" que utilizamos para atribuir estados mentales a otros también se usa en la auto-atribución.

que las definiciones funcionales recogen el significado de nuestros conceptos mentales ordinarios, la "teoría" en que se basan nuestras explicaciones y predicciones mentalistas (*mindreading*) se identifica con el conjunto de perogrulladas sobre lo mental.

Según las propuestas mencionadas, la teoría otorga significado a los conceptos de estados mentales estableciendo las relaciones legaliformes con la conducta y los estímulos sensoriales, tal como estos se conciben desde el sentido común. Dado que las relaciones legaliformes establecen conexiones entre los *inputs* sensoriales, los *outputs* conductuales y otros estados internos, la teoría se entiende como funcionalista. Más específicamente, los funcionalistas conciben los estados y eventos mentales como los mediadores causales entre los *inputs* sensoriales del sujeto y sus *outputs* conductuales. De este modo, tanto el significado de los términos mentales como lo que hace que un estado mental sea de determinado tipo, proviene de sus relaciones funcionales con los estados perceptuales, las respuestas conductuales y los otros estados mentales del sujeto (Lycan 1994). Usualmente, se ha concebido a esta teoría como una colección de perogrulladas sobre la mente, o más bien, como generalizaciones que sistematizan tales perogrulladas. Por "perogrulladas" se entienden todos aquellos principios que todas las personas estarían dispuestas a aceptar, que pueden conocer fácilmente y que les parecen obvios. Esta perspectiva se ha denominado "enfoque de TT como platitudes".

Sin embargo, no todos los funcionalistas adhieren a una teoría de sentido común y se ha establecido una distinción entre el "funcionalismo conceptual" y el "psicofuncionalismo" (Block 1994). Según el funcionalismo conceptual, los términos de estados mentales recogen su significado de una teoría que sistematiza las nociones de sentido común, mientras que, según el psicofuncionalismo, la teoría implicada en *mindreading* es una teoría psicológica empírica y no de sentido común[17].

[17] Otra clasificación de las versiones del funcionalismo distingue entre el funcionalismo filosófico, el funcionalismo *a priori* o conceptual y el funcionalismo psicológico o psicofuncionalismo (Bermúdez 2005). Para el funcionalismo filosófico las generalizaciones causales son transparentes al sujeto psicológico ordinario. La captación de los principios de la teoría psicológica ordinaria o psicología de sentido común es directa, a menos que seamos autistas o niños pequeños. La psicología de sentido

En concordancia con el psicofuncionalismo, Stich & Nichols (2003) postulan la existencia de un cuerpo de conocimiento sobre lo mental representado internamente en la mente/cerebro. Este cuerpo de conocimiento interno guía los procesos mentales que generan nuestras atribuciones, predicciones y explicaciones sobre los estados mentales y la conducta (*mindreading*). Según este enfoque, a medida que avance la investigación de las habilidades mentalistas se irá descubriendo el cuerpo de información subyacente, aunque es discutible que éste se constituya como algo que merezca llamarse una teoría. Esta propuesta destaca la ventaja de postular un cuerpo de información representado internamente frente a la propuesta clásica de una teoría de sentido común subyacente, constituida por perogrulladas. A continuación me referiré a esta ventaja.

Esta propuesta se apoya, principalmente, en evidencia empírica. El estudio empírico de habilidades cognitivas, equiparables en complejidad a la habilidad de *mindreading*, sugiere que la información que utiliza la mente para llevar adelante sus procesos es, en su mayor parte, inaccesible a la conciencia. En base a esto, Stich & Nichols (2003) sugieren que es probable que la información utilizada por los procesos cognitivos que subyacen a *mindreading* tampoco sea accesible a la conciencia. Este presupuesto parece verse apoyado especialmente por estudios que muestran que las personas realizan atribuciones de estados mentales basadas en indicadores sobre lo que ellas mismas no son conscientes. Por ejemplo, se ha mostrado la existencia de un

---

común es lo que todos compartimos y podemos hacer explícito y, mientras ésta es una guía acerca de la naturaleza de los estados mentales, la psicología científica revelará las leyes causales isomórficas a las leyes causales, de nivel personal, de la psicología de sentido común (Fodor 1987). Esta postura se distingue del funcionalismo *a priori* o conceptual, cuyo supuesto básico consiste en que el marco conceptual de la psicología de sentido común puede hacerse manifiesto sin investigación empírica o científica, y que está implícito en la práctica cotidiana. El funcionalismo psicológico o psicofuncionalismo, a diferencia del funcionalismo filosófico, considera que no es fácil identificar la estructura teórica de la psicología de sentido común. Así, cuando se explica el comportamiento no se llevan a cabo subsunciones explícitas bajo generalizaciones sino, más bien, se proponen candidatos (actitudes proposicionales) para dar cuenta del comportamiento. La idea es que, implícitamente, suponemos generalizaciones que vinculan actitudes proposicionales y comportamiento, pero no es fácil determinar cuáles son estas generalizaciones causales. Se necesita investigación para derivar la taxonomía de generalizaciones de sentido común a partir de las prácticas explicativas cotidianas. Esta investigación será llevada a cabo por la psicología científica. La postura escéptica del psicofuncionalismo parece encontrar apoyo en el hecho de que hay pocas leyes en psicología y que las leyes son vistas, más bien, como *explananda*.

amplio rango de "indicadores de engaño", desde inflexiones en la voz hasta cambios en la expresión facial y en la postura corporal, que nos conducen a creer que una persona no cree en lo que nos está diciendo. Sin embargo, las personas no tienen conciencia de que utilizan estos indicadores en sus apreciaciones (Ekman 1985). Contrariamente, el enfoque de las platitudes asume que a *mindreading* subyace una teoría constituida por aquello que las personas pueden conocer más fácilmente (*i.e.* las perogrulladas). En suma, el enfoque de TT no sólo asume que utilizamos una "teoría" en *mindreading*, sino que esta teoría es la que suministra las descripciones del rol funcional/causal que otorgan significado a los distintos tipos de conceptos mentales. La TT asociada al funcionalismo resulta ser una respuesta posible a la cuestión de la naturaleza de los conceptos mentales. Según la versión del funcionalismo de que se trate, las relaciones funcionales pueden estar estipuladas en términos de una teoría de sentido común o bien de un cuerpo de conocimiento sobre lo mental representado internamente en la mente/cerebro. La principal diferencia entre estas dos versiones reside en que el enfoque de las platitudes presupone que los principios que guían *mindreading* son transparentes al agente. Contrariamente, según el enfoque de Stich & Nichols (2003), la información utilizada por la habilidad cognitiva de *mindreading* no es accesible de manera consciente y, de este modo, da cuenta de un hecho ampliamente aceptado en las ciencias cognitivas.

En esta sección, al tratar la relación entre la TT y el funcionalismo, he desarrollado una versión de la TT conocida como el "enfoque del funcionalismo filosófico". Se puede resumir este enfoque en las siguientes tesis. Los términos mentales son términos teóricos (Sellars 1956). Los términos mentales, tales como "creencia" y "deseo", en tanto tales se definen por medio de leyes o generalizaciones (de rol funcional) que conectan *inputs* sensoriales, *outputs* conductuales y otros estados internos (Lewis 1970, 1972)[18]. Este conjunto de generalizaciones legaliformes

[18] Así, este cuerpo de generalizaciones recoge el modo en que están causalmente relacionados los estados mentales con los estímulos sensoriales (*inputs*), otros estados mentales y las respuestas motoras (*outputs*). Esto implica que los conceptos de estados mentales adquieren su significado por el rol funcional/causal que ocupan en la teoría. Según los distintos tipos de funcionalismo, las relaciones nomológicas entre estímulos sensoriales, otros estados mentales y respuestas motoras pueden ser de

conforman una teoría. Las generalizaciones establecen las conexiones legaliformes tal como se conciben desde el sentido común. Así, se trata de perogrulladas, lo más conocido para las personas (*i.e.* las conexiones legaliformes son transparentes al agente). La tarea de atribuir estados mentales a otros consiste en realizar inferencias guiadas por tales generalizaciones a partir de comportamientos observados, condiciones estimulares y estados mentales antecedentes que han sido determinados previamente. En este sentido, los compromisos del funcionalismo filosófico implican una versión estrecha de la TT, que asume la presencia de elementos conceptuales y generalizaciones legaliformes. En este respecto, se distingue de la otra versión filosófica que es la inclusiva o genérica de TT, que presentaré en la sección 3.1.

3. Las teorías de la teoría

La TT no es unívoca, más bien refiere a un conjunto de "teorías de la teoría" que comparten la idea básica de que las personas poseen una teoría psicológica rudimentaria, una "teoría de la mente", que utilizan cotidianamente en la atribución de estados mentales, y en la explicación y la predicción de la acción y el pensamiento en términos mentalistas (*mindreading*). Algunas versiones de TT provienen de la filosofía y otras de la psicología del desarrollo de orientación cognitiva.

En particular, los psicólogos del desarrollo han tratado de determinar el modo de adquisición de la teoría subyacente a *mindreading* a la que denominan "Teoría de la Mente" (en adelante TdM). Los psicólogos adhieren a un enfoque de TT al asumir que los conceptos mentales ordinarios se constituyen como elementos teóricos. Pero, además, las propuestas más cognitivas en el estudio de la adquisición de TdM concuerdan con la TT en otro respecto. Estos enfoques consideran a la TdM un sistema conceptual. La TdM es el sistema que permite asimilar ordenadamente la conducta de

---

sentido común, esto es, transparentes para el agente (funcionalismo de sentido común o de perogrulladas), pueden ser descubiertas desde el sillón, sin investigación empírica ni científica (funcionalismo *a priori*), o pueden llegar a ser descubiertas con investigación empírica o científica puesto que no se conocen (psicofuncionalismo).

los congéneres, de modo tal que cuando le atribuimos a alguien creencias y deseos organizamos su conducta en función de nuestros conceptos (Leslie 1987; Wellman 1990; Perner 1991). Así, adquirir una TdM es adquirir un sistema de conceptos, sin el cual no es posible entender, ni predecir la conducta (Wellman 1990). Sin embargo, estos enfoques han recibido críticas. Una de las principales señala su carga intelectualista. Para algunos resulta difícil adherir a la idea de que las destrezas mentalistas observadas en niños menores de 5 años sean conceptuales.

En principio, no hay acuerdo respecto del modo en que se adquiere la TdM. Para algunos, la información psicológica representada internamente se estructura al modo de una teoría científica, ya que parece estar sujeta a un proceso de cambio conceptual similar al que sucede en la ciencia (Gopnik & Wellman 1992; Gopnik & Meltzoff 1997). Para otros, la información psicológica es innata, y está almacenada en módulos mentales que sólo interactúan de una manera limitada con la información almacenada en otros componentes de la mente (Scholl & Leslie 1999). En este sentido, la disputa entre los defensores del enfoque modular y los defensores del enfoque del cambio conceptual reside en que si el desarrollo resulta ser el rasgo característico de la TdM, en principio, un enfoque modular no parece adecuado. Se asume que los módulos cognitivos se consideran estáticos y, en virtud de esto, no parecen armonizar con el desarrollo. Asimismo, es preciso señalar que la cuestión de la adquisición de la TdM es de interés filosófico en la medida que ilumina respuestas en relación a cuáles son los procesos que intervienen en *mindreading*. A continuación presento los distintos enfoques cognitivos de la TT: la TT genérica o inclusiva, el enfoque del "niño científico" y el enfoque nativista/modular.

3.1. La Teoría-Teoría genérica o inclusiva

Stich, Nichols y sus colegas (Stich & Nichols 1992; 1995; Nichols, Stich, Leslie & Klein 1996) adoptan un enfoque científico cognitivo. En adhesión al objetivo de la ciencia cognitiva contemporánea definido por Cummins (1983) como la explicación de

las habilidades o capacidades cognitivas, estos proponen una explicación para un grupo de habilidades. Entre éstas, se incluyen la descripción de las personas y sus comportamientos en términos intencionales, y la utilización de tales descripciones para la predicción y explicación del comportamiento. Siguiendo "la estrategia explicativa dominante en las ciencias cognitivas", Stich & Nichols (1992: 35) postulan "una estructura de conocimiento" representada internamente, que sirve como guía de la ejecución de la capacidad a ser explicada. En este sentido, el mecanismo cognitivo que subyace a la habilidad de describir, explicar y predecir el comportamiento de las personas (*mindreading*) es una teoría representada internamente (Stich & Nichols 1992: 36).

De acuerdo a esta versión de la TT, la teoría psicológica representada internamente es en su mayor parte "tácita", en el sentido de que, la mayor parte de las veces, el agente no tiene acceso consciente al conocimiento que guía su comportamiento (Stich & Nichols 1992: 36), ni es consciente de estar utilizando tal conocimiento (Stich & Nichols 1992: 39). Según Stich & Nichols (1992: 36) se trata de una teoría tácita en el sentido de Chomsky (1965) o "subdoxástica" en el sentido de Stich (1978). De este modo, puede considerarse que los contenidos de la teoría psicológica representada internamente no son accesibles así como no lo son las reglas de la gramática. Me ocuparé con más detalle de esta analogía lingüística en la sección 4.

No obstante, la noción de "estructura de conocimiento" se entiende en un sentido amplio o genérico. Se trata simplemente de una estructura de conocimiento (interna) estructurada de algún modo. Esta noción de "teoría" o "estructura de conocimiento" es amplia en, al menos, tres aspectos. En primer lugar, Stich & Nichols admiten que cualquier cuerpo de información sobre cuestiones psicológicas, sea correcto o incorrecto cuenta como una teoría (Stich & Nichols 1992: 67, 69; 1995: 88). El segundo aspecto concierne al término "teoría". Cualquier estructura de conocimiento cuenta como una teoría, esté constituida por leyes psicológicas o no:

> [H]ay muchos dominios de conocimiento de sentido común en los cuales es casi imposible suponer que la "estructura de conocimiento" representada internamente incluya constructos teóricos vinculados de modo legaliforme. El conocimiento sobre cocinar o sobre temas de actualidad son los candidatos, así como el conocimiento que subyace a nuestros juicios sobre lo que es cortés o descortés en nuestra cultura. Y es totalmente posible que el conocimiento de psicología de sentido común termine pareciéndose a las estructuras de conocimiento de los juicios relacionados con la cocina o con lo cortés más que a las estructuras de conocimiento que subyacen a las predicciones y explicaciones científicas producidas por físicos y químicos competentes. (Sitch & Nichols 1995: 88, mi traducción)

Stich & Nichols quieren destacar que el hecho de que sea poco plausible que los dominios del sentido común impliquen la representación interna de constructos teóricos vinculados de manera legaliforme no afecta su propuesta. Al no comprometerse con un tipo de estructura específica, la teoría psicológica representada internamente puede incluir desde generalizaciones legaliformes, tal como sugiere el funcionalismo, o constructos teóricos, tal como sugiere la analogía con el cambio teórico en ciencia, hasta reglas heurísticas y casos prototípicos. En este sentido, este enfoque se diferencia del funcionalismo filosófico, que hemos visto en la sección anterior, que establece un compromiso fuerte con la presencia de enunciados legaliformes.

El tercer aspecto de esta "amplitud" concierne a la noción de "representación interna". Stich & Nichols (1992, 1995) consideran que la información en la "caja de psicología de sentido común" puede estar representada en un rango amplio de formatos. Desde un formato oracional explícito o basado en reglas explícitas, hasta formatos no oracionales como la codificación en redes neuronales, donde el

conocimiento está almacenado en la fuerza de las conexiones entre los nodos de la red. De este modo, no queda excluida la posibilidad de que las redes conexionistas constituyan el conocimiento tácito de una teoría aún cuando las redes no hacen uso de representaciones (explícitas) en formato linguaforme. Incluso, queda abierta la posibilidad de que el conocimiento pueda estar almacenado en otros formatos, como la codificación en modelos mentales al estilo de Johnson-Laird (1983).

La motivación de Stich & Nichols (1992) para adoptar una noción amplia de teoría representada internamente reside en estructurar el debate respecto del mecanismo cognitivo subyacente a *mindreading*, de modo tal que la TT inclusiva y la TS constituyan las únicas dos alternativas (Stich & Nichols 1995: 90). Así, el propósito de esta propuesta es encontrar una estrategia para ganar el debate entre TT y TS. En este sentido, una definición muy amplia de conocimiento permite a Stich & Nichols considerar a la TT y la TS mutuamente exclusivas, ya que cualquier uso de información en los procesos de mentalización se considera como incompatible con un enfoque de TS (Nichols *et al.* 1996). No obstante, es discutible que exista una noción de "teoría" tan inclusiva que permita abarcar todo el espacio lógico que no es ocupado por la simulación. Es más, sólo la versión "radical" de la TS excluye el recurso a los conceptos en la atribución mentalista (Gordon 1995a; 1995b; 1995c; 1995d). Las versiones restantes de TS asumen que, en alguna medida, el procedimiento simulacional está mediado por conceptos (Goldman 1993a; 1993b; Heal 1995b; 2003). Por esto, en principio, resulta una simplificación caracterizar el debate teoría-simulación como un enfrentamiento entre "uso de información" *versus* "no uso de información", tal como parece proponer el enfoque genérico.

3.2. El enfoque del "niño científico"

El enfoque del niño científico proviene de la psicología del desarrollo de tendencia cognitiva y asume que el desarrollo cognitivo implica la internalización de una psicología intencional. Según ésta, en concordancia con la teoría psicológica

ordinaria, el comportamiento del agente es causado por deseos y creencias. Este enfoque postula una teoría tácita subyacente al entendimiento temprano de la mente o "Teoría de la Mente" (TdM). La teoría tácita es análoga a las teorías científicas en tanto la estructura conceptual de los niños es teórica (Gopnik & Meltzoff 1997: 11), en el sentido de que los conceptos mentales ordinarios son constructos teóricos, relacionados entre si de manera legaliforme, de modo que permiten explicaciones causales y facilitan predicciones (Gopnik & Wellman 1992: 147). El enfoque del niño científico se caracteriza por afirmar que ocurren cambios en el entendimiento mentalista de los niños y que estos cambios pueden entenderse como cambios conceptuales similares a los que tienen lugar en ciencia (Gopnik & Wellman 1992: 145). Es más, el argumento que sostiene el carácter teórico de los conceptos mentales descansa, particularmente, en la afirmación de que estos cambian durante el desarrollo infantil de manera análoga al cambio conceptual en las teorías científicas.

Ahora bien, qué implica que los conceptos mentalistas tempranos estén sujetos al cambio conceptual y que este cambio conceptual sea equiparable al que ocurre en las teorías científicas. Por un lado, los niños son vistos como pequeños científicos que construyen y revisan su pensamiento en relación a varios dominios de conocimiento (física, biología, psicología). En este sentido, el desarrollo de *mindreading* parece depender de habilidades generales para teorizar, que en la infancia se utilizan para formar y cambiar la TdM y, en la adultez, para formar y cambiar teorías científicas (Gopnik & Meltzoff 1997). Por otro lado, implica que el rasgo clave de *mindreading* es el desarrollo, esto es, el cambio de una teoría mentalista rudimentaria a una más compleja. A continuación me ocuparé, en primer lugar, de los rasgos de la comprensión mentalista temprana que permiten establecer una analogía con las teorías científicas y el cambio conceptual asociado a éstas. En segundo lugar, caracterizaré las sucesivas "teorías de la mente" en la comprensión mentalista infantil, tal como las conciben los teóricos del enfoque del niño científico.

En principio, la comprensión mentalista temprana es asimilable a una teoría en la medida que presenta rasgos que pueden considerarse característicos de una teoría

científica (Gopnik & Meltzoff 1992: 146-149). Ésta, como las teorías científicas, apela a constructos teóricos que proveen explicaciones causales de fenómenos observables, y están relacionados entre si de manera legaliforme. En este caso, se trata de los estados intencionales (deseos, creencias) y las generalizaciones que establecen las relaciones de estos entre sí y con el comportamiento. Como sucede en las teorías científicas, la conexión legaliforme entre conceptos le otorga a la comprensión mentalista, fuerza explicativa, capacidad de predicción y capacidad de interpretación de la evidencia. De modo que un niño provisto con una teoría rudimentaria puede interpretar hechos fundamentales de modo diferente a cómo puede interpretarlos un niño que posee una teoría completamente desarrollada.

Este enfoque otorga mayor importancia a los rasgos asociados al cambio conceptual en las teorías científicas. Desde el punto de vista del desarrollo, los rasgos asociados a los procesos de formación y cambio en las teorías son más relevantes porque refieren a aquello sobre lo que el "desarrollo" mismo quiere dar cuenta: el cambio. En principio, no se ha establecido un algoritmo para el cambio conceptual. No obstante, pueden señalarse ciertos procesos intermedios que típicamente ocurren en una transición entre teorías. Gopnik & Wellman (1992) están pensando en los procesos asociados a la acumulación de contraevidencia. En este caso, la reacción inicial típica es la negación de la misma. Los mecanismos interpretativos de la teoría tratarán a la contra evidencia, por ejemplo, como "ruido". Posteriormente, la teoría desarrollará hipótesis para dar cuenta de la contra evidencia de manera *ad hoc*. Particularmente, estas hipótesis son útiles porque expresan la contra evidencia en términos del lenguaje de la teoría. Finalmente, se requiere la formulación de un modelo alternativo a la teoría original.

Gopnik & Wellman (1992: 149) proponen que estos rasgos dinámicos de las teorías aparecen en la transición de una TdM rudimentaria a una TdM completamente desarrollada. Al parecer, primero, los niños ignoran cierto tipo de contra evidencia. Luego, utilizan hipótesis auxiliares para dar cuenta de la misma. Posteriormente, los niños utilizan nuevas ideas en contextos acotados y, finalmente, reorganizan su

conocimiento de modo tal que las nuevas entidades tengan un rol central en la teoría. A continuación me referiré al cambio observado en el desarrollo de la comprensión mentalista temprana donde, según Gopnik & Wellman, se aprecian los rasgos mencionados.

El desarrollo infantil, entre los 2 años y medio y los 4 años, presenta los siguientes procesos intermedios. Alrededor de los 2 años, la comprensión mentalista infantil se estructura en términos de estados internos no representacionales, específicamente, los deseos y las percepciones. En este marco, los deseos se entienden como un impulso hacia los objetos (Wellman & Woolley 1990) y la percepción, simplemente, como "darse cuenta" de los objetos (Flavell 1988). De este modo, los niños no tratan a las percepciones y a los deseos como estados representacionales, sino como vínculos causales simples con el mundo. Esto es, si alguien desea un objeto, actuará para obtenerlo. Si un objeto está en el campo visual de un agente, entonces éste lo ve. Estos constructos causales son simples, pero tienen poder predictivo. En conjunto proveen una forma inicial de silogismo práctico: "si un agente desea *X,* y ve *X,* hará algo para obtener *X"*. Esto resulta suficiente para que los niños puedan entender que los deseos modifican el mundo y que el mundo modifica las percepciones.

A los 3 años, los niños se encuentran en una etapa intermedia en la que tienen cierta comprensión representacional de las percepciones y los deseos, por ejemplo, son capaces de entender que las personas pueden tener diferentes deseos. Pero no tienen una comprensión representacional de las creencias. La comprensión de las creencias en esta etapa se acota a la relación directa entre las creencias y el mundo. Este entendimiento es considerado no representacional en tanto los niños aún no son capaces de comprender que las creencias pueden representar al mundo de manera incorrecta. No obstante, si a esta edad los niños son guiados hacia el reconocimiento de creencias "representacionales", son capaces de comprender que existen. Por ejemplo, los niños son capaces de explicar acciones que no han podido ser completadas a raíz de estar basadas en creencias falsas. Sin embargo, aún no son capaces de predecir una creencia, ni de diagnosticar su contenido en base a creencias

falsas. Según Gopnik & Wellman (1992), este entendimiento acotado de las creencias falsas revela que el concepto es aún periférico a su teoría central, es decir, que lo entienden al modo de una hipótesis auxiliar.

Finalmente, a los 4 años, los niños ya poseen una TdM completamente desarrollada. Ésta les permite concebir los estados mentales, en general, como representacionales. Los niños ya son capaces de advertir que el actor piensa y que las acciones no están determinadas por el mundo, sino por la representación que el actor tiene del mundo. Gopnik & Wellman (1992) consideran que, en este momento del desarrollo, se reorganiza la teoría central de los niños como una psicología representacional. Posteriormente, esta teoría se conserva y se seguirá sofisticando en la adultez.

No obstante, este enfoque presenta dificultades en el orden conceptual y empírico. En el plano conceptual, se cuestiona la falta de analogía entre el cambio conceptual en ciencia y el cambio conceptual en *mindreading*. En general, aún no es claro cómo opera el cambio conceptual en ciencia. Pero si concedemos que se pueden determinar los rasgos señalados por Gopnik & Wellman (1992), surgen algunas dificultades. En primer lugar, una de las características distintivas de la ciencia profesional reside en que sus practicantes generan teorías diferentes. No disponemos de evidencia respecto de que, en la ciencia profesional, cada científico particular arribe a los mismos principios teóricos (Goldman 2006: 85). No obstante la convergencia entre científicos puede existir pero luego de largos períodos de debate y experimentación, pero las carreras de los científicos se desarrollan a partir de los desacuerdos con otros en cuestiones teóricas. En contraste, todos los niños desarrollan su propia teoría, tal como ocurre con los científicos pero, a diferencia de estos, los niños convergen en una misma teoría. Y, además, lo hacen en un período temporal breve. En este respecto, la analogía entre el desarrollo de *mindreading* y la ciencia profesional parece difícil de establecer. En segundo lugar, el cambio conceptual en ciencia asume que los científicos modifican sus teorías cuando advierten su falsedad o, al menos, cuando advierten que éstas conducen a predicciones falsas. Sin embargo, el

pequeño niño científico no estaría equipado para decidir que su teoría de la mente es falsa. Según este enfoque, los niños manejan un concepto rudimentario de estado mental que no los capacita para concebir la posibilidad de representaciones incorrectas del mundo. De este modo, su déficit conceptual parece impedirles concebir la falsedad (Carruthers 1996).

Las dificultades de orden empírico están asociadas a cierto cúmulo de evidencia problemática para este enfoque. Un primer conjunto de hallazgos empíricos desafía la idea de que el desempeño infantil en la TFC sólo pueda ser explicado por un déficit conceptual, esto es, por la presencia de una concepción temprana no-representacional de los estados mentales. En primer lugar, se ha observado que los niños pequeños mejoran su desempeño si se les brinda ayuda relacionada con la memoria. Por ejemplo, en la tarea del recipiente engañoso (ver la nota 6 de la introducción), si se les brinda ayuda relacionada con la memoria, los niños pueden recordar y reportar su predicción inicial falsa, esto es, responden que la primera vez que vieron el tubo de "*smarties*" dijeron que en su interior había confites y no un lápiz (Mitchell & Lacohee 1991). En este sentido, es posible que el desempeño deficitario en la tarea del tubo engañoso refleje demandas altas en relación a la memoria para niños pequeños y que no esté específicamente relacionado con un déficit conceptual en *mindreading*. En segundo lugar, las TFC parecen requerir funciones ejecutivas que pueden resultar demandas excesivas para niños tan pequeños. Se ha observado que cuando, en la TFC de cambio de lugar (ver la sección 1.1 de la introducción), se les presenta a los niños la realidad de manera menos saliente, estos son capaces de dar la respuesta correcta. Por ejemplo, en lugar de que los niños vean por sí mismos dónde está el chocolate, sólo se les comunica verbalmente la ubicación del mismo (Zaitchik 1991).

Asimismo, hay evidencia adicional que sugiere que el problema de los niños de 3 años está relacionado con dificultades en el control inhibitorio (Carlson & Moses 2001). El control inhibitorio es una habilidad ejecutiva que permite ignorar el estímulo saliente o dominante, en este caso, la realidad tal como el niño la percibe. La TFC requiere obviar el estímulo saliente, lo que el niño conoce sobre la realidad, y llevar

esto a cabo puede resultar una tarea muy difícil para un niño de 3 años. Un año más tarde, a los 4 años, las habilidades ejecutivas se encuentran más maduras y esto puede resultar crucial en el desempeño en la TFC, sin que este mejoramiento implique un cambio conceptual en los niños. Incluso, cierta evidencia sugiere que hay una comprensión temprana de las creencias falsas. Onishi & Baillargeon (2005) han empleado un paradigma experimental con demandas muy reducidas. Se trata de la TFC no verbal, que fue diseñada con el propósito de evaluar la posibilidad de que niños de apenas 15 meses de edad puedan apreciar las creencias falsas. Los resultados de este estudio aportaron indicadores positivos de entendimiento y sugieren que hay entendimiento de las creencias mucho antes de lo contemplado por los teóricos del enfoque del niño científico (ver nota 7 de la Introducción).

Un segundo conjunto de evidencia problemática proviene de estudios del síndrome de Williams, que muestran una disociación entre las capacidades cognitivas de cambio conceptual y mentalista. Los pacientes con síndrome de Williams tienen retraso mental, alcanzan un coeficiente intelectual promedio de 50 e, incluso siendo adolescentes y adultos, parecen ser incapaces de experimentar el cambio conceptual asociado con la construcción de teorías (Johnson & Carey 1996). Sin embargo, los niños con síndrome de Williams comienzan a explicar las acciones en términos de deseos y creencias a la misma edad que los niños con un desarrollo típico (Tager-Flusberg 2000). Esta evidencia señala una disociación entre las habilidades para teorizar y *mindreading*, y sugiere que *mindreading* no es el producto de una habilidad general para teorizar.

3.3. El enfoque nativista/modular

El enfoque nativista o modular se asocia a la propuesta de un mecanismo de "Teoría de la Mente" (MTdM), sostenida por Leslie y sus colegas (Leslie 1987, 1994; Leslie & Thaiss 1992; Leslie & Pollizzi 1998; Leslie & Scholl 1999; Scholl & Leslie 2001; Leslie, Friedman & German 2004; Leslie, German & Pollizzi 2005). Este enfoque se caracteriza por postular la existencia de un mecanismo específico subyacente a

mindreading, en contraposición con el enfoque del niño científico que postula habilidades cognitivas de dominio general. Este mecanismo es un módulo que procesa espontánea y post perceptualmente los comportamientos, permite que estos sean atendidos y computa los estados mentales que contribuyen a los mismos (Scholl & Leslie 2001: 697). Al hacer esto, un concepto innato de creencia es puesto a disposición del niño mucho antes de que éste haya adquirido otros conceptos abstractos mediante la construcción general de teorías. Como resultado de esto, el MTdM provee, al niño pequeño, de un *insight* intencional respecto del comportamiento de los otros.

Este enfoque asume explícitamente una noción fodoriana de módulo (Fodor 1983)[19]. Según Scholl & Leslie (1999: 134), la aplicación de la noción de modularidad al dominio de *mindreading* resulta en la afirmación de que esta competencia cognitiva tiene una base innata específica. "Base" en tanto *mindreading* no es modular en su totalidad sino que sólo lo es la competencia temprana de *mindreading*. El MTdM intenta captar el origen y no el rango total de las actividades maduras que emplean la habilidad de *mindreading* (Scholl & Leslie 1999, 2001). "Específica" en el sentido de que el mecanismo utiliza representaciones especiales (metarrepresentaciones), que no se aplican a otros dominios cognitivos y que pueden estar dañadas selectivamente[20].

[19] Me ocuparé de esta noción más adelante en esta sección. De manera que, sólo para tener una idea, los módulos son sistemas computacionales de dominio específico caracterizados por el encapsulamiento informativo, la operación a alta velocidad, el acceso restringido a las metas y los propósitos del organismo, la operación de manera obligatoria, la generación de *outputs* superficiales, una arquitectura neuronal fija, un patrón de daño específico, y una secuencia y ritmo de maduración característicos. Según Fodor (1983), la modularidad de los sistemas analizadores de *input* consiste en la posesión de la mayoría o la totalidad de las propiedades mencionadas. No obstante, algunas de las propiedades son esenciales a la modularidad, por ejemplo el encapsulamiento informativo que consiste en que los sistemas de *input* no tienen acceso a las expectativas ni a las creencias del organismo para generar sus *outputs*, y es probable que un mecanismo psicológico que no cumpla con la misma no constituya un módulo (Fodor 2000).

[20] "Metarrepresentación" es un término técnico en este modelo. Según Leslie (1994), una metarrepresentación está conformada por una actitud (*e.g.* creencia, deseo, fingimiento) seguida de tres argumentos que especifican un agente, un anclaje o algún aspecto de la situación real (*e.g.* el agente sostiene una banana) y un estado imaginario o fingido (*e.g.* el agente sostiene un teléfono). El ejemplo típico de metarrepresentación es "mamá finge que esta banana es un teléfono". Las metarrepresentaciones se caracterizan por ser construcciones opacas, esto es, construcciones donde las relaciones referenciales están suspendidas. Contrariamente, las representaciones primarias se definen por su relación directa con el mundo. Una relación que es adecuada, fiel y literal. Las habilidades perceptuales son, según Leslie (1994), una manifestación de la capacidad de representación primaria del niño.

"Innata" en el sentido de que los conceptos básicos de *mindreading* (creencia, deseo y ficción) son parte del equipamiento genético innato que es desencadenado por factores ambientales apropiados, tal como la pubertad se desencadena y no se aprende (Scholl & Leslie 1999). Este enfoque considera que la "teoría" de la mente en su estado final se aprende, pero su base, los conceptos de creencia, deseo y ficción, son innatos y no se adquieren por aprendizaje. En este sentido, se asume que la adquisición normal de la competencia de *mindreading* se debe, en parte, a la operación del MTdM.

No obstante, el MTdM tiene poderes limitados. Leslie & Thaiss (1992) sostienen que el MTdM necesita ser complementado en ciertas situaciones para que el niño pueda seleccionar el contenido correcto de los estados mentales, particularmente, en el caso de tener que elegir contenidos de creencias que son falsas. Este enfoque propone que el MTdM es complementado por el procesador de la selección (PS), un proceso ejecutivo general requerido en muchas situaciones para inhibir respuestas salientes pero indeseadas. Este complemento es necesario porque el modelo asume, conceptualmente, que el MTdM atribuye automáticamente creencias verdaderas. La estrategia por *default* del MTdM pretende reflejar el hecho de que las creencias deben ser verdaderas y que, usualmente, lo son. No obstante, la misma implica el costo de que, en situaciones de creencia falsa, la respuesta preponderante necesita ser inhibida (Leslie & Thaiss 1992; Leslie & Pollizzi 1998; Scholl & Leslie 1999). Específicamente, en la TFC estándar, para captar el contenido de la creencia del protagonista respecto de la ubicación del objeto/blanco (la tarea correspondiente al MTdM) es preciso inhibir primero la respuesta privilegiada por el MTdM asociada a la ubicación actual y saliente del objeto. Este es el trabajo que realiza el PS.

En base a esto, el enfoque del MTdM/PS da cuenta del desarrollo infantil de *mindreading* entre los 2 y medio y los 4 años. Esto está relacionado con el éxito en la TFC mencionado anteriormente, a partir de la distinción entre competencia y desempeño. Si bien la competencia está intacta y los niños pequeños poseen el concepto de creencia, los errores en la TFC se deben a dificultades en el desempeño.

Éstas están relacionadas con otras habilidades cognitivas aún no maduras y requeridas para llevar a cabo esta tarea de manera exitosa (Fodor 1992; Scholl & Leslie 1999), en particular, aún no han madurado las funciones ejecutivas asociadas al PS. Al mismo tiempo, el enfoque del MTdM/PS da lugar a un modelo sobre el desarrollo típico y no típico. Se asume que el MTdM está selectivamente dañado en el autismo y que, por esta razón, los niños con autismo fallan en la TFC a pesar de tener intacto el PS. Contrariamente, los niños con desarrollo típico menores de 3 años tienen un MTdM intacto, pero fallan en la TFC porque aún no han madurado sus habilidades ejecutivas de control inhibitorio.

La evidencia empírica que brinda apoyo al enfoque modular se relaciona con estudios que sugieren que *mindreading* presenta un patrón de daño específico: el autismo (Baron-Cohen, Leslie & Frith 1985; Baron-Cohen, Leslie & Frith 1986; Leslie 1991; Roth & Leslie 1991; Leslie & Thaiss 1992; Baron-Cohen 1995). Por un lado, según los investigadores, la especificidad de dominio es sugerida por la disociación hallada entre las habilidades de inteligencia y la mentalista. El estudio seminal de Baron-Cohen, Leslie & Frith (1985) mostró que un alto porcentaje de niños con desarrollo típico y con síndrome de Down, todos con edad mental de 4 años, se desempeñó exitosamente en la TFC, mientras que sólo el 20 por ciento de los niños con autismo tuvieron éxito. Un segundo estudio mostró un resultado similar (Baron-Cohen, Leslie & Frith 1986). Los niños con autismo se mostraron incapaces de ordenar historietas mentalistas y no pudieron contar estas historias con coherencia. Contrariamente, ordenaron correctamente las historias mecánicas y comportamentales que, además, pudieron relatar con sus propias palabras. En base a estos resultados, los investigadores propusieron la hipótesis de un daño específico de *mindreading* en los niños con autismo, que sugiere la existencia de un mecanismo específico. Por otro lado, la relación entre las habilidades mentalistas y las habilidades de control inhibitorio señaladas por el modelo MTdM/PS encuentra apoyo en evidencia que sugiere una relación entre de *mindreading* y la función ejecutiva de control inhibitorio

(Hugges 1998; Carlson & Moses 2001). Me ocuparé de la relación entre estas habilidades con detalle en el capítulo 5 (sección 3).

Más allá de los hallazgos empíricos que parecen apoyar al enfoque nativista/modular, éste presenta dificultades de orden conceptual. La principal objeción se relaciona con la noción de módulo utilizada. Como mencioné, este enfoque adhiere explícitamente a una noción fodoriana de módulo. Particularmente, se asume que las interpretaciones, explicaciones y predicciones de *mindreading* parecen ser específicas de dominio, en el sentido de que implican tipos especiales de representación (las metarrepresentaciones) y de computaciones que no se aplican a otros dominios cognitivos. Además, éstas son rápidas y ocurren sin esfuerzo cognitivo, y son obligatorias, en el sentido de que no se puede decidir dejar de interpretar muchas situaciones como implicando agentes intencionales, aunque es posible ignorar esta interpretación. Asimismo, parece existir un patrón específico de daño del MTdM: el autismo (Scholl & Leslie 1999: 134-135).

Sin embargo, en mi opinión, el MTdM subyacente a *mindreading* no reúne las propiedades de un módulo en sentido fodoriano. Para mostrar esto, a continuación introduciré la noción fodoriana de módulo. Según Fodor, los sistemas analizadores de *input* son los mecanismos cognitivos que pueden ser tratados como módulos. Los sistemas de *input* analizan los *outputs* de los transductores de modo tal de generar representaciones en un código que pueda ser procesado por los sistemas centrales. Al realizar esto, llevan a cabo computaciones que proceden asignando análisis intermedios de la estimulación próxima. En cada nivel de representación se llevan a cabo inferencias no demostrativas, que van generando representaciones más abstractas en relación al nivel de representación inmediatamente anterior. Es preciso señalar que no se trata de una simple traducción del estimulo próximo, ya que el *output* de los módulos representa la disposición de las cosas en el mundo (el estímulo distal).

La afirmación de que los sistemas de *input* están encapsulados informativamente refiere a que los sistemas de *input* no tienen acceso a las

expectativas, ni a las creencias del organismo, para generar sus *outputs*. Esto no es un problema para la operación de los módulos porque los mismos almacenan la información que precisan para realizar sus computaciones y sólo esta información es la que afecta la generación de hipótesis al interior del módulo. En este sentido, el encapsulamiento informativo implica que las operaciones de los sistemas de *input* no tienen acceso a toda la información que el organismo tiene representada internamente y que los datos que interesan para la confirmación de las hipótesis perceptuales incluyen, en general, mucho menos que lo que el organismo sabe (o sus creencias y expectativas).

En base a esto, a mi entender, resulta difícil concebir al MTdM como un módulo fodoriano encapsulado informativamente puesto que, en principio, el entendimiento mentalista no parece estar aislado de otros mecanismos de procesamiento cognitivo (sean centrales o modulares). Al explicar y predecir el comportamiento de una persona hacemos uso de toda la información disponible que podamos recolectar. Sin embargo, la posibilidad de acceder a información perteneciente a otros sistemas cognitivos es característica de los sistemas centrales pero no de los sistemas modulares. En este sentido, el MTdM no parece cumplir con la propiedad de encapsulamiento informativo, que, según Fodor (1983: 71), constituye la esencia de la modularidad.

A su vez, en la caracterización de Scholl & Leslie (1999) se afirma la especificidad de dominio de la TdM. Según Fodor, la especificidad de dominio refiere al hecho de que sólo una clase relativamente restringida de estímulos puede "encender" el analizador de *input*. En este sentido, existe un dominio estimular específico que es analizado por el correspondiente mecanismo psicológico de análisis de *input*. Sin embargo, en el caso de *mindreading*, resulta difícil demarcar el campo de las situaciones sociales e identificar el estímulo relevante que desencadene la operación del módulo de TdM. En mi opinión, es dudoso que exista una clase de estímulo que satisfaga esta condición. Incluso, tal como señala Bermúdez (2005: 182), aunque hubiera un dominio social específico, aún resta la cuestión de que el entendimiento

mentalista resulta sensible al contexto, y el contexto no es puramente social. En este sentido, el MTdM no satisface la propiedad de especificidad de dominio.

No obstante, quizás pueda concedérsele a Scholl & Leslie (1999, 2001) que las descripciones, predicciones y explicaciones de *mindreading* se llevan a cabo, en algunos casos, rápidamente y sin esfuerzo cognitivo, que parecen ser obligatorias, que parece haber una secuencia y ritmo característicos de maduración, que parece existir un patrón de daño específico (el autismo) y que, para algunos, hasta existe una arquitectura cerebral fija (Frith & Frith 2003). Si bien todos estos rasgos son compartidos con los módulos, no parecen ser suficientes para afirmar la existencia de un MTdM, puesto que éste no satisface algunas de las propiedades más esenciales. Tal como se mencionó, el MTdM no satisface las propiedades de encapsulamiento informativo, ni de especificidad de dominio, que resultan más esenciales que los rasgos mencionados (Fodor 2000). Además, a mi entender hay otra dificultad para considerar al MTdM como un módulo fodoriano. A saber, Scholl & Leslie (1999) se refieren explícitamente al MTdM como un mecanismo post perceptual y esto no es compatible con la noción fodoriana de módulo. Según Fodor, los candidatos a ser tratados como módulos son los analizadores de *input* que son mecanismos perceptuales que, si bien no se identifican con la percepción sino con estadios tempranos de la misma, no se trata de mecanismos post perceptuales. Justamente, todos aquellos procesos que operan post perceptualmente son asociados, en el enfoque fodoriano, a los procesos centrales. Por todo lo mencionado, es probable que el mecanismo de TdM no sea modular en el sentido fodoriano. Quisiera mencionar que si bien me he detenido esta dificultad de la propuesta nativista/modular, en principio, ésta no afecta directamente a las propuestas de los enfoques híbridos de teoría y simulación que analizaré y evaluaré en esta investigación.

4. Una noción relevante de "teoría" para la competencia mentalista

En general, cuando en ciencia cognitiva se estudia la capacidad cognitiva de atribuir estados mentales a uno mismo y a otros, la cuestión nuclear reside en determinar cómo las personas ejecutan esta capacidad cognitiva. Es decir, cómo sus sistemas cognitivos llevan a cabo la tarea de formar creencias o juicios sobre los estados mentales de los otros, estados que no son observables directamente[21]. En relación a esto, los "teóricos de la teoría" de inclinación cognitivista han tratado de ofrecer un enfoque de "teoría" que sea relevante para dar sentido a la afirmación de que a *mindreading* le subyace un cuerpo de conocimiento. Para esto, por un lado, se ha establecido una analogía entre la psicología de sentido común y la lingüística, a partir de la noción chomskiana de conocimiento tácito (Stich & Nichols 1992). Por el otro, se ha propuesto un paralelo entre la estructura y la forma de las explicaciones de psicología de sentido común, y la estructura y la forma de las explicaciones nomológico-deductivas, consideradas las explicaciones características de la ciencia (Fodor 1987). A continuación, me ocuparé de estas analogías.

Como he mencionado en la sección 3, los "teóricos de la teoría" asumen que *mindreading* es una habilidad que involucra conceptos y que, particularmente, requiere la posesión del concepto de creencia. Así, se considera que el desempeño exitoso en la TFC requiere que los niños sean capaces de tener pensamientos con la forma general "él *cree* que *p*" y se asume que para esto, los niños necesitan poseer el concepto de creencia. Ahora bien, poseer el concepto de creencia implica algo más que tener creencias y estar simplemente en estados de creencia. Poseer el concepto de creencia implica que el sujeto debe tener un cuerpo de conocimiento psicológico. Los "teóricos de la teoría" asumen que tal cuerpo de conocimiento puede ser considerado una teoría psicológica (Davies & Stone 1995a: 2-3). Sin embargo, dado que no es claro en qué sentido este cuerpo de conocimiento psicológico ordinario puede ser análogo a

[21] Otras cuestiones estudiadas en ciencia cognitiva son cómo se adquiere esta capacidad mentalista y en qué arquitectura cognitiva o neurocognitiva se apoya, si se recurre a los mismos mecanismos que se utilizan para pensar sobre los objetos en general o se utilizan mecanismos dedicados de dominio específico y cómo se relaciona esta capacidad con otros procesos relativos a la cognición social, tales como la imitación o la empatía.

una teoría científico-empírica, los filósofos y psicólogos han recurrido a otras estrategias para establecer un significado relevante de "teoría".

La analogía entre la capacidad de *mindreading* y la capacidad lingüística está basada en la noción chomskiana de conocimiento tácito. Esta noción surge con el propósito de explicar la capacidad que tiene un hablante ordinario de producir y entender un número indeterminado de oraciones de su lengua materna[22]. En este sentido, los lingüistas generativos postulan una gramática del lenguaje que siendo poseída y utilizada por el hablante de la lengua, resultará en que éste será capaz de producir y entender las oraciones que, de hecho, entiende y produce.

No obstante, se afirma algo más. El hablante "conoce" el lenguaje en virtud de estar en posesión del cuerpo de conocimiento expresado en la gramática. Esta gramática expresa el conocimiento que el hablante tiene efectivamente de su lengua. Sin embargo, esto no es lo mismo que decir que el hablante es consciente de las reglas de la gramática o que puede llegar a ser consciente de las mismas. La evidencia indica que los hablantes, en general, no pueden reportar sus estados lingüísticos y, cuando lo intentan, la mayoría de las veces suelen estar equivocados. En este sentido, los procesos mentales que intervienen en la capacidad del lenguaje están más allá del nivel actual o potencial de conciencia (Chomsky 1965: 8). Además de la evidencia a favor de que los estados lingüísticos son inconscientes, los lingüistas han recurrido, en realidad, a una inferencia a la mejor explicación para defender que hay estados y mecanismos inconscientes: "Dado cierto fenómeno que hay que explicar y dado que, de todas las explicaciones en competencia, la explicación que postula estados y procesos inconscientes es la mejor, según ciertos criterios de elección entre explicaciones, es probable que esta explicación sea la adecuada" (Skidelsky 2007: 34). Al parecer, no hay ninguna teoría que, sin postular un mecanismo inconsciente, pueda dar cuenta de la numerosa evidencia que explica la teoría que lo postula.

[22] Los lingüistas generativos pretenden que la explicación de la capacidad del lenguaje también pueda dar cuenta del desarrollo característico del lenguaje. Esto es, un desarrollo rápido, sin instrucción formal, con independencia del desarrollo de habilidades intelectuales generales y del desarrollo de otras capacidades simbólicas.

De este modo, la estrategia explicativa de los lingüistas chomskianos concuerda con el tipo de explicación habitual y dominante en las ciencias cognitivas (Stich & Nichols 1992). Esta "estrategia explicativa dominante" procede de la siguiente manera:

> [P]ostula una "estructura de conocimiento" representada internamente –generalmente un cuerpo de reglas o principios o proposiciones– que sirve de guía para la ejecución de la capacidad a ser explicada. Estas reglas, principios o proposiciones son a menudo descriptas como la "teoría" del agente sobre el dominio en cuestión. En algunos casos, la teoría puede ser en parte accesible a la conciencia; el agente puede enunciar algunas de las reglas o principios que está utilizando. Más a menudo, sin embargo, el agente no tiene acceso consciente al conocimiento que guía su comportamiento. (Stich & Nichols 1992: 35, mi traducción)

Los "teóricos de la teoría" emplean la analogía de base lingüística para sostener que las personas desempeñan la habilidad *mindreading* utilizando una teoría asimilable a la gramática, en tanto que ésta guía el desempeño de su habilidad. Ahora bien, para los lingüistas la gramática que guía el desempeño del hablante ordinario es conocida y utilizada tácitamente. De este modo, la analogía de base lingüística implica que el cuerpo de conocimiento que guía la habilidad de *mindreading*, también se conoce y se utiliza tácitamente. Sin embargo, asumir un cuerpo de conocimiento tácito en relación a la posesión y uso de conceptos psicológicos en *mindreading* es problemático.

Según Davies & Stone (1995a: 12), si el cuerpo de conocimiento tácito es una teoría conocida tácitamente esto implica una dificultad para aquello asumido por los "teóricos de la teoría" respecto del vínculo entre la posesión de conceptos psicológicos y la posesión de un cuerpo de conocimiento psicológico. Como mencioné, se asume que la habilidad de *mindreading* requiere la posesión del concepto de creencia y que para tener el concepto de creencia es preciso poseer un cuerpo de conocimiento

psicológico (Davies & Stone 1995a: 3). Ahora bien, no se ve cómo se pueden conocer los conceptos implicados en una teoría que se conoce y se utiliza tácitamente, cuando:

> [U]na de las marcas del conocimiento tácito es que su contenido no precisa ser conceptualizado por el agente…Así poseer conocimiento tácito apenas puede constituir la captación de los conceptos. De la misma manera, no es inmediatamente obvio cómo la codificación interna no conceptualizada de los principios de una teoría psicológica puede constituir la captación de conceptos psicológicos. (Davies & Stone 1995a: 12, mi traducción)

Creo que lo que Davies & Stone (1995a) quieren sostener puede entenderse en el marco de la concepción de Davies respecto de qué es lo que está involucrado en la posesión de conceptos. La postura particular de Davies (1989) sobre la conceptualización está ligada a la cuestión de proveer un criterio que permita distinguir entre estados doxásticos y subdoxásticos. A continuación me referiré a esta distinción con el objetivo de introducir la noción de conceptualización que, creo, que subyace a la tensión entre conocimiento tácito y conceptual planteada por Davies & Stone (1995a). Esto permitirá echar luz respecto de si la analogía de base lingüística es apta para dar cuenta de la afirmación de que una teoría subyace a la habilidad de *mindreading*.

Los filósofos acuerdan en la existencia de una distinción intuitiva entre estados doxásticos y subdoxásticos (Stich 1978). Los estados doxásticos o intencionales (EI) son estados como los de creencia. Los estados subdoxásticos (ES) son estados informacionales que desempeñan un papel en la historia causal próxima de las creencias pero que no son creencias. Estos estados psicológicos almacenan información sobre, por ejemplo, las reglas de la gramática que dan lugar a los juicios lingüísticos o los rasgos retinales que dan lugar a los juicios perceptuales. Davies (1989) considera que los criterios de accesibilidad a la consciencia y de integración inferencial

propuestos por Stich (1978) no son suficientes para fundamentar la distinción entre EI y ES. Tales criterios no se sostienen por sí mismos y, según Davies, están basados en el criterio de conceptualización, sobre el que descansa la distinción en última instancia. Particularmente, Davies considera que el requisito de generalidad de Evans (1982) explicita lo que está involucrado en la posesión de conceptos.

> Así, si a un sujeto se le atribuye el pensamiento de que *a* es *F*, entonces debe tener los recursos conceptuales para tener el pensamiento de que *a* es *G*, para cualquier propiedad de ser *G* de la cual tiene una concepción. Ésta es la condición que llamo "el requisito de generalidad." (Evans 1982: 104, traducción Skidelsky 2007: 45)

Según el requisito de Evans, tener creencias depende de habilidades conceptuales estructuradas. Esto es, las creencias (pensamientos, en general) son estados estructurados en el sentido de ser un complejo de habilidades, por ejemplo, creer *Fa* consiste en el ejercicio de dos habilidades, la correspondiente a *a* y a *F*. Así, el pensamiento *Fa* puede concebirse como la intersección de dos series, la de *Fa*, *Fb*,…, y la de *Fa*, *Ga*,… Si un sujeto tiene los pensamientos estructurados de que *a* es *F* y *b* es *G*, no habría en principio ninguna barrera conceptual para tener el pensamiento de que *a* es *G* o que *b* es *F*. En esta capacidad de ejercitar habilidades cognitivas consiste la posesión de conceptos (Skidelsky 2007: 45).

Siguiendo la propuesta de Evans (1982: 104, n 22) de que este requisito se aplica al ámbito intencional pero no a los estados informacionales, Davies (1989) considera que el requisito de generalidad provee el criterio para distinguir entre EI y ES porque explicita lo que está involucrado en la posesión de conceptos. Así, los EI tienen contenido conceptual y es preciso ejercitar habilidades conceptuales para estar en un EI. Contrariamente, los ES tienen contenido no conceptual y no se requiere que el sujeto ejercite habilidades conceptuales para estar en un estado informacional.

Mencionado esto, voy a retomar la dificultad que encuentran Davies & Stone (1995a) en relación a una teoría psicológica tácita. Creo que esta dificultad puede entenderse de la siguiente manera. Los ES son estados representacionales con contenido no conceptual y no se requiere que el sujeto ejercite habilidades conceptuales para estar en un estado informacional. En principio, no hay ninguna dificultad para considerar que la gramática (o la habilidad lingüística) está conformada por estados informacionales. Usualmente, se asume que los estados informacionales (ES) son estados psicológicos que almacenan información sobre, por ejemplo, las reglas de la gramática que dan lugar a los juicios lingüísticos. Además, el hecho de no poder acceder conscientemente al conocimiento efectivo que se tiene del lenguaje es concordante con la idea de que no es necesario ejercitar habilidades conceptuales para estar en un estado informacional. De este modo, no parece existir tensión alguna al considerar que la gramática se conoce tácitamente, tal como asume la lingüística chomskiana.

Sin embargo, siguiendo el criterio de Davies no puede decirse lo mismo de la teoría subyacente a *mindreading*. Ésta está conformada por EI y no por estados informacionales. De este modo, surge una tensión entre concebir un cuerpo de conocimiento como tácito y, a la vez, conceptual. La tensión señalada por Davies & Stone (1995a) puede reformularse como la tensión que implica concebir a la teoría subyacente a *mindreading* como tácita y, en este sentido, implicando representaciones con contenido no conceptual cuando, en realidad, es inherente a la misma que sus representaciones sean de tipo conceptual. Esto es así, por un lado, porque la teoría subyacente a *mindreading* está conformada por conceptos psicológicos y principios que establecen relaciones entre los mismos. Por otro lado, porque la habilidad de *mindreading* implica la posesión de conceptos psicológicos como el de creencia. En este sentido, la analogía de base lingüística no logra proveer un sentido de "teoría" subyacente a *mindreading*, que dé cuenta de la posesión de conceptos psicológicos.

En relación a la segunda analogía, para algunos es posible establecer una analogía con la ciencia pero limitándola a ciertos rasgos de las teorías científico-

empíricas. Según Fodor, existe una similitud entre las explicaciones psicológicas de *mindreading* y las explicaciones científicas. Esta similitud es en dos respectos. Primero, las generalizaciones que subyacen a la teoría se definen sobre no observables. Segundo, esas generalizaciones conducen a sus predicciones cuando se iteran e interactúan más que cuando se ejemplifican directamente (Fodor 1987: 24)[23]. Así, como sucede en ciencia, la explicación mentalista incluye generalizaciones que postulan entidades inobservables y estas entidades inobservables juegan un rol explicativo como portadoras de poderes causales. Usualmente, estas generalizaciones recogen el rol causal de las creencias y las preferencias para la decisión, la intención y la acción. En este sentido, se considera una generalización típica de *mindreading*, por ejemplo, "una persona que quiere *P* y cree que *Q* es suficiente para obtener *P*, si no tiene otros deseos que entren en conflicto u otras estrategias preferidas, tratará de obtener *Q*" (Churchland 1988).

No obstante, Fodor sostiene algo más respecto de la relación entre la teoría que subyace a *mindreading*, o la psicología de sentido común, y la ciencia, cree que la psicología científica está en condiciones de reivindicar a la psicología de sentido común (Fodor 1987: 10). Más precisamente, cierta teoría resulta ser, al mismo tiempo, la mejor explicación científica disponible respecto de la mente y la teoría empírica que está comprometida ontológicamente con los estados mentales tal como son concebidos por la psicología de sentido común. La teoría en cuestión es la Teoría Representacional de la Mente (TRM) asumida, según Fodor (1975), por la psicología cognitiva. La TRM permite, según Fodor, dar cuenta de las características de lo mental señaladas por la psicología de sentido común a las que me referiré a continuación.

La psicología de sentido común es la teoría compleja e implícita que conecta los deseos, las creencias y las acciones. Según esta teoría, los estados psicológicos son actitudes proposicionales. Específicamente, estos consisten en una actitud (creencia,

[23] Además, Fodor (1987: 7) asume que las explicaciones de psicología de sentido común tienen la estructura de las explicaciones nomológico-deductivas características de la ciencia. Las generalizaciones de sentido común relacionan estados de actitud proposicional, portadores de poderes casuales, y que participan en la decisión, intención y acción.

deseo) dirigida hacia un contenido que tiene una forma proposicional. Según la psicología de sentido común los estados psicológicos poseen las siguientes propiedades: (a) son semánticamente evaluables, esto es, las creencias pueden ser verdaderas o falsas en virtud de su relación con el mundo. De modo que la evaluación semántica y el contenido de una creencia están íntimamente relacionados. Si se conoce el contenido (proposicional) de una creencia, entonces se conoce qué es lo que en el mundo determina la evaluación semántica de la creencia, y (b) son causalmente eficaces en la predicción de la conducta. De esta manera, según Fodor, el rasgo característico de la psicología de sentido común es atribuir contenido y poderes causales a las mismas entidades mentales que considera semánticamente evaluables (las creencias). Y además, es característico que las relaciones causales entre las actitudes proposicionales respeten las relaciones de contenido. Más aún, las explicaciones de *mindreading* se basan en esto. Así, dado que para la psicología de sentido común los estados mentales están conectados, semántica y causalmente, la psicología científica que reivindique la psicología de sentido común debe poder dar cuenta de estos rasgos. Según Fodor (1987), la TRM es la teoría que muestra cómo se podría tener una ciencia cuya ontología reconozca estados que exhiben las propiedades (a) y (b), que la psicología de sentido común atribuye a las actitudes proposicionales.

Según la TRM, los estados mentales son tipos de relaciones con representaciones mentales (Fodor 1998). En este sentido, tener una determinada actitud proposicional es estar en una relación determinada con una representación. Más precisamente, para cada evento que consiste en que el organismo posea una actitud proposicional con el contenido *P* (del tipo Ana cree que *P* en el tiempo *t*) hay un evento correspondiente que consiste en que el organismo se relaciona, de un modo característico, con una instancia de representación mental que tiene el contenido *P* (Fodor 1998: 8). Según Fodor, el modo característico de estar en relación con una representación mental puede entenderse como que, para cada evento de creer que *P*, hay un evento correspondiente de tener una representación mental en “la caja de

creencias", que significa que *P*. La "caja de creencias" no alude a cajas literalmente o a algo que tenga interior sino al rol funcional/causal de las actitudes. A pesar de esta especificación, sin embargo, no queda del todo claro en qué consiste una actitud. Un sentido en que esto puede entenderse es que la representación mental está localizada en un lugar apropiado de la organización funcional de modo tal de generar y procesar creencias (Crane 1990: 189).

En relación a los procesos mentales, la TRM afirma que son secuencias causales de casos de representaciones mentales (Fodor 1987). Esto es, cuando alguien tiene en su cabeza una secuencia de creencias, por ejemplo, la creencia de "si *p* entonces *q*" más la creencia de "que *p*", que conducen a creer "que *q*", lo que sucede en su cabeza es la secuencia causal de las siguientes representaciones mentales: "si *p* entonces *q*", "*p*" y "*q*". Así, las representaciones mentales funcionan como los objetos inmediatos de las actitudes proposicionales y como los dominios de los procesos mentales (Fodor 1987).

Hasta aquí me he referido al modo en que, según Fodor, la TRM concibe los estados y procesos mentales. Ahora bien, el modo como la TRM trata a las actitudes es el que permite dar cuenta de la relación de simetría, asumida por la psicología de sentido común, entre las relaciones causales entre los estados mentales y las relaciones semánticas entre las proposiciones. Para dar cuenta de esta propiedad, Fodor combina la postulación de representaciones mentales con la idea de computación a la Turing (Turing 1950) en términos de un lenguaje del pensamiento.

Según Fodor, la TRM postula un lenguaje del pensamiento (LdP), que consiste en un conjunto finito de símbolos o representaciones mentales que tienen propiedades semánticas y causales que están vinculadas por sus propiedades sintácticas. Este vínculo tiene lugar de la siguiente manera. En el LdP, la sintaxis es equiparable a la forma del símbolo (Fodor 1987: 18) en el sentido de que, la sintaxis superviene a la forma. Esto es, no habrá diferencia en la sintaxis sin diferencia en la forma (Crane

1990: 195)[24]. A su vez, la forma es un potencial determinante del rol causal del símbolo, ya que la causalidad tiene que ser vía propiedades intrínsecas y la forma constituye una propiedad intrínseca del símbolo. De este modo, según LdP, un símbolo tiene poderes causales en virtud de su forma/sintaxis.

Una vez relacionados la sintaxis y los poderes causales, resta mostrar cómo se vinculan las propiedades sintácticas y semánticas del símbolo. Tal como señalan los hallazgos de la lógica simbólica moderna, la mayor parte del razonamiento deductivo puede ser formalizado. Esto es, la mayor parte de las relaciones semánticas entre símbolos pueden ser capturadas por las propiedades sintáctico/formales de los símbolos y de las relaciones entre los mismos (Aydede 2010). De este modo, si la mayoría de las relaciones semánticas que tienen lugar entre los símbolos pueden ser imitadas por sus relaciones sintácticas, quedan conectadas las propiedades semánticas y sintácticas. Finalmente, las propiedades sintácticas, que son las que poseen los roles causales, tal como se verá a continuación, establecen el vínculo entre las propiedades causales y las semánticas.

Según la idea de la computación a la Turing, la computación consiste en la manipulación formal de símbolos no interpretados de acuerdo a algoritmos. Estos algoritmos constituyen las instrucciones necesarias para generar un resultado mediante un número finito de pasos. Según Fodor, al combinar esta idea con el LdP se logra la siguiente concepción de los procesos cognitivos. Un proceso cognitivo es un proceso computacional que implica la manipulación de cadenas de símbolos interpretables semánticamente, que son procesados de acuerdo a algoritmos. Esta combinación provee un medio en el cual las operaciones consisten en transformaciones de símbolos y donde el sistema sólo es sensible a las propiedades sintácticas de los símbolos. De este modo, las operaciones están dedicadas a alterar la

[24] Contrariamente, en los lenguajes naturales la sintaxis no superviene a la forma. En este sentido, dos oraciones pueden tener la misma sintaxis pero distinta forma (considérese la diferencia entre “Al papa le gusta el pescado” y “AL PAPA LE GUSTA EL PESCADO”). Es más, una oración puede tener una misma forma y dos análisis sintácticos correctos (considérese “Ana vio un tigre con binoculares”). Sin embargo, como señala Crane (1990), en el caso del pensamiento no hay ambigüedad, puesto que el modo de desambiguar una oración del lenguaje natural consiste, justamente, en establecer a qué pensamientos está asociada.

forma de los símbolos. A su vez, si el sistema está diseñado para transformar símbolos sólo si las proposiciones expresadas por los mismos guardan ciertas relaciones semánticas, éste resulta ser el medio en el que la sintaxis de un símbolo determina su rol causal de manera que respeta el contenido. De este modo, Fodor propone una solución al problema de la mediación entre las propiedades causales y las propiedades semánticas de los símbolos, a través de las propiedades sintácticas. Los sistemas físicos aludidos son las computadoras, y permiten explicar cómo los estados mentales pueden ser intencionales, tal como afirma la psicología de sentido común.

En contraposición con la analogía lingüística, esta analogía con la ciencia no presenta dificultades en relación al elemento conceptual. Como he mencionado, la psicología de sentido común se constituye como una psicología de actitudes proposicionales y la TRM proporciona un enfoque sobre las mismas como representaciones con contenido conceptual. Según la TRM, el contenido proposicional de las actitudes proposicionales está conformado por representaciones mentales (liguaformes) que constituyen el vehículo del contenido semántico y, a la vez, son las portadoras de los poderes causales. Los conceptos serían representaciones del LdP, esto es particulares mentales. Por todo lo mencionado, en principio, esta analogía con la ciencia parece proveer un sentido de teoría subyacente a la habilidad de *mindreading* más relevante, al menos, en tanto da cuenta de la posesión de conceptos mentalistas.

5. Conclusión

Las distintas versiones de la TT proponen una respuesta a la cuestión de los procesos subyacentes a *mindreading*. Si bien los cuatro enfoques mencionados, el funcionalismo filosófico, la TT genérica, el enfoque del niño científico y el enfoque nativista, afirman que un cuerpo de información subyace a *mindreading*, existen diferencias importantes entre los mismos. Éstas se pueden ordenar alrededor de tres ejes: el compromiso conceptual con la presencia de conceptos y enunciados

legaliformes, el tipo de adquisición de *mindreading* (aprendida o innata) y el tipo de recurso o mecanismo cognitivo postulado (de propósito general o específico).

Los enfoques provenientes de la filosofía, tal como se mencionó, son el funcionalismo filosófico y la TT genérica o inclusiva. La principal diferencia entre los mismos reside en los compromisos conceptuales. La TT genérica adopta una noción amplia de "teoría" representada internamente que no está comprometida con la existencia de una conexión legaliforme entre los elementos conceptuales que conforman el cuerpo de conocimiento psicológico. Contrariamente, el funcionalismo filosófico está comprometido con la presencia de generalizaciones legaliformes. La TT genérica asume que la cuestión de la presencia o no de generalizaciones legaliformes en el cuerpo de conocimiento psicológico interno subyacente a *mindreading* se definirá empíricamente.

En concordancia con la ciencia cognitiva, la TT genérica asume que la información utilizada por la competencia mentalista es, en su mayor parte, inaccesible a la conciencia. De este modo, este enfoque no está comprometido con la idea de que el cuerpo de conocimiento representado internamente está conformado por perogrulladas sobre la mente, conocidas por todas las personas. En este respecto, se distingue del funcionalismo filosófico de sentido común. Por su parte, el enfoque del niño científico también adhiere a la estrategia dominante en ciencia cognitiva y postula una teoría tácita subyacente. En este aspecto se asemeja a la TT genérica aunque se diferencia de la misma por ser un enfoque comprometido fuertemente con la presencia de generalizaciones legaliformes.

Si bien el funcionalismo filosófico es un enfoque empírico, en rigor, no es estrictamente un enfoque sobre los procesos cognitivos que subyacen a la competencia mentalista. Sin embargo, la propuesta clave del funcionalismo filosófico de considerar los términos mentales como los términos teóricos de una teoría que aluden a las entidades inobservables que intervienen en la explicación causal de lo observable, es de suma importancia en los enfoques provenientes de la psicología cognitiva. Particularmente, el enfoque del niño científico asume que *mindreading*

invoca entidades inobservables y que estos constructos teóricos están vinculados entre sí de modo legaliforme (Gopnik & wellman 1995:233; Wellman 1990: 6-11, 325), en una marcada concordancia con el funcionalismo filosófico.

No obstante, el enfoque del niño científico es más bien un enfoque sobre la adquisición de la competencia mentalista. En principio, el funcionalismo filosófico no se pronuncia respecto de la adquisición, aunque si hubiera algún compromiso sería con la adquisición de principios legaliformes. Por su parte, la TT genérica no se encuentra comprometida con alguna forma particular de adquisición. En este sentido, si el enfoque del niño científico está en lo cierto empíricamente, la TT genérica no colapsaría, ya que no excluye la posibilidad de generalizaciones legaliformes presentes en el cuerpo de conocimiento psicológico interno.

Respecto al tipo de recursos cognitivos, el enfoque del niño científico postula recursos cognitivos de tipo general. Según este enfoque, los niños utilizan capacidades conceptuales generales para la formación de la teoría subyacente a *mindreading*. Contrariamente, el enfoque nativista/modular propone la utilización de recursos cognitivos específicos de dominio. Particularmente, este enfoque postula un mecanismo específico de TdM subyacente a *mindreading*, que pone a disposición del niño pequeño conceptos mentalistas innatos (creencia, deseo, ficción) antes que haya desarrollado otros conceptos mediante habilidades conceptuales generales. Más allá de las objeciones, este mecanismo se concibe como un módulo que permite atender a los comportamientos y computa los estados mentales que contribuyen a estos. En el caso de la TT genérica o inclusiva, también estamos en presencia de mecanismos cognitivos específicos, en tanto, el cuerpo de conocimiento psicológico representado internamente guía la competencia mentalista de manera análoga a como la gramática guía la competencia lingüística, según la propuesta chomskiana de 1965.

Dado que existen diferentes versiones de la TT es necesario establecer algunos compromisos mínimos para caracterizarla de modo paradigmático. Así, consideraré como las tesis básicas de la TT, que luego retomaré para analizar los enfoques híbridos, en los capítulos 4, 5 y 6, las siguientes afirmaciones:

1. Las atribuciones de estados mentales se llevan a cabo utilizando un cuerpo de conocimiento sobre lo mental, y las predicciones y explicaciones de la conducta se realizan en base a inferencias deductivas o subsunciones de los casos particulares bajo principios generales.
2. El cuerpo de conocimiento acerca de lo mental está compuesto por un conjunto de conceptos mentales y un conjunto de principios o generalizaciones que ponen en relación tales conceptos.
3. Ciertos procesos o mecanismos cognitivos utilizan este cuerpo de conocimiento para llevar a cabo las atribuciones de estados mentales, las predicciones y las explicaciones de la conducta y el pensamiento en términos mentalistas.
4. El cuerpo de conocimiento sobre lo mental es una psicología de actitudes proposicionales.

La tesis 1 alude a la asunción general de la TT de que un cuerpo de conocimiento subyace a la capacidad de atribución mentalista, así como a la explicación y predicción del comportamiento en términos mentalistas (*mindreading*). La tesis 2 señala la composición del cuerpo de conocimiento sobre lo mental: conceptos mentalistas y principios que los relacionan. Principalmente, la TS surge en oposición a esta tesis. O, más específicamente, la TS postula que no es necesario recurrir a generalizaciones sobre lo mental para llevar a cabo *mindreading*. La postura de la TS en relación a la utilización de conceptos mentales es más compleja y la desarrollaré en el capítulo 2. Según la tesis 3, habría mecanismos inferenciales que manipularían ese cuerpo de conocimiento, sin comprometerse con que sean generales o específicos de dominio.

La tesis 4 se refiere a que la TT asume una psicología de actitudes proposicionales. En relación a ésta tesis, existe una objeción que sólo he mencionado de manera tangencial en la sección 1 de este capítulo. Se ha señalado que la concepción de la psicología de sentido común como una psicología de actitudes proposicionales es estrecha, puesto que el repertorio de estados mentales que

interviene en el entendimiento y la coordinación social es más amplio que el de las actitudes proposicionales, e incluye otros estados psicológicos tales como las emociones, las sensaciones, los sentimientos, los dolores y demás. En este sentido, se ha sugerido que un enfoque completo de *mindreading* tendría que poder dar cuenta también de otros tipos de estados mentales tales como las sensaciones y emociones (Goldman 2006: 20). En principio, el alcance de una psicología de actitudes proposicionales es acotado en este respecto. No obstante, los enfoques que asumen una psicología de actitudes proposicionales, al menos, dan cuenta de aquella porción de las habilidades sociales relacionadas con las actitudes proposicionales. En este sentido, los enfoques puros de teoría y simulación, en tanto asumen la psicología de actitudes proposicionales, pueden considerarse explicaciones parciales de las habilidades sociales.

CAPÍTULO 2. EL ENFOQUE DE LA "TEORÍA DE LA SIMULACIÓN"

En este capítulo caracterizaré el enfoque de la "teoría de la simulación". Este enfoque asume que no es necesario recurrir a conocimiento psicológico alguno para llevar a cabo la atribución de estados mentales, y la explicación y la predicción de la acción y el pensamiento en términos mentalistas, sino que, la mayor parte de las veces, basta con la simulación mental[25]. Como he mencionado en la introducción, las distintas versiones de la TS difieren ampliamente respecto del modo en que se articula la simulación mental, y las propuestas resultan difíciles de reconciliar entre sí. Aquí, no intentaré reconciliar las diferentes versiones sino sólo me ocuparé de aquellas que se formulan en términos de procesos cognitivos subyacentes a la capacidad de *mindreading*[26]. Como también mencioné en la introducción, me interesa evaluar si los enfoques híbridos de teoría y simulación logran brindar una descripción y una explicación satisfactoria de los procesos subyacentes a la capacidad de *mindreading*. En este sentido, me ocuparé particularmente de aquellas versiones de la TS que se retoman en la formulación de las propuestas híbridas de teoría y simulación: la propuesta simulacional pura de Goldman (1993b, 1993b, 1995a, 1995b, 2000) y la caracterización de la simulación "*off-line*" por parte de los teóricos de la teoría (Stich & Nichols 1992). Este último enfoque tiene un rol importante en los enfoques híbridos de *mindreading*. Por un lado, Goldman (2006: 20) adhiere explícitamente a esta caracterización científico-cognitiva de la simulación. Por el otro, Stich & Nichols (2003)

[25] Cuando los "teóricos de la simulación" sostienen que no es necesario recurrir a conocimiento psicológico para llevar a cabo *mindreading*, usan el término "conocimiento" en un sentido débil y laxo, sin entender por el mismo "creencia verdadera justificada". Además, cuando aluden a este aspecto de la simulación suelen utilizar de manera intercambiable una serie de términos que no son sinónimos, tales como "conocimiento", "generalizaciones", "teoría", "cuerpo de información", "cuerpo de conocimiento", "información". Siguiendo a los "teóricos de la simulación", utilizaré los términos de esta misma manera.

[26] Otras versiones de la simulación llevan a cabo propuestas muy diferentes y hasta irreconciliables con las propuestas empíricas de procesos cognitivos subyacentes a *mindreading*. Gordon (1986) ha propuesto a las "rutinas de ascenso", formuladas por Evans (1982), como el método para llevar a cabo la simulación mentalista, sin recurrir a conceptos, teorías. Por su parte, Heal (1995a, 1995b) propone que, *a priori,* la capacidad intelectual de pensar acerca de lo posible puede utilizarse para obtener un *insight* sobre las otras personas, y en esto consiste la estrategia imaginativa de replicación (o simulación).

adhieren a una noción de simulación que desarrollan a partir de esta caracterización que va a tener impacto en su propuesta híbrida (Nichols & Stich 2003).

En la sección 1 me ocuparé de algunas cuestiones generales de la propuesta de simulación mental. Considero que es preciso realizar algunas aclaraciones respecto de la noción de simulación, que ha sido utilizada con distintos sentidos por los teóricos de la simulación y que, a su vez, no debe confundirse con su uso en el ámbito de las ciencias de la computación. En la sección 2, reconstruiré la propuesta de un proceso de simulación puro subyacente a *mindreading* (Goldman1993b, 1995a; Gallese & Goldman 1998), que tiene varios puntos de conexión con la propuesta híbrida (Goldman 2006, 2009) de la que me ocuparé en el capítulo 5. En la sección 3, analizaré la propuesta de simulación *off-line,* formulada por defensores de la TT como una caracterización de la simulación que reúne los puntos principales de las propuestas de Gordon (1995a) y Goldman (1995a), junto con algunas especificaciones (Stich & Nichols 1992, 1995, 1996, 1997, Nichols *et al.* 1996). Llamativamente, esta caracterización, que ha sido aceptada y apropiada por algunos teóricos de la simulación (Gallese & Goldman 1998; Goldman 2006), se ha convertido en la caracterización estándar de este enfoque en el ámbito de las ciencias cognitivas. Sin embargo, Stich & Nichols (1992, 1997, 2003) consideran que es implausible que la simulación *off-line* subyazca a *mindreading*. No obstante, en la sección 3.2, me ocuparé de cierto tipo de proceso simulacional más plausible para dar cuenta de ciertas atribuciones a otras personas, según estos autores. Finalmente, en la sección 4 de conclusiones señalaré las tesis básicas de TS que luego serán retomadas para el análisis de los modelos híbridos.

1. La simulación mental

La explicación usual en ciencias cognitivas recurre a la postulación de cuerpos de información subyacentes a las competencias cognitivas, como señalo en el capítulo 1 (sección 3.1) (Nichols & Stich 1992; Davies & Stone 1995a). En concordancia con este tipo de explicación, la TT postula un cuerpo de información psicológico representado

internamente que guía las capacidades de descripción, predicción y explicación del comportamiento en términos mentalistas, que conforman la competencia cognitiva para *mindreading*. Contrariamente, el enfoque de la TS asume que no es necesario emplear conocimiento psicológico alguno al ejecutar tales capacidades, sino que basta con la simulación mental para que tenga lugar la comprensión psicológica cotidiana de las otras personas. De este modo, la TS se establece como una desviación de la explicación usual en ciencias cognitivas y, en este sentido, se constituye como una alternativa al enfoque de TT. De este enfrentamiento surge el debate teoría-simulación.

En principio, el surgimiento de la simulación tiene, al menos, dos consecuencias inmediatas. Por un lado, si el enfoque de TS constituye una alternativa explicativa auténtica, la estrategia explicativa usual en ciencias cognitivas dejaría de serlo porque, al menos, algunas competencias cognitivas se explican de otro modo (Stich & Nichols 1992). Por otro lado, el surgimiento de este enfoque supone una consecuencia para el enfoque eliminativista de los estados mentales (Churchland, 1988). Si la TS está en lo cierto y no hay ninguna teoría subyacente a la comprensión mentalista, el eliminativismo deja de tener asidero puesto que no existe una teoría psicológica que pueda resultar falsa (Bermúdez 2005b).

No obstante, el enfoque de la simulación presenta algunas dificultades. La principal dificultad reside en que sus defensores recurren a diferentes nociones de simulación y no es claro que éstas puedan reconciliarse entre sí, ni que el término "simulación" recoja una categoría interesante[27]. A modo de simplificación, sin embargo, puede afirmarse que todas las versiones comparten, en mayor o en menor medida, la idea de que los mismos recursos mentales que son utilizados en nuestro propio pensamiento, toma de decisiones o respuestas emocionales son reutilizados para proveer el entendimiento de los pensamientos, las decisiones y las respuestas

[27] Una lista de los distintos sentidos en que se ha aplicado el término "simulación" en la literatura de *mindreading* puede encontrarse en Nichols & Stich (2003: 132-134). No obstante, algunos consideran que se trata de un término muy arraigado en la literatura y que no sería útil deshacerse del mismo. Particularmente, Goldman (2006) propone afinar el término con el objetivo de mostrar que hay cierta unidad en la categoría.

emocionales ajenas. Al parecer este solapamiento nos ahorra la necesidad de poseer conocimiento antecedente sobre cómo funcionan típicamente los procesos psicológicos.

Asimismo, resulta conveniente hacer cierta salvedad respecto de la noción de simulación utilizada en la literatura de *mindreading*. Esta noción debe distinguirse de otros usos de la noción proveniente de la ciencia de la computación, que se asocia a la inteligencia artificial y, particularmente, a la modelización en ciencia. En la simulación computacional se intenta predecir el comportamiento de un sistema utilizando un modelo computacional del mismo. Tales modelos pueden ser caracterizados por ecuaciones matemáticas o reglas. Justamente, estos modelos suponen cuerpos de información sobre el sistema a ser simulado, por ejemplo, la simulación computacional de un huracán se lleva a cabo mediante un programa computacional que tiene información sobre las leyes de aerodinámica e hidrodinámica que gobiernan los huracanes (Haugeland 1985). En este sentido, este tipo de simulaciones están basadas en información. Contrariamente, el enfoque de la simulación mental postula la posibilidad de realizar predicciones del comportamiento de un sistema sin recurrir a cuerpos de información sobre su funcionamiento.

El sentido con que se utiliza la noción particular de "simulación mental" en la literatura de *mindreading*, surge de un análisis preteórico de la noción de simulación (Stich & Nichols 1992; Goldman 1995a; Davies & Stone 1998, 2000). Este desarrollo conduce a la distinción entre simulación conducida por proceso o por teoría (Goldman 1995a: 85)[28]. Al parecer, la concepción detrás de la simulación mental es que para

[28] Goldman (1995a) propone la distinción entre simulación conducida por teoría o por proceso en respuesta a la objeción que sugiere que llevar adelante una simulación requiere necesariamente de teoría (Dennett 1987). Si uno se imagina, por ejemplo, ser un puente colgante y quiere averiguar cómo se comportará si sopla viento, la idea que uno pueda hacerse va a depender críticamente de sus conocimientos sobre la ingeniería y la física de los puentes colgantes (Dennett 1987). Sin embargo, éste es un modo posible de llevar adelante una simulación. En otras ocasiones, por ejemplo, cuando el dispositivo que simula un determinado sistema mantiene un isomorfismo relevante con el comportamiento del sistema a ser simulado, la teoría no parece ser necesaria (Goldman 1995a). Más precisamente, este tipo de simulación requiere el cumplimiento de dos condiciones. Primero, que el proceso que conduce a la simulación en el dispositivo debe ser el mismo proceso, o un proceso relevantemente similar, al que conduce al sistema. Segundo, que los estados iniciales del dispositivo simulador deben ser los mismos estados, o estados relevantemente similares, a los estados del sistema

predecir el comportamiento de un objeto cualquiera disponemos de, al menos, tres modos de realizar predicciones: que las mismas pueden estar basadas en una teoría o en una simulación y, a su vez, la simulación puede llevarse a cabo en la realidad o en la imaginación (Davies & Stone 2000).

Así, la predicción del comportamiento de un objeto, por ejemplo, de un avión sometido a ciertas condiciones de viento, puede realizarse recurriendo a una teoría pertinente (la teoría aerodinámica) que guíe los cálculos correspondientes. Estos, pueden llevarse a cabo de manera consciente y esforzada, consultando un manual y utilizando papel y lápiz. O bien, puede ser el caso que la teoría esté tan memorizada e internalizada, que los cálculos se realicen mentalmente (Stich & Nichols 1992). No obstante, en ambos casos, la teoría tiene un rol esencial para la predicción del comportamiento del avión. Se trata de una predicción basada completamente en una teoría.

Asimismo, es posible predecir el comportamiento de un avión construyendo un objeto similar a éste, un modelo cuyo comportamiento se pone a prueba en circunstancias similares a las que nos interesan, por ejemplo, en un túnel de viento. En este caso, si bien es preciso recurrir a la teoría para determinar qué características se quieren replicar en el modelo, en la simulación misma tiene lugar un proceso relevantemente similar al comportamiento que se busca simular. Por este motivo, en este caso la predicción consiste en inferir el comportamiento del avión a partir de un proceso que tiene lugar en la realidad, a saber, el comportamiento del modelo en el túnel de viento. En este sentido, se trata de una simulación en la realidad "conducida por proceso", que permite predecir el comportamiento de un objeto.

Ahora bien, este mismo tipo de simulación se puede llevar a cabo completamente en la imaginación. Por ejemplo, puedo imaginarme un modelo del

---

simulado. Si se cumplen estas condiciones, entonces se trata de una simulación conducida por proceso. En este sentido, cuando una persona simula una secuencia de estados mentales de otra persona, puesto que comparten un mismo sistema cognitivo, si los estados iniciales son relevantemente similares, es esperable que los estados finales también resulten relevantemente similares. Así, se distingue entre simulación conducida por teoría o conducida por proceso (Goldman 1995a). Particularmente, la teoría de la simulación se distingue por postular que la simulación mental en la imaginación puede ser conducida por proceso.

avión con tales y cuales características, y un túnel de viento que lo somete a unas circunstancias de interés. Sin embargo, para que esto sea posible se requiere la posesión de información sobre cómo suele comportase un avión con ciertas características, en determinadas circunstancias. En este sentido, la predicción del comportamiento dependerá completamente del conocimiento antecedente que poseamos al respecto y así, el rol de la simulación se vuelve insignificante. Dado que, en este caso, la teoría tiene un rol preponderante para realizar la predicción, se trata de una simulación en la imaginación conducida por teoría. En este sentido, este caso no constituye una alternativa a las predicciones basadas en teoría.

Estos tres modos de realizar predicciones pueden aplicarse a la predicción del comportamiento de otras personas. Supongamos, por ejemplo, que queremos predecir el comportamiento de cierta figura política en ascenso cuando alguien con autoridad le pide, por ejemplo, que administre descargas eléctricas a una persona amarrada a una silla en la habitación contigua (Stich & Nichols 1992). Un modo posible de llevar a cabo esta predicción consiste en recolectar información sobre la historia y la personalidad de la persona en cuestión y luego, consultar la mejor teoría sobre los determinantes del comportamiento en tales circunstancias. O bien, se pueden reclutar sujetos con rasgos de personalidad similares a los del agente de interés y someterlos a la prueba de Milgram (1963). En el primer caso, la predicción se obtendrá a partir de una teoría y, en el segundo caso, a partir de una simulación en la realidad conducida por proceso. No obstante, una tercera alternativa es posible. Simplemente podemos imaginar a los sujetos siendo sometidos a la prueba de Milgram. En este caso, la predicción dependerá absolutamente del conocimiento precedente que tengamos de la prueba y del comportamiento que es usual en las personas con ciertos rasgos de personalidad. Es más, usualmente este tipo de cálculos y anticipaciones se utilizan para idear experimentos psicológicos. En este último caso, la predicción se basa en una simulación en la imaginación conducida por teoría.

No obstante, es preciso destacar que no toda simulación en la imaginación es asimilable a una simulación conducida por teoría. Precisamente, el enfoque de TS

asume que existe un cuarto modo de realizar predicciones del comportamiento de los seres humanos, a saber, mediante una simulación en la imaginación conducida por proceso. Para sostener esto, el enfoque simulacional asume una serie de supuestos que sustentan la postulación de la simulación mental:

1. Los seres humanos compartimos el mismo sistema cognitivo, es decir, no existen diferencias relevantes entre los sistemas cognitivos.
2. Algunos procesos cognitivos funcionan normalmente cuando son reutilizados en la simulación, al menos, éste es el caso para el razonamiento teórico y práctico.
3. Es posible determinar los estados mentales iniciales de las otras personas.

De manera trivial, la TS asume que los seres humanos compartimos el mismo sistema cognitivo. Además, se asume que algunos procesos o sistemas cognitivos operan del mismo modo cuando realizan sus funciones propias y cuando son re-utilizados para proveer el entendimiento de los pensamientos, las decisiones y las respuestas emocionales ajenas. Especialmente, esto se considera plausible respecto del funcionamiento del razonamiento práctico y del razonamiento teórico. Dados 1 y 2, si, además, estos sistemas son alimentados con *inputs* (creencias y deseos) relevantemente similares a los del agente blanco (por 3), obtendremos resultados similares a los que ocurren en la mente del agente blanco. Es en este sentido que debe entenderse la afirmación de que la simulación mental brinda un entendimiento del comportamiento ajeno a partir del funcionamiento cognitivo propio, que es utilizado como "modelo" del funcionamiento cognitivo ajeno. De este modo, la simulación mental se constituye como una alternativa a la predicción basada en teoría en la medida en que en este procedimiento no interviene ningún tipo de conocimiento del funcionamiento psicológico típico, sino que se trata de una simulación conducida por proceso (Goldman 1995a).

2. La simulación pura

En *Interpretation Psychologized*, Goldman defiende un enfoque de simulación para dar cuenta de cómo se las ingenia el intérprete ingenuo para llevar a cabo sus juicios sobre las actitudes proposicionales de las otras personas y hace hincapié en que los filósofos que se han ocupado de contestar esta pregunta "… no han sido suficientemente psicológicos, o cognitivistas, ni siquiera aquellos que tienen una inclinación por la psicología" (1995a: 74). Contrariamente, su modo de afrontar el problema lo construye deliberadamente como una pregunta empírica sobre la psicología de los que realizan atribuciones: ¿Qué es lo que de hecho sucede en las cabezas de los que realizan atribuciones que da cuenta de sus atribuciones?...Deberíamos tratar de identificar los procesos cognitivos implicados en esta tarea" (1995b: 186). Así, la simulación se concibe como el proceso primordial de atribución de estados mentales a otras personas.

En principio, esta propuesta comparte el supuesto mencionado en la introducción, común a las distintas versiones de la simulación, de la similitud psicológica entre las personas. Al menos, en el sentido mínimo de que los sistemas cognitivos operan según el mismo conjunto de restricciones psicológicas, por ejemplo, los mismos procesos de formación de creencias. No obstante, este enfoque se distingue del resto de las versiones de la simulación en tanto postula que las atribuciones a otras personas están basadas en el entendimiento previo de uno mismo y, en este sentido, la atribución a otras personas se vuelve dependiente de la autoatribución. En particular, se asume que la atribución de estados mentales a uno mismo está mediada por conceptos mentales, y esto otorga un rol central a tales conceptos. Este rol consiste, en primer lugar, en la necesaria identificación y clasificación de los estados mentales para poder establecer los parámetros de una simulación, esto es, los estados iniciales del blanco y, en segundo lugar, en la identificación y categorización de los estados mentales propios al finalizar una

simulación, antes de adscribírselos al blanco. De este modo, en la medida en que los conceptos mentales son esenciales en la autoatribución, tienen un lugar central en la simulación[29].

No obstante, ningún enfoque de simulación es adecuado para ofrecer una teoría de la naturaleza de los conceptos mentales y, en este punto, este enfoque tendrá que recurrir al auxilio de alguna otra teoría. En particular, esta propuesta se inclina por la introspección y la idea de que los conceptos están determinados por sus aspectos cualitativos, este aspecto será desarrollado en la sección 2.2. Antes, en la sección 2.1, me ocuparé de la simulación como el proceso subyacente a la atribución de estados mentales ajenos y de la relación estrecha de este proceso con la autoatribución.

2.1. La atribución de estados mentales a otras personas

Según esta propuesta, las personas "…adscriben estados mentales a otras personas fingiendo o imaginándose a ellos mismos en los zapatos del otro, construyendo o generando los estados (subsiguientes) en los que se encontrarían, y adscribiendo tales estados al otro. En síntesis, *simulamos* la situación de los otros, y los interpretamos de acuerdo a ésta" (Goldman 1995a: 81). En este sentido, la esencia de la simulación consiste en reproducir o replicar en uno mismo un evento o secuencia de eventos que ocurren, o se piensa que ocurren, en otra persona con el propósito de

[29] No todas las versiones de la simulación otorgan un rol a los conceptos mentales. El enfoque ofrecido por Gordon (1995a, 1995d, 1996) niega la intervención de conceptos mentales en la simulación. Gordon propone las rutinas de ascenso (Evans 1982) como un método para dar cuenta de las atribuciones sin hacer uso de la introspección, ni de los conceptos, ni de teoría (capítulo 1, sección 4). Específicamente, las rutinas de ascenso consisten en el procedimiento de responder a una pregunta sobre los estados mentales de uno mismo ¿vos crees que Neptuno tiene anillos?, respondiendo a una pregunta de nivel semántico inferior ¿Neptuno tiene anillos? Esta última pregunta es acerca de los hechos y su respuesta está basada en la evaluación de los mismos sin preguntar acerca de uno mismo, ni acerca de los estados mentales. Este modo de adscripción de estados mentales, basado en la capacidad de expresar estados mentales, se extiende a la simulación de estados mentales ajenos (Gordon 2009). Así, en el caso de la simulación, se responde a la pregunta de si el blanco cree que *p* preguntando, simplemente, si es el caso que *p*. Y, reportando qué hay ahí, está reportando las creencias del blanco. Así, reportar las creencias del blanco consiste sólo en reportar lo que está ahí en el mundo (Gordon 1995d: 60).

generar predicciones y explicaciones del comportamiento ajeno (Goldman 2000: 186). Así, se atribuyen estados mentales a otras personas mediante la adopción de la perspectiva de estos para observar, luego, qué otros estados surgen a partir de ubicarse en tal posición. Es preciso mencionar que uno de los argumentos esgrimidos para afirmar la plausibilidad de la simulación apela a la experiencia subjetiva de que en ocasiones uno puede ponerse en el lugar del otro con el objetivo de anticipar o comprender su comportamiento.

En particular, la simulación puede utilizarse para predecir los estados mentales de otros o su comportamiento. Por ejemplo, en un juego de ajedrez un jugador puede anticiparse a la jugada del otro imaginándose a sí mismo en la situación de su contrincante y decidiendo qué elegir hacer desde esta posición. Asimismo, la simulación puede utilizarse para la explicación del comportamiento. Este uso explicativo, o "retrodictivo", tiene lugar generalmente cuando el punto de partida es el comportamiento ajeno observado y el simulador se remonta hacia atrás para determinar los estados mentales que le dieron lugar. Por ejemplo, el simulador puede preguntarse por el objetivo del blanco para llevar a cabo la acción *X*. En este caso, el simulador conjetura el objetivo *Y*, y genera el objetivo ficticio *Y* que se utilizará para poner en marcha la simulación. Se asume que, en base a esto, la simulación genera la decisión de realizar *X* y que el simulador usará este resultado para concluir que el objetivo del blanco es *Y* (Gallese & Goldman 1998).

Ahora bien, a mi entender, es preciso notar que, en todos estos casos la inferencia de un estado mental ajeno supone que la otra persona posee ciertos estados mentales previos. En este sentido, colocarse en la situación del otro mediante la adopción de sus deseos y creencias, para observar qué otros estados suceden luego, implica suponer estados mentales previos de los que se sigue el estado de interés. Según esta propuesta, para determinar estos estados mentales, se toma como punto de partida la situación perceptual del otro y, a partir de esto, se infieren ciertas experiencias perceptuales o creencias, las mismas que uno tendría en su situación puesto que la simulación asume la similitud entre los sistemas cognitivos en la medida

en que todos están sometidos a las mismas restricciones psicológicas, por ejemplo los mismos procesos formadores de creencias para determinar los estados mentales. Además de esta similitud, este enfoque asume que todas las personas tienen las mismas necesidades básicas de comida, amor, calor y demás, salvo que se tenga información contraria (Goldman 1995a: 82). No obstante, este procedimiento no se considera infalible y, en este sentido, las fallas al tratar de establecer los estados previos del blanco se consideran una de las fuentes de error en la simulación.

Este enfoque contempla el uso de la simulación atendiendo a las diferencias relevantes entre individuos, y llevando a cabo ajustes en relación a las destrezas y conocimiento que poseen los otros para establecer los parámetros de la simulación. Sin embargo, no siempre es posible llevar a cabo tales ajustes y, usualmente, no se tiene en cuenta esta información para llevar a cabo una simulación. En principio, esto no se considera una desventaja del enfoque dado que no se asume que seamos simuladores óptimos, ni que la simulación sea el único método para la comprensión y la coordinación sociales. En este sentido, se acepta que ciertas regularidades en el comportamiento así como algunas diferencias interindividuales puedan ser aprendidas inductivamente (Goldman 1995a: 83). Por ejemplo, si las personas que entran en un auto y se sientan en el asiento del conductor tienden, usualmente, a poner el auto en marcha, esta información pude usarse de base para generar expectativas sobre el comportamiento de las personas, sin que sea necesario recurrir a simulación alguna.

De este modo, el primer paso en la simulación consiste en generar en uno mismo estados mentales ficticios (creencias, deseos, preferencias) que surgen de adoptar la perspectiva de la otra persona. Por ejemplo, se generan preferencias entre jugadas de ajedrez. Estas creencias y preferencias ficticias alimentan como *input* al propio sistema de toma de decisiones, cuyo *output* será una decisión ficticia, que se utiliza como base para predecir la conducta del blanco (Gallese & Goldman 1998). Ahora bien, dado que los *inputs* ficticios pueden alimentar al sistema de toma de decisiones tal como lo hacen los *inputs* propiamente dichos, la propuesta simulacional asume que los estados mentales ficticios guardan cierta similitud relevante con los

estados mentales propiamente dichos. Así, surge la cuestión acerca de en qué medida los estados mentales ficticios son similares a los estados mentales no ficticios.

En principio, la similitud y diferencia entre los estados mentales ficticios y los propiamente dichos se considera una cuestión empírica a determinar (Goldman 2000: 186). En relación a esto, la homología entre los estados mentales ficticios y los no ficticios parece encontrar apoyo en la literatura sobre la imaginería visual y motora, en la que se asume que las mismas consisten respectivamente en "fingir que se ve" y "fingir que se hace" (Currie & Ravenscroft 1997). No obstante, la homología visual y motora no permite afirmar nada respecto de si otros estados mentales tales como las creencias y los deseos ficticios, son funcionalmente homólogos respecto de las creencias y los deseos propiamente dichos.

Además, esta propuesta no sólo pretende dar cuenta de estados mentales tales como las creencias y demás actitudes proposicionales sino, también, de otros estados mentales tales como el dolor, las cosquillas o las emociones (Goldman 1995a, 2000). En principio, este enfoque parecería ser adecuado para esto en la medida en que la simulación produce, en la mente del simulador o intérprete, una secuencia de estados que se parece a los estados del blanco de la atribución en aspectos cognitivos cruciales. Al parecer, este tipo de rastreo no sólo permitiría detectar creencias y demás estados doxásticos sino, también, indicadores de sentimientos, comportamientos o síntomas fisiológicos que un rastreo intelectual no podría detectar (Goldman 2000). Por su parte, los defensores de la TT estarían de acuerdo con que la interpretación del blanco implica, en cierta medida, un rastreo de estados mentales, aunque no aceptarían que esto involucre reproducir, en el propio sistema cognitivo, los mismos eventos que ocurren en el sistema del blanco. En este sentido, si bien se acepta que un rastreo implica conjeturas o creencias sobre los estados en los que puede estar el blanco, sin embargo, no se acepta que esto implique "réplicas" de estos estados. En este sentido, este tipo de rastreo no permitirá experimentar estados mentales similares a los del blanco. En principio, esto puede considerarse una limitación del enfoque de TT en relación a su alcance, mientras que la propuesta simulacional parece aventajarlo en

este respecto, en la medida en que puede dar cuenta de la atribución de un rango mayor de estados mentales. Por supuesto, para que esto sea así la simulación tiene que ser un proceso plausible.

Como señalé anteriormente, establecer los parámetros de la simulación implica establecer la situación perceptual del blanco, atender a diferencias interindividuales, suponer la similitud entre los sistemas cognitivos y en relación a las necesidades básicas primordiales. A mi entender, establecer estas cuestiones podría no ser tan simple como sugiere este enfoque. Sin embargo, esto no se cuestiona en la simulación, ni desde la perspectiva del enfoque de la TT, puesto que, según este, también se considera posible establecer los estados iniciales del blanco. En este último caso, la información sobre tales estados iniciales se utiliza junto con el cuerpo de información psicológica, para inferir los estados mentales y explicar y predecir el comportamiento. No obstante, me interesa señalar que en este enfoque simulacional particular resulta esencial rastrear e identificar los estados mentales del blanco para poner en marcha la simulación, y para esto son necesarios los conceptos mentales. Asimismo, también es preciso identificar y clasificar el estado mental generado por simulación antes de su proyección, aunque este aspecto no se desarrolla en detalle hasta la propuesta híbrida (Goldman 2006) de la que me ocuparé en el capítulo 5. De modo que, puesto que es necesario determinar e identificar estados mentales, esta propuesta precisa, según Goldman (2006) la ayuda de algún enfoque sobre la naturaleza de los mismos, ya que ningún enfoque simulacional la proporciona. Además, será preciso que esta teoría pueda dar cuenta satisfactoriamente de la identificación de los estados mentales propios, puesto que este enfoque particular asume que los estados mentales a identificar se generan en uno mismo, ya sea como estados ficticios para iniciar la simulación, ya sea como el producto de la simulación (antes de ser proyectados). A continuación me ocuparé de la propuesta de autoatribución que da cuenta de estos aspectos de la simulación.

2.2. La cuestión de la autoadscripción de estados mentales

Es preciso señalar que el enfoque de simulación no ofrece una respuesta directa a la cuestión de cómo las personas detectan y clasifican sus propios estados mentales ocurrentes. Además, ningún teórico de la simulación ha sostenido que la simulación medie la autoatribución de estados mentales ocurrentes aunque, en principio, puede ser usada para la autoadscripción de estados mentales pasados, hipotéticos o futuros (Goldman 2006: 24). No obstante, la propuesta de Goldman (1993b) otorga un rol central al fenómeno cognitivo de la autoatribución en la medida en que la atribución de tercera persona depende de la de primera persona para establecer los parámetros de la simulación y antes de realizar la proyección a otras personas. Mínimamente, los estados mentales deben etiquetarse como estados del simulador al inicio y al final de la simulación (Goldman 2000: 184). Así, dado el rol central de la autoatribución de estados mentales se requiere que este enfoque de cuenta de este aspecto de las atribuciones. Para esto, se asume la ocurrencia genuina de eventos de conciencia fenoménica caracterizada en términos intrínsecos y no relacionales. En este sentido, un proceso especial interviene en la atribución en primera persona denominado "introspección", "sentido interno" o "auto-monitoreo" (Goldman 1993a; 1993b; 2000). Así, la autoatribución tiene un rol central en un sentido más fundamental, en virtud del compromiso con la concepción de los conceptos mentales como especificados por sus propiedades fenoménicas o cualitativas, *i.e.*, sus *qualia*.

En particular, Goldman adhiere a la introspección porque considera al funcionalismo como un enfoque insatisfactorio para dar cuenta del rol de los conceptos mentales en la autoatribución, por las siguientes razones. Como señalé en el capítulo 1 (sección 2), la propuesta funcionalista implica que las personas entienden cada concepto de un estado mental como captando un rol funcional específico, esto es, un conjunto de relaciones funcionales/causales entre el estímulo o *input*, el *output* conductual u otros estados internos que le siguen y otros estados mentales intermedios. De acuerdo con esto, la tarea de categorizar un estado mental propio requiere, según la propuesta funcionalista de los procesos inferenciales que

intervienen en la autoatribución, decidir qué roles funcionales están instanciados por los propios estados mentales ocurrentes. Sin embargo, según Goldman (1993b), si se analiza el modo en que se lleva a cabo la tarea de reconocer y clasificar un estado mental, por ejemplo "sediento", puede advertirse que el enfoque funcionalista es insatisfactorio.

Según el análisis de Goldman (1993b), si el estado mental ocurrente en que uno se encuentra se clasifica según los *inputs* que lo preceden y los *outputs* o estados internos que le siguen, el estado "sediento" se entiende en términos de sus relaciones con ciertos *inputs*, por ejemplo no haber ingerido líquido por un período prolongado. Asimismo, los *inputs* se entienden según sus relaciones con ciertos *ouputs,* por ejemplo el deseo de beber, y con otros estados mentales, por ejemplo, el deseo de ir a tomar líquido se relaciona con la creencia de que el recipiente que tengo en la mano contiene agua potable y la creencia de que el agua calma la sed. Sin embargo, según Goldman (1993b), de este panorama surgen al menos tres dificultades que sugieren que no es plausible que las personas ejecuten la tarea de autoadscripción de modo puramente relacional.

En primer lugar, parece posible auto-adscribirse un estado mental sin recurrir a información sobre sus causas o efectos (el problema de la "evidencia insuficiente"). Si alguien se despierta con una jaqueca matutina, éste podría reconocerla sin necesidad de recordar alguna circunstancia que pueda considerarse como una causa usual de dolor de cabeza (por ejemplo, haber ingerido alcohol). Asimismo, podría reconocerla con independencia de algún *output* asociado típicamente a las jaquecas, por ejemplo levantarse y tomar una aspirina. Es más, el reconocimiento de una jaqueca puede estar acompañado del deseo de mitigarla pero dado que este deseo es compartido por los dolores en general resulta no ser suficiente para reconocer el estado específico de "jaqueca". De este modo, si el recurso a información de tipo relacional fuera esencial para la tarea de clasificación de estados mentales, ésta no podría llevarse a cabo en absoluto, ni con la rapidez con la que es ejecutada. Sin embargo, constantemente se

identifican y clasifican los estados mentales, por lo tanto, debemos estar apelando más bien a rasgos intrínsecos del estado en cuestión y no a información relacional.

Como es sabido, el funcionalismo no sólo asume que los estados se clasifican según aquellos eventos que los preceden o suceden sino que, además, la identidad de los mismos depende de sus propiedades disposicionales. Por ejemplo, el comportamiento u otros estados mentales que éste pudiera producir. Sin embargo, según Goldman (1993b), esto introduce el problema de la ignorancia de las propiedades disposicionales. Retomando el ejemplo de la jaqueca, probablemente uno sabe que el estado en que se encuentra es tal que causará tomar una aspirina. No obstante, la cuestión es cómo se sabe esto sin haber clasificado previamente el estado en que uno se encuentra, como "una jaqueca" (Goldman 1993b: 378). Si esto es así, resulta que la información de propiedades disposicionales no se utiliza para clasificar los estados sino, contrariamente, la información de clasificación se utiliza para inferir los estados disposicionales.

En tercer lugar, surge la dificultad de que para identificar de manera relacional un ejemplar de estado interno es preciso, a su vez, identificar el tipo de muchos otros estados internos (Goldman 1993b). La naturaleza relacional de los estados mentales, tal como los concibe el funcionalismo, implica que la identidad-tipo de un ejemplar de estado mental depende de la identidad tipo de los estados con los que está relacionado (sus *relata*) y la identidad-tipo de los *relata*, a su vez, depende de los estados mentales con los que estos están relacionados. Claramente, esto lleva a una explosión combinatoria usualmente señalada como una desventaja del enfoque funcionalista sobre los conceptos de estados mentales.

En base a estas dificultades, Goldman (1993b) concluye que las rutinas de autoadscripción no deben estar basadas en información relacional/causal o disposicional sino, más bien, en propiedades que son categoriales e intrínsecas al propio estado, esto es, propiedades que los estados tienen por sí mismos más que en virtud de sus relaciones con otros estados. Las candidatas para cubrir estas

propiedades son dos, las propiedades neuronales y las propiedades cualitativas o fenoménicas (Goldman 1993b: 378).

En principio, las propiedades neuronales no resultan satisfactorias porque el sistema que clasifica los estados mentales no tiene acceso a estas propiedades. Por un lado, las personas no tienen acceso personal a las propiedades neuronales. Por el otro, si bien todo proceso en el cerebro es neuronal en el nivel más inferior, los contenidos o significados codificados por eventos neuronales son contenidos neuronales y estos no generan etiquetas que puedan ser reconocidas como mentales. Básicamente, Goldman está pensando en las actividades homeostáticas o en la respuesta pupilar a la iluminación. En todos estos casos, el procesamiento ocurre en el nivel subpersonal sin que tales estados sean reconocibles en el nivel personal, ni en la conciencia. En general, el proceso etiquetador mental espontáneo no tiene acceso a información puramente subpersonal, salvo cuando los eventos fisiológicos o neurológicos dan lugar a sensaciones conscientes, tales como la sed, en cuyo caso una etiqueta primitiva es introducida o aplicada (Goldman 1993b: 379). Por estas razones, Goldman se inclina por las propiedades fenoménicas o *qualia* como las propiedades intrínsecas que se detectan en la clasificación mentalista, aunque si bien estas propiedades pueden ser detectadas o monitoreadas directamente, no son necesariamente infalibles.

Como consecuencia de esta propuesta, la conciencia fenoménica deviene esencial en la explicación de una tarea cognitiva ordinaria como la autoadscripción de estados mentales. Como es sabido, las sensaciones y las experiencias varían ampliamente. En cada uno de estos casos, uno es sujeto de un estado mental con un carácter subjetivo distintivo. A menudo, los filósofos utilizan el término "*qualia*" para referirse a los aspectos introspectivamente accesibles y fenoménicos de la vida mental (Tye 2013). Si bien el término "*qualia*" se usa en varios sentidos, Goldman (1993b) adhiere particularmente a un sentido amplio de "*qualia*" como carácter fenoménico. Puede decirse que el carácter fenoménico de una experiencia es lo que subjetivamente se siente al estar en ella. Sin embargo, la propuesta de Goldman (1993b) es controversial por varias razones. En primer lugar, se asume que además de las

sensaciones y percepciones, consideradas usualmente fuente de los *qualia*, las actitudes proposicionales también poseen ciertas cualidades que permiten su reconocimiento y clasificación[30]. Goldman (1993b) considera que aplicar el término "*qualia*" sólo a cualidades sensoriales quizás sea una cuestión puramente terminológica y que, quizás, este término pueda ser aplicado a estados que acarrean un sentimiento subjetivo de algún tipo que pueda abarcar a las actitudes proposicionales. Además de Goldman, otros filósofos sostienen que hay experiencias del tipo de entender una oración, de pensar súbitamente en algo o de recordar súbitamente algo, que tienen una sensación subjetiva asociada o *qualia* (Strawson 1994).

Una posible objeción a esta posición señala que la experiencia no provee un sentimiento subjetivo que pueda asociarse a las actitudes proposicionales sino más bien, que las sensaciones subjetivas deben provenir de imágenes lingüísticas o verbales (Tye 2013). Éstas portan la estructura sintáctica y fonológica del habla del sujeto, y usualmente están acompañadas de detalles sobre el énfasis y la entonación. De modo que, al leer, fenoménicamente parece que se está hablando a uno mismo. A menudo se escucha una voz interna, e incluso se pueden experimentar emociones y sentimientos tales como aburrimiento, excitación y demás. Ahora bien, si uno removiera todas estas reacciones junto con las imágenes provenientes de la voz interior y de las sensaciones visuales producidas al leer, no queda ningún resto de fenomenología. Es más, ni siquiera es claro que este tipo de imágenes y sensaciones estén usualmente presentes en el pensamiento, ni que le sean esenciales. Algo similar ocurre con los deseos y otros estados.

En segundo lugar, en contraposición a lo que usualmente se sostiene, este enfoque implica que los *qualia* tienen poderes causales en tanto que, generalmente, producen autoatribuciones verbales e, internamente, desencadenan la actividad de

[30] Usualmente, se considera estados que poseen *qualia* a las experiencias perceptuales, por ejemplo, escuchar trompetas fuertes, las sensaciones corporales, por ejemplo, sentir picazón o calor. Algunos extienden esta lista e incluyen las reacciones, emociones o pasiones, por ejemplo, sentir miedo, y los estados de ánimo, por ejemplo, sentirse eufórico (Tye 2013).

monitoreo (Goldman 1993b). En contraposición, muchos filósofos asumen que los rasgos intrínsecos de las experiencias sensoriales sólo son responsables del carácter fenoménico de las mismas (Tye 2013). De este modo, la propuesta de Goldman se enmarca en la doctrina de acceso privilegiado a los propios estados mentales, que supone la asimetría entre la atribución de primera y tercera persona. Según la tesis de la asimetría, los seres humanos podemos saber lo que pensamos, creemos, deseamos, de una manera distinta a la manera en que conocemos los estados mentales ajenos. Específicamente, Goldman adhiere a la idea de que existe un método especial para detectar los estados mentales propios, que no puede ser aplicado a los estados mentales ajenos. El método especial, como mencioné anteriormente, es la introspección. No obstante, ante la propuesta de introspección, surgen cuestionamientos de orden filosófico y psicológico.

Desde el punto de vista filosófico, una de las objeciones al enfoque de método especial, señala que la fuente del acceso privilegiado reside más bien en una inmunidad metafísica al error y no en un método especial. Además, se ha señalado que el contenido de las actitudes proposicionales es individuado, en parte, por factores ambientales y, por lo tanto, no superviene de las propiedades intrínsecas del pensador (la tesis del externalismo de contenido, ver nota 34) (Putnam 1975; Burge 1979). Desde el punto de vista psicológico, el acceso privilegiado causa escepticismo en virtud de la falibilidad de los reportes introspectivos (Watson 1913). Además, la evidencia sugiere que puede llevarse a cabo la autoatribución por otros procesos que no sean la introspección, a saber, la confabulación (*i.e.* la fabricación de historias explicativas) (Gazzaniga 1995; Gazzaniga & Baynes 2000).

A mi entender, el problema de este argumento reside en que el ejemplo de categorización utilizado (“jaqueca”) por Goldman implica una sensación asociada con jaqueca. Como es sabido los *qualia* se asocian a percepciones y sensaciones y, en este caso, se trataría de la sensación de dolor asociada a la jaqueca. Así, el caso es claro y convincente. Sin embargo, no es claro que esto se pueda extender a otros estados mentales tales como las actitudes proposicionales. Como señalé anteriormente, no es

claro que existan sensaciones subjetivas asociadas a las mismas. No parece ser el caso de que la creencia de que "2+2 es 4" se adscriba en base a los rasgos intrínsecos de este estado, y no es claro a qué otro tipo de estado subjetivo distinto de los *qualia* asociados a las percepciones y sensaciones, podrían asociarse las actitudes proposicionales. Pero, por otro lado, puede concedérsele a Goldman que el funcionalismo presenta ciertas dificultades que impiden que pueda considerarse como un enfoque apropiado para dar cuenta de la autoatribución de estados mentales. No obstante, esta propuesta brinda un rol central a la introspección y los *qualia* que es, al menos, igualmente difícil de defender y hasta, incluso, más controversial. En este sentido, no es claro que este enfoque introspeccionista sobre los estados mentales sea una buena alternativa. En todo caso, éste enfoque tendría que ofrecer una propuesta positiva sobre algún otro tipo de sensación subjetiva a la que se puedan asociar las actitudes proposicionales para ser satisfactorio.

3. La simulación "*off-line*"

Algunos defensores de la TT sostienen que las propuestas empíricas de simulación son muy esquemáticas y, en virtud de esto, Stich & Nichols (1992, 1995, 1996, 1997, Nichols *et al.* 1996) proponen la "simulación *off-line*" como una caracterización más detallada que reúne los puntos principales de las propuestas de Gordon (1995a) y Goldman (1995a) junto con algunas especificaciones[31]. El punto de partida de esta caracterización es el funcionamiento del sistema de toma de decisiones (STD) o de razonamiento práctico, en condiciones normales. A saber, este sistema es alimentado por creencias y deseos como *input*. Las creencias pueden haber sido generadas perceptualmente o por inferencia. Los deseos tienen su origen, en cambio, en sistemas que monitorean los estados corporales, *e.g.* el deseo de calmar la sed, o bien son generados por el STD mismo como medios o metas intermedias para alcanzar

[31] Esta caracterización es aceptada y apropiada por Goldman (Goldman 1995c), mientras que Gordon dice suponerla aunque, a su entender, ésta no recoge el aspecto más relevante de la simulación (Gordon 1995c).

fines, *e.g.* el deseo de dirigirse a la cocina (Stich & Nichols, 1992: 39). Además de generar metas intermedias, este sistema genera decisiones respecto de qué acción realizar y éstas constituyen su *output* característico. A su vez, las decisiones alimentan como *input* a los sistemas motores controladores de la acción, esto es, a los mecanismos cognitivos responsables de la planificación y coordinación para ejecutar decisiones (Stich & Nichols 1992: 39-41).

Ahora bien, el enfoque de la simulación se caracteriza por postular que el STD funciona, además, de manera "*off-line*". Este modo de funcionamiento se diferencia del uso corriente, en la medida en que posibilita que el STD se desacople de los sistemas controladores de la acción y genere decisiones que no serán ejecutadas, sino utilizadas para realizar atribuciones al blanco. A su vez, se asume que, bajo este modo de funcionamiento, el STD puede ser alimentado con creencias y deseos hipotéticos o "ficticios". Esto es, por creencias y deseos relevantemente similares a los del blanco y que el simulador puede compartir o no. En este sentido, durante la simulación, el STD funciona de un modo no estándar: es alimentado por *inputs* ficticios y está desacoplado de los sistemas motores. Así, el STD es alimentado con creencias y deseos muy similares a los del blanco y, puesto que se asume la similitud de los sistemas cognitivos, es esperable que la decisión generada por simulación sea similar a la decisión generada por el blanco. En este sentido, la predicción basada en la simulación mental utiliza parte del propio sistema cognitivo como "modelo" del sistema cognitivo del blanco y, de este modo, parece prescindir del recurso a una estructura interna de información sobre la psicología humana.

En este sentido, según la simulación "*off-line*", considero que postular este mecanismo o proceso cognitivo subyacente implica las siguientes tesis:

1. Los seres humanos poseen sistemas cognitivos relevantemente similares.
2. Es posible establecer las creencias y deseos que son el punto de partida de las decisiones ajenas (sean similares o divergentes a las del simulador).

3. El sistema de toma de decisiones puede utilizarse de manera *off-line*, esto es, desacoplado de los sistemas controladores de la acción, de modo que su *output* constituye una decisión que no se lleva a la acción.
4. El sistema de toma de decisiones puede ser alimentado con *inputs* hipotéticos o ficticios (creencias y deseos que el simulador puede o no tener de hecho, y que se pretende (por 2) que sean relevantemente similares a las del blanco).
5. Dados 1-4, se sigue que uno puede dejar que su propio STD genere una decisión, de modo que ésta será, a menudo, similar a la decisión que genere el STD del blanco.

Asimismo, es preciso señalar que la simulación *off-line* no se restringe al reclutamiento del STD o del sistema inferencial. En principio, cualquier mecanismo mental podría utilizarse de manera *off-line* (*i.e.* desconectado de los sistemas de acción) siempre y cuando éste sea alimentado por creencias y deseos (u otras actitudes proposicionales) como *inputs*, y produzca cualquier tipo de estado mental como *output* (Stich & Nichols 1997: 309). Asimismo, el enfoque de TS no sólo asume que la simulación mental subyace a la predicción del comportamiento y el pensamiento, sino también que ésta tiene un rol en la descripción intencional y la explicación mentalista. Según esta caracterización, el rol de la simulación mental en la explicación intencional del comportamiento ajeno tiene como punto de partida la explicación de la predicción mencionada anteriormente. Se asume que las explicaciones intencionales pueden generarse por medio de una estrategia de análisis por síntesis[32]. La idea es que puedan encontrarse creencias y deseos hipotéticos que, alimentando al sistema de toma de

[32] La estrategia de análisis por síntesis, también llamada "método de prueba y testeo", es propuesta para dar cuenta de cómo un proceso simulacional da lugar a la explicación del comportamiento (Stich & Nichols 1992; Goldman 1995a, 2006; Nichols & Stich 2003). Según este método, primero se generan alternativas de estados mentales que puedan haber dado lugar al comportamiento en cuestión y luego, se ponen a prueba mediante un proceso de simulación. Esto es, se utilizan como *input* para una simulación y, si el resultado de la misma coincide con el comportamiento que quiere ser explicado, el proceso finaliza. Si no, continúa la evaluación de cada alternativa hasta hallar los estados que dieron origen a la conducta de interés.

decisiones, puedan producir una decisión que dé lugar al comportamiento que se quiere explicar.

Generalmente existirán numerosas explicaciones alternativas y habrá que determinar la más plausible. Los teóricos de la simulación sugieren distintas estrategias que permiten acotar las alternativas. Como mencioné en la sección 2.1, los defensores de la simulación proponen estrategias para determinar los estados iniciales del blanco tales como el conocimiento previo de la personalidad del blanco, de su situación perceptual y de sus creencias y deseos, que también se pueden utilizar con el propósito de acotar las explicaciones alternativas. Sin embargo, estas pueden no ser suficientes y una vez aplicadas, aún pueden subsistir numerosas alternativas. En este caso, según los defensores de la simulación, aún resta el recurso de asumir que el blanco es psicológicamente similar al simulador y, así, rechazar aquellas alternativas que el propio simulador considere menos "naturales" según su propia psicología (Goldman 1995a: 90; "el principio de menor simulación" Gordon 1995a: 65).

A su vez, la descripción intencional o la atribución de creencias y deseos a otras personas, funciona de manera similar a la explicación por simulación. Uno de los modos de determinar qué creencias y deseos atribuir a los otros se basa en observar su comportamiento y luego atribuir los estados intencionales que mejor expliquen tal comportamiento. Otra estrategia de atribución basada en la simulación, reside en focalizar en la situación perceptual del agente e inferir, a partir de ésta, las creencias y las experiencias perceptuales correspondientes a la situación. A su vez, esto se puede determinar, como se mencionó, asumiendo que todos tenemos las mismas necesidades básicas de comida, amor, calor y demás (Goldman 1995a: 82).

Ahora bien, en relación al proceso propuesto hay que señalar que es más específo y esto permite notar que la simulación *off-line* postula mecanismos y capacidades mentales en cierta medida novedosas. Primero, se postula algún mecanismo de generación de creencias y deseos ficticios, que construya, a partir de la creencia de que el blanco cree que *p*, la creencia ficticia de que *p**. Segundo, se asume que los sistemas cognitivos reclutables para la simulación no sólo operan sobre *inputs*

estándar sino también sobre *inputs* ficticios. Tercero, hay procesos o mecanismos cognitivos que son "reclutables" para la simulación, esto es, que pueden utilizarse de modo *off-line*. Así, se postula la capacidad de los sistemas cognitivos reclutados para desacoplarse de los sistemas controladores de la acción y utilizar su *output* para realizar *mindreading* (atribución, explicación o predicción mentalista). En este sentido, debe existir algún sistema cognitivo que se encargue de la proyección o adjudicación al blanco del estado mental producto del sistema cognitivo reclutado en la simulación, aquel sistema que operó de modo *off-line* y fue alimentado con *inputs* ficticios. En otras palabras, se asume que, cuando los mecanismos cognitivos son reclutados para la simulación, estos funcionan de manera no estándar.

A mi entender, esta propuesta en relación a los procesos o mecanismos subyacentes tiene la siguiente consecuencia. Si bien la propuesta simulacional parecería ser más económica que una propuesta rica en información en tanto se ahorra la posesión de cuerpos de conocimiento psicológico, sin embargo es compleja en otro aspecto. Postula una arquitectura cognitiva donde existen dos procesos o mecanismos cognitivos que un enfoque de TT no supone, por ejemplo, el generador de *inputs* ficticios y el generador de proyecciones. A continuación me ocuparé de la principal objeción que ha recibido la propuesta de simulación *off-line*.

3.1. La objeción de la "penetrabilidad cognitiva"

Los seres humanos cometen fallas sistemáticas en la predicción del comportamiento ajeno y, según los defensores de la TT, un enfoque de simulación tiene dificultades para dar cuenta de este fenómeno. El argumento de la penetrabilidad cognitiva afirma que el único modo de explicar los errores sistemáticos en las predicciones y las inferencias es apelando a cuerpos de información incompletos o parcialmente erróneos que puedan dar lugar a predicciones incorrectas. En cambio, un enfoque simulacional no dispone de una explicación al respecto. Esto, al menos, por dos razones. Por un lado, el enfoque simulacional considera que la información

psicológica es irrelevante para llevar a cabo *mindreading* y, por otro lado, al parecer sólo puede dar cuenta de eventos de *mindreading* correctos. A continuación me ocuparé en detalle de esta objeción a la simulación.

Según Stich & Nichols (1992, 1995; Nichols *et al.* 1996), la diferencia entre un enfoque rico en información y un enfoque simulacional reside en que, para el primero, la información que posea el sujeto sobre los principios que gobiernan el funcionamiento psicológico es crucial para llevar a cabo *mindreading*, mientras que para el segundo es irrelevante. En este sentido, ambos enfoques difieren en sus expectativas respecto del impacto que tendrá en *mindreading* el conocimiento psicológico poseído por el sujeto. Si bien no se han diseñado experimentos para estudiar la penetrabilidad cognitiva de las predicciones del comportamiento, la literatura de psicología social provee numerosos ejemplos que sugieren que las predicciones son cognitivamente penetrables (Stich & Nichols 1995b: 100). Stich & Nichols (1992, 1995, 1996, 1997, 2003; Nichols & Stich, 2003) recurren a esta literatura y la utilizan como "evidencia indirecta" de fallas sistemáticas en las atribuciones mentalistas y las predicciones comportamentales.

Uno de los estudios más mencionados por los autores es sobre el fenómeno de la "persistencia de las creencias". Se ha notado que tanto en situaciones de laboratorio, así como en la vida real, puede presentársele a una persona evidencia persuasiva que indica que posee un rasgo particular pero inesperado, por ejemplo, mediante los resultados de un examen médico. En virtud de la evidencia presentada, la persona tiende usualmente formarse la creencia de que posee tal característica. El fenómeno en cuestión reside en que si posteriormente se presenta información convincente que desacredita el primer cuerpo de evidencia, por ejemplo se informa que los resultados del examen médico pertenecen a otra persona o que no se realizó ningún examen, la mayoría de las personas, y de los psicólogos sociales con anterioridad a estos estudios, esperan que las creencias desacreditadas sean descartadas. Sin embargo, ciertos estudios muestran que éste no es el caso. Una vez que un sujeto cree que posee cierto rasgo, el sólo hecho de mostrarle que la evidencia

que lo ha convencido es falsa no resulta suficiente para descartar la creencia original (Ross *et al.* 1975; Nisbett & Ross 1980).

Según Stich & Nichols (1992), el hallazgo del fenómeno de la perseverancia de las creencias puede considerarse un ejemplo de la posesión de una teoría psicológica incompleta o errónea. Se trata de aquella parte del sistema de atribución de creencias que condujo a las personas, y a los psicólogos, a esperar que las creencias fueran descartadas luego de ser desacreditadas. Esta teoría contiene información errónea respecto del proceso de perseverancia de las creencias, o bien carece de la misma, y conduce a atribuciones de creencias sistemáticamente equivocadas así como a la sorpresa frente a los resultados de los estudios mencionados. De este modo, mostrar que una capacidad cognitiva es cognitivamente penetrable sugeriría que ésta deriva de una base de información más que de una simulación *off-line* (Nichols *et al.* 1996).

De esta manera, según el enfoque rico en información, si poseemos cuerpos de información equivocados o incompletos podemos esperar que las predicciones sean incorrectas en tales áreas. Por su parte, el enfoque simulacional no niega la posesión de una teoría psicológica internalizada sino que asume, más bien, su irrelevancia. De modo que, si la información psicológica es errónea o incompleta esto no afectará la corrección de las predicciones en la medida en que para llevar a cabo *mindreading* por simulación recurrimos al propio sistema cognitivo que funciona como un modelo del blanco. Además, en el argumento de la penetrabilidad cognitiva se asume que la simulación mental sólo puede producir predicciones correctas respecto de las otras personas. No obstante, se reconocen algunas fuentes de error en la simulación. Específicamente, el simulador debe identificar correctamente las actitudes iniciales del blanco y no deben existir diferencias en el funcionamiento básico de los procesos cognitivos entre el blanco y el simulador (*i.e.* alguna enfermedad o déficit cognitivo). A mi entender, estas dos fuentes de error se consideran, a su vez, condiciones para la simulación. En este sentido, si no se cumplen, esto conducirá a una simulación errónea. Sin embargo, este tipo de error no da cuenta de las fallas sistemáticas en la predicción sino, más bien, de la falla ocasional del proceso de simulación para generar

predicciones correctas. En favor de esta interpretación, puede señalarse que los críticos de la TS usan ejemplos de fallas en las predicciones donde tales condiciones parecen cumplirse en la medida en que es plausible que, por ejemplo, no existan diferencias en el funcionamiento de los procesos cognitivos del blanco y del simulador ya que todos los participantes se han elegido de manera aleatoria, entre la misma población de sujetos. Con esto los oponentes creen poder mostrar que, aún cuando la simulación mental es un método posible para realizar predicciones mentalistas, no es el método que la gente usa.

En suma, "penetrabilidad cognitiva" se entiende en el sentido de que una capacidad se ve afectada por el conocimiento o ignorancia que el sujeto tenga de determinado dominio (Nichols *et al.* 1996). Así, según el argumento de la penetrabilidad cognitiva las fallas sistemáticas sugieren un proceso de bases de información subyacente, mientras que los aciertos en las predicciones del comportamiento ajeno no permiten distinguir entre tipos de procesos, porque ambos pueden subyacer a las predicciones exitosas (Stich & Nichols 1996: 160). De este modo, si las predicciones mentalistas se basan en un cuerpo de conocimiento sobre procesos psicológicos, entonces el uso de una teoría incorrecta puede conducir a predicciones erróneas, y las fallas en la predicción tienen una explicación en términos de un enfoque rico en información. No obstante, tal como argumentaré en el capítulo 4, considero que es posible que un proceso de tipo simulacional pueda conducir a predicciones fallidas dadas ciertas condiciones.

3.2. Un enfoque de simulación plausible

Stich & Nichols se mantienen escépticos respecto de la posibilidad de que la simulación *off-line* sea el proceso cognitivo subyacente a *mindreading*. Sin embargo, consideran que es plausible que otro tipo de simulación subyazca a ciertas atribuciones (Stich & Nichols 1995, 1997; Nichols *et al.* 1996; Nichols 2003). A continuación trataré de echar luz respecto de esta noción más plausible de simulación. Esta tarea es

importante porque considero que esta es la noción de simulación que estará a la base del modelo híbrido de teoría y simulación de Nichols & Stich (2003), que analizaré en el capítulo 4.

El origen de la noción de simulación con la que simpatizan Stich & Nichols puede rastrearse en el siguiente argumento de Harris (1995). Supongamos que se lleva a cabo un estudio psicolingüístico en el cual los hablantes de un idioma (*e.g.* español) juzgan la gramaticalidad de ciertas oraciones. Si, luego de realizar esto, se le pide a otro hablante del español que prediga qué decisiones tomaron los participantes del estudio, se observará una tasa de aciertos altísima. En la mayoría de los casos, este hablante podrá decir si la mayoría de los participantes juzgó una oración como gramatical o no. Es más, si se le pide al hablante que explique sus predicciones, éste lo hará indicando las mismas construcciones y morfemas en las oraciones agramaticales que las que fueron señaladas por los participantes del estudio.

Según Harris (1995), esto se puede explicar de dos maneras. O bien, el hablante se pregunta a sí mismo si la oración es gramatical o no y, luego, asume que los otros hablantes del idioma realizarán los mismos juicios recurriendo a las mismas razones. O bien, se postula la posesión de dos representaciones tácitas de la gramática. Una representación de primer orden que se utiliza para realizar los juicios propios. Y una representación de las representaciones de las otras personas, diseñada para producir juicios equivalentes, que se utiliza para predecir los juicios que realizan los otros. Según Harris, esta última es la respuesta que un "teórico de la TT" tiene a la mano. Dado que esta última explicación es compleja y poco parsimoniosa en relación a la primera, debe preferirse la primera (Harris 1995: 210-211).

Así, según el argumento de Harris, la explicación a disposición de los teóricos de la teoría implica la postulación de gramáticas duplicadas. Esto es, una gramática para realizar los juicios lingüísticos propios, y una teoría sobre cómo los hablantes del español usan la gramática para realizar juicios lingüísticos. Frente a esto, la posibilidad de postular el recurso a la propia gramática para realizar juicios lingüísticos, que luego se proyectarán de alguna manera en el otro, resulta una explicación más sencilla en

tanto no supone una reduplicación. Según Harris (1995), este argumento brinda apoyo al enfoque simulacional.

Stich & Nichols (1995, 1997; Nichols *et al*. 1996; Nichols 2002) consideran muy plausible que ciertas capacidades cognitivas, como la predicción de los juicios lingüísticos ajenos, estén basadas en este tipo de simulación. La principal razón para sostener la plausibilidad de este tipo de simulación frente a la simulación *off-line* reside en que, al parecer, la simulación *á la* Harris es esencialmente diferente de la simulación *off-line* en tanto no recurre a la ficción (Nichols 2002: 8). En este sentido, mientras que se considera poco plausible que la simulación *off-line* subyazca a *mindreading*, la simulación por "situación real" puede ser útil para predecir las creencias del blanco en ciertas ocasiones. Puesto que este tipo de simulación requiere ponerse en la situación real del otro es plausible que pueda guiar ciertas predicciones como, por ejemplo, la predicción del resultado de una operación matemática llevada a cabo por otro agente (o la predicción de una inferencia, o qué piensa el blanco). A su vez, este proceso se considera plausible respecto de la predicción de deseos y emociones ajenas. Por ejemplo, podemos tomar una droga psicoactiva, observar qué efectos produce en nosotros (deseos y emociones) y extenderlos al caso del blanco (Stich & Nichols 1997: 300-301). Nótese que en todos estos casos el agente simula al blanco poniéndose en una situación (real) similar, sin necesidad de alimentar el proceso con *inputs* ficticios. En mi opinión, "real" puede entenderse también como el requisito de compartir las mismas creencias, por ejemplo las que se pueden compartir al realizar una operación matemática.

De este modo, Stich & Nichols (1997) caracterizan los distintos tipos de simulación a partir del tipo de *input* que la alimenta, si es ficticio o no. Así, distinguen entre la simulación "*off-line* conducida por ficción" y la simulación "conducida por situación real", que es el tipo de simulación que subyace a la predicción de juicios gramaticales ajenos (Harris 1995). Estos tipos de simulación difieren en la medida en que la simulación por situación real no postula los mecanismos y las capacidades "extra" que postula la simulación *off-line* y esto es considerado una ventaja (Stich &

Nichols 1997). Sin embargo, este tipo de simulación requiere que lleguemos a tener los mismos deseos y creencias que tienen los otros, y este requisito es, a menudo, muy difícil de satisfacer. Por esta razón, los autores consideran que la simulación por situación real no puede constituirse como un proceso que se utiliza habitualmente. Así, en virtud de que este requisito es difícil de satisfacer, según Stich & Nichols (1997), queda patente que la mayoría de los casos de predicción de la acción y el pensamiento ajeno no son casos de este tipo de simulación. En el sentido de que, por un lado, consideran esto absurdamente impráctico y, por el otro, psicológicamente imposible en la medida en que no hay dos sistemas de creencias iguales.

La diferencia entre un proceso simulacional alimentado por *inputs* ficticios o reales va a ser una constante en la propuesta de Stich & Nichols. Incluso, como señalaré en el capítulo 4, en la propuesta híbrida (Nichols & Stich 2003) se considera que las creencias y deseos ficticios (los particulares en la caja de mundos posibles) no tienen las mismas propiedades que las creencias propiamente dichas (los particulares en la caja de creencias). Es más, este aspecto será cuestionado por Goldman (2006, 2009) y motivará su cuestionamiento del carácter híbrido de la propuesta de Nichols & Stich (2003), particularmente, cuestiona el rol de la simulación en la propuesta. Según Goldman (2006), para la simulación es esencial que las creencias propiamente dichas y las creencias ficticias tengan el mismo rol funcional puesto que, de otro modo, se desdibuja el carácter simulacional del proceso. De esta cuestión me ocuparé en detalle en el capítulo 4.

4. Conclusión

En este capítulo me he ocupado de las versiones empíricas de la simulación de tendencia científico-cognitivas por dos razones. Por un lado, porque estos enfoques puros postulan procesos cognitivos subyacentes y esto está en concordancia con el objetivo de la tesis de evaluar si los enfoques híbridos de teoría y simulación logran brindar una descripción y una explicación satisfactoria de los procesos subyacentes a

mindreading. Por el otro, porque estas caracterizaciones se retoman en las propuestas híbridas que evaluaré en los capítulos 4 y 5 de esta tesis (Nichols & Stich 2003; Goldman 2006, respectivamente). En particular, Goldman acepta explícitamente la caracterización *off-line* a la base de su propuesta híbrida (Goldman 2006: 20). A su vez, es preciso señalar que la simulación sólo puede dar cuenta de la predicción, descripción y explicación del comportamiento y el pensamiento de otras personas. En este sentido, la propuesta simulacional necesita ser complementada con alguna propuesta sobre la autoatribución para dar cuenta de *mindreading*. Esto se evaluará para cada propuesta híbrida particular en los capítulos 4 y 5.

Considero que las tesis de la simulación que se van a retomar para los enfoques híbridos coinciden con las tesis de la simulación "*off-line*", que constituye la propuesta más detallada, presentada en la sección 3:

1. Los seres humanos poseen sistemas cognitivos relevantemente similares.
2. Es posible establecer las creencias y deseos que son el punto de partida de las decisiones ajenas.
3. El sistema de toma de decisiones puede utilizarse de manera *off-line*, esto es, desacoplado de los sistemas controladores de la acción, de modo que su *output* constituye una decisión que no se lleva a la acción.
4. El sistema de toma de decisiones puede ser alimentado con *inputs* hipotéticos o ficticios (creencias y deseos que el simulador puede o no tener de hecho, y que se pretende (por 2) sean relevantemente similares a las del blanco).
5. Dados 1-4, se sigue que uno puede dejar que su propio STD genere una decisión, de modo que ésta será, a menudo, similar a la decisión que genere el STD del blanco.

Es necesario agregar que, en principio, cualquier sistema que pueda ser alimentado con actitudes proposicionales como *input* y que produzca estados mentales como *output* puede ser reclutable para la simulación. En el capítulo 3, me ocuparé de las razones en las que se basa el vuelco reciente, más o menos masivo, hacia posiciones híbridas de teoría-simulación en el estudio de la competencia mentalista. Este vuelco,

está motivado, por un lado, por una necesidad del complemento de simulación por parte de la TT y del complemento de teoría por parte de la TS que parece surgir de las insuficiencias de los enfoque puros que he adelantado en los capítulos 1 y 2. Por el otro, porque ambos enfoques, teoría y simulación, parecen tener casos a su favor. Asimismo, me ocuparé de los requisitos de mínima que tiene que cumplir una propuesta híbrida de teoría y simulación para dar cuenta de *mindreading*.

CAPÍTULO 3. REQUISITOS PARA UN ENFOQUE HÍBRIDO DE TEORÍA Y SIMULACIÓN

En este capítulo me ocuparé de los requisitos para evaluar los enfoques híbridos de *mindreading*. Previamente, en la sección 1, definiré con más precisión qué entiendo por *mindreading*, en relación a las cuestiones y las capacidades que abarca. En la sección 2, presentaré lo que considero que han sido las motivaciones para postular enfoques híbridos de teoría y simulación en la literatura de *mindreading*. A saber, por un lado, la posibilidad de complemento entre teoría y simulación y, por otro lado, los casos a favor de cada enfoque. La sección 3 está dedicada a los requisitos. En la sección 3.1, analizaré los requisitos para la atribución de estados mentales a otras personas. En la sección 3.2, abordaré los requisitos para la autoatribución de estados mentales. En la sección 3.3, propondré como requisito la necesidad de proporcionar un criterio para distinguir el tipo de proceso subyacente a un caso de *mindreading*. Considero que estos tres son los requisitos fundamentales para cualquier enfoque híbrido o que postula más de un tipo de proceso cognitivo subyacente a *mindreading*.

1. ¿Qué es *Mindreading*?

En esta sección retomaré una serie de consideraciones sobre cuestiones generales de las que ya me he ocupado en cierta medida en la introducción y el capítulo 1. Como he mencionado, "*mindreading*" y "psicología de sentido común" no son términos equivalentes. Existe un amplio acuerdo respecto de que la psicología de sentido común consiste en el conjunto de principios que el común de la gente utiliza cotidianamente para entender, explicar y predecir la conducta, y los estados mentales, propios y ajenos. Usualmente, tal caracterización de la psicología de sentido común está estrechamente asociada a la idea de que este cuerpo de conocimiento acerca de la mente constituye una "teoría". Esto es sostenido, particularmente, por los teóricos de la TT en las versiones del funcionalismo (Sellars 1956, Lewis 1970, 1972) y del niño científico (Gopnik 1993, Gopnik & Wellman 1992, 1994; Gopnik & Meltzoff 1997). Una consecuencia de esta perspectiva reside en tener que dar cuenta de si el conocimiento

de sentido común acerca de la mente puede considerarse, y en qué medida, como un conjunto de proposiciones que puedan conformar una teoría.

Como también ya he mencionado, dado que han surgido enfoques alternativos a la TT para dar cuenta de la psicología de sentido común, parece apropiado abandonar todas aquellas etiquetas que hagan referencia al elemento de teoría. Así, etiquetas como "psicología de sentido común" o "Teoría de la Mente" han sido recientemente reemplazadas por rótulos como "*mentalizing*" o "*mindreading*", que tienen la ventaja de ser más neutrales en la medida en que no aluden a un elemento teórico subyacente. Por esta razón, elegí utilizar el término neutral "*mindreading*".

Además del viraje hacia términos más neutrales, ha tenido lugar un cambio de foco en la discusión de *mindreading* en la medida en que se separan dos cuestiones que antes estaban relacionadas estrechamente (Davies & Stone 1995). La idea de que una teoría psicológica de sentido común, constituida por generalizaciones legaliformes, guía las atribuciones mentalistas de la gente está estrechamente relacionada con la idea de que una teoría de sentido común es la que otorga significado a los conceptos mentales. Sin embargo, se trata de dos cuestiones distintas, que pueden tratarse independientemente y esta distinción se ha entendido como la diferencia entre "tener conocimiento" y "usar conocimiento". Así, mientras la cuestión de los conceptos mentales implica dar cuenta del modo de "tener conocimiento" de los conceptos mentales, la cuestión de la práctica de la psicología de sentido común implica dar cuenta del modo de "usar el conocimiento" (Davies & Stone 1995).

Como vimos en el capítulo 1 (sección 2), la cuestión de cómo los conceptos mentales adquieren su significado ha sido tratada, principalmente, de manera conceptual por los filósofos. Los psicólogos que han adherido a un enfoque de TT, no se han detenido en esta cuestión, aunque plausiblemente sus afirmaciones sean compatibles con una teoría funcionalista de los conceptos. De este modo, la tendencia en el estudio de *mindreading* consiste en asumir el conocimiento de los conceptos y desarrollar la cuestión del uso de los mismos. Así, al ocuparme de *mindreading* estaré dejando de lado cuestiones tales como la del estatus de teoría y de los conceptos

mentales en deferencia de la cuestión de los procesos cognitivos subyacentes a *mindreading*.

Asimismo, desde una perspectiva filosófica, hay acuerdo respecto de que "*mindreading*" abarca un conjunto de habilidades o capacidades cognitivas entre las que pueden enumerarse: (1) reconocer estados mentales, (2) atribuir estados mentales, (3) describir intencionalmente a las personas, (4) describir intencionalmente los comportamientos, (5) realizar inferencias sobre estados mentales, (6) realizar predicciones de estados mentales basadas en estados mentales, (7) realizar predicciones del comportamiento basadas en estados mentales, (8) realizar explicaciones de estados mentales basadas en estados mentales, (9) realizar explicaciones del comportamiento basadas en estados mentales, (10) especular, describir, evaluar, recordar estados mentales, (11) especular, describir, evaluar, recordar disposiciones para tener estados mentales, (12) especular, describir, evaluar, recordar disposiciones para actuar, (13) anticipar el comportamiento sin decirlo verbalmente, (14) reconocer generalizaciones de psicología de sentido común (Davies & Stone 1998; Ravenscroft 2010; Stich & Nichols 1992; 1995; 2002; Stich & Ravenscroft 1994).

En líneas generales, se puede entender *mindreading* en un sentido amplio que incluya todas estas capacidades. Sin embargo, la lista exhaustiva de capacidades mentalistas puede ser discutible. Por esta razón, utilizaré un sentido mínimo de "*mindreading*" según el cual las capacidades que se agrupan bajo este término son, al menos, las de (2) atribuir estados mentales, (6) y (7) predecir estados mentales y comportamiento en base a estados mentales. Es probable que las capacidades adicionales (1, 3-5, 8-12) estén basadas en las del "sentido mínimo", pero esto no puede asumirse simplemente. En este sentido, considero que *mindreading* abarca al menos, de manera fundamental las capacidades de atribución de estados mentales, propios y ajenos, y de predicción y explicación de estados mentales y comportamiento en términos mentalistas.

2. Razones para un enfoque híbrido de teoría y simulación

Como ya se ha mencionado en la introducción, la postulación de los enfoques híbridos de teoría y simulación de *mindreading* está en concordancia con el supuesto ampliamente aceptado de que *mindreading* es un fenómeno complejo y, en este sentido, es probable haya más de un mecanismo o proceso cognitivo subyacente. No obstante, creo que pueden señalarse, al menos, dos fuentes de motivación provenientes del debate teoría y simulación, que han colaborado en el vuelco hacia los enfoques híbridos en la literatura de *mindreading*. Por un lado, parece probable que teoría y simulación puedan complementarse de modo que algunos de los aspectos defectuosos de los enfoques puros puedan rectificarse. Por otro lado, quizás el factor que ha tenido mayor peso, parecen existir casos a favor de cada uno de los enfoques. Al parecer, la simulación subyace plausiblemente a cierto tipo de predicciones sobre los juicios de otras personas (Harris 1995; Nichols & Stich 2003; Stich & Nichols 2003; Goldman 2006; Apperly 2008) y, por su parte, la TT parece disponer de la mejor explicación sobre las fallas sistemáticas en *mindreading* (Stich & Nichols 1992, 1997, 2003; Apperly 2008). A continuación me ocuparé de estas dos fuentes de motivación.

2.1. El complemento de teoría y simulación

Se ha sugerido que algunas deficiencias de los enfoques puros pueden subsanarse con el auxilio del otro enfoque. En este sentido, cabe la posibilidad de que los enfoques híbridos no hereden las desventajas de los enfoques puros sino que, más bien, las compensen. A continuación voy a presentar algunos aspectos en los que se ha sugerido que teoría y simulación se complementan.

2.1.1. La simulación precisa el complemento de teoría

El uso de la simulación mental para realizar predicciones sobre otras personas asume dos cuestiones. Primero, que los seres humanos poseen sistemas cognitivos relevantemente similares. Segundo, que las otras personas son relevantemente similares al simulador con respecto a los deseos y las creencias que poseen (Davies & Stone 2000). En relación a estas afirmaciones se determinan dos fuentes de error. Una simulación puede resultar errónea si, por alguna razón, el funcionamiento de los sistemas cognitivos del *mindreader* y del blanco no es relevantemente similar. Si los deseos y las creencias del *mindreader* y del blanco difieren, el *input* de la rutina simulacional puede no ser el indicado y generar un *output* que no coincide con el estado mental del blanco. En relación a estas dos fuentes de error la teoría puede ayudar a la simulación.

Si es el caso que una persona es relevantemente diferente al simulador, será preciso evaluar en qué aspecto lo es para llevar a cabo los ajustes necesarios. El problema para la simulación reside en que, en este caso, al simulador no le alcanza con simular la situación del blanco para realizar los ajustes en relación a la diferencia relevante, de modo que no queda claro cómo determinará el simulador qué experiencias del blanco hay que simular. La TS no tiene una respuesta a la mano y, en cambio, la TT tiene una simple. No hace falta partir del propio caso para realizar luego los ajustes necesarios, sino que simplemente se comienza desde la tercera persona. Así, el hecho de que la persona resulte relevantemente diferente no implica un problema, puesto que sus deseos y creencias se pueden inferir usando generalizaciones sobre el comportamiento e información de lo que sabemos sobre su situación. Por ejemplo, es posible determinar qué es lo que hará mi compañero de *trekking* que es miope y camina sin anteojos junto a mí, al toparnos con un animal en el camino de montaña. Esto se puede inferir a partir de información sobre su situación, su condición física, cierta teoría respecto de la visión miope, de este modo, la falencia de la simulación puede compensarse mediante el recurso a un cuerpo de conocimiento psicológico respecto de cuáles son las creencias y deseos asociados característicamente a ciertos comportamientos y situaciones perceptuales.

Asimismo, puede ser el caso que falle la primera asunción. Esto es, pueden existir diferencias relevantes entre los procesos mentales del blanco y del simulador. Por ejemplo, puede ser el caso que el blanco haya ingerido una droga que afecte el funcionamiento normal de los procesos de razonamiento teóricos o prácticos. En este caso, no es posible para el simulador compensar su diferencia con el blanco por medio de imaginar la ingesta de la droga. Más bien, será preciso utilizar información empírica sobre cómo la droga influencia el razonamiento o el comportamiento. De este modo, una vez más es preciso invocar un cuerpo de conocimiento para establecer los estados mentales del blanco (Davies & Stone 2000). En este caso, posiblemente la simulación mental sea utilizada para generar la predicción inicial que luego será modificada a la luz de la información empírica sobre los efectos de la droga.

Así, en el primer caso, el auxilio de la teoría interviene para establecer los parámetros para iniciar una simulación. En el segundo caso, el auxilio de la teoría actúa entre el *output* del mecanismo de toma de decisiones y la proyección del mismo al blanco, al final de la simulación. En ambos casos, la teoría no reemplaza completamente a la simulación sino que, más bien, la auxilia. No obstante, en mi opinión, no es claro que se trate de un complemento de teoría y simulación. Si bien puede concederse que cierto cuerpo de conocimiento viene a auxiliar a la simulación, no es claro que se trate del cuerpo de conocimiento psicológico que tienen en mente los defensores de la TT. No es claro que los principios a los que se acude en ambos casos, a saber, principios sobre la visión miope y sobre los efectos de cierta droga, formen parte de un cuerpo de conocimiento sobre el funcionamiento de la mente que cualquier persona podría poseer; más bien parece conocimiento de expertos.

2.1.2. La teoría precisa el complemento de la simulación

En la literatura se han propuesto algunos casos de funcionamiento de TT con el auxilio de la simulación. Básicamente, el aporte de la simulación está vinculado con la posibilidad de acceder, por medio de la misma, a nuestros propios hábitos y

disposiciones que vienen en auxilio de las “lagunas” en la teoría psicológica de sentido común. Así, a la hora de la formación de creencias en base a la percepción, en algunas ocasiones parece razonable considerar las propias percepciones con el propósito de establecer las creencias que forman las personas. Por ejemplo, cuando el simulador comparte el mundo circundante con el blanco, en principio, parece derrochador asumir que un sujeto precisa recurrir a una teoría de la opacidad de los cuerpos para determinar si la persona con la que está sentada a la mesa, puede ver lo que está debajo de la misma. Más bien, poniéndose en el lugar del otro pueden determinarse las creencias que pueda tener el blanco desde su perspectiva perceptual. En la medida en que la simulación permite la incorporación de los propios hábitos y disposiciones, este uso de la simulación puede auxiliar a la TT en este respecto.

Asimismo, se ha señalado que el enfoque de la TT enfrenta el problema de tener que dar cuenta de cómo seleccionar ciertos deseos y creencias como relevantes entre una vasta red de creencias y deseos, sin que haya podido establecerse hasta el momento una teoría de la relevancia. En este sentido, se ha sugerido que la simulación podría ayudar en esta tarea porque, en la medida en que la simulación implica usar los recursos cognitivos propios como modelo del blanco, permitiría que nuestras propias intuiciones sobre lo que es relevante puedan guiarnos en esta tarea (Heal 2003). Si se concede que es posible usar nuestras intuiciones de relevancia, en tanto la simulación permite usar nuestros hábitos y disposiciones, éstas funcionarían cognitivamente al modo de heurística o atajos para calcular lo relevante, sin necesidad de poseer una teoría de la relevancia.

Sin embargo, a mi entender estas consideraciones son demasiado especulativas. Por un lado, no hay evidencia, al menos que yo conozca, que avale la posibilidad de usar las propias “intuiciones sobre lo relevante” para determinar las creencias relevantes de los otros. En este sentido, esto permanece como una mera posibilidad. Por otro lado, en mi opinión, la simulación enfrenta dificultades similares. En particular, ciertas versiones de la TS han propuesto la estrategia de análisis por síntesis para dar cuenta de cómo un proceso simulacional da lugar a la explicación del

comportamiento (Stich & Nichols 1992; Goldman 1995a, 2006; Nichols & Stich 2003). Según este método, primero se generan alternativas de estados mentales que puedan haber dado lugar al comportamiento en cuestión y luego, se ponen a prueba mediante un proceso de simulación. Esto es, se utilizan como *input* para una simulación y, si el resultado de la misma coincide con el comportamiento que quiere ser explicado, el proceso finaliza. Si no, continúa la evaluación de cada alternativa hasta hallar los estados que dieron origen a la conducta de interés. Como es sabido, numerosos estados mentales pueden dar lugar a un único comportamiento. Según los teóricos de la simulación, esto implica que, probablemente en esta estrategia sea preciso evaluar numerosas alternativas antes de encontrar la adecuada. La cuestión es que si numerosos estados dan lugar a un mismo comportamiento, cualquiera de estos estados dará lugar al comportamiento que se busca. Este explicará la conducta del blanco. Sin embargo, a la simulación le interesa saber cuáles fueron los estados particulares del blanco que dieron lugar a su comportamiento, no le interesa cualquier estado mental que pueda originarlo. De modo que surge el problema de cómo se determina que los estados que alimentaron la rutina de análisis por síntesis sean los estados adecuados. Acaso ¿se le pregunta al blanco por los mismos? Más allá de esta dificultad, además, ningún defensor de la simulación ha dado cuenta de cómo este método se las arregla con las alternativas numerosas. En caso de que fuera preciso acudir a una teoría de la relevancia para acotar las alternativas, el enfoque simulacional se encontraría en la misma situación que la TT.

En líneas generales, si bien creo hay una posibilidad de complementación, no creo ésta pueda cubrir todas las falencias de los enfoques puros. En este sentido, no creo que la complementación constituya la razón principal para postular enfoques híbridos de teoría y simulación. A continuación, me ocuparé de otra motivación que, a mi entender, ha tenido mayor incidencia en el vuelco hacia las propuestas híbridas.

2.2. Los casos a favor de la "Teoría de la Teoría" y la "Teoría de la Simulación"

Otra motivación para los enfoques híbridos de teoría y simulación son los casos a favor de cada enfoque. Se trata de argumentos que, en la literatura, se consideran convincentes y encuentran una amplia adhesión. En el caso de la TT, este enfoque parece disponer de la mejor explicación de las fallas sistemáticas en las predicciones y las atribuciones mentalistas (Stich & Nichols 2003; Apperly 2008). Como he señalado en el capítulo 2 (sección 3.1), según Stich & Nichols (1992), la diferencia entre un enfoque rico en información y un enfoque simulacional reside en que, para el primero, la información que posea el sujeto sobre los principios que gobiernan el funcionamiento psicológico es crucial para llevar a cabo *mindreading*, mientras que para el segundo es irrelevante (aunque no niega su existencia). En este sentido, ambos enfoques difieren en sus expectativas respecto del impacto que tendrá, en *mindreading*, el conocimiento psicológico poseído por el sujeto. Si la información psicológica es errónea o incompleta esto no afectará la corrección de las predicciones generadas mediante simulación en la medida en que, en este proceso, se recurre al sistema cognitivo propio que funciona como un modelo del blanco. Si poseemos cuerpos de información parcialmente equivocados o incompletos podemos esperar que las predicciones sean incorrectas en tales áreas. Así, "penetrabilidad cognitiva" se entiende en el sentido de que una capacidad se ve afectada por el conocimiento o ignorancia que el sujeto tenga de determinado dominio (Nichols *et al.* 1996).

En este sentido, si hay casos donde las personas fallan sistemáticamente al predecir el comportamiento hay razones para pensar que a estas fallas sistemáticas les subyace un cuerpo de información parcialmente incompleto o erróneo. Así, mostrar que una capacidad cognitiva es cognitivamente penetrable indica que ésta deriva de una base de información más que de una simulación *off-line* (Nichols *et al.* 1996). Si bien no se han diseñado experimentos para estudiar la penetrabilidad cognitiva de las predicciones del comportamiento, en la literatura de psicología social pueden rastrearse fallas sistemáticas en las atribuciones mentalistas y predicciones comportamentales, que pueden explicarse como predicciones generadas por teorías defectuosas. En el capítulo 2 (sección 3.1), me ocupé del fenómeno de la perseverancia

de las creencias que puede considerarse como el producto de la posesión de una teoría psicológica defectuosa de atribución de creencias. Esta conduce (erróneamente) a esperar que las creencias se descarten luego de ser desacreditadas, sin embargo, las creencias persisten a pesar del descrédito. Así, la TT puede dar cuenta de las fallas sistemáticas en las predicciones como el producto de cuerpos de información defectuosos, mientras que la simulación no puede recurrir al argumento de "predicciones erróneas generadas por teorías equivocadas" en la medida en que considera que los cuerpos de información son irrelevantes para llevar a cabo *mindreading*. No obstante, la simulación presenta una ventaja. Es ampliamente aceptado que cuando anticipamos el juicio de alguien sobre la gramaticalidad de una oración utilizamos nuestras intuiciones gramaticales (no teóricas), y esto provee un caso convincente de simulación (Nichols & Stich 2003; Stich & Nichols 2003; Goldman 2006; Apperly 2008). El argumento de Harris (1995) comienza con la propuesta del siguiente experimento mental. Supongamos que se ha llevado a cabo un estudio psico-lingüístico en el cual hablantes de un idioma (*e.g.* español) juzgaron la gramaticalidad de ciertas oraciones. Si luego se le pide a otro hablante del español que prediga qué decisiones tomaron los participantes del estudio, se observará una tasa de aciertos altísima. En la mayoría de los casos este hablante podrá decir si la mayoría de los participantes juzgó una oración como gramatical o no. Es más, si se le pide que explique sus predicciones lo hará indicando las mismas construcciones y morfemas en las oraciones agramaticales que las que fueron señaladas por los participantes del estudio (Harris 1995: 210).

Un modo plausible de explicar esto, según Harris (1995) el más plausible, sugiere que el hablante se pregunta a sí mismo si la oración es gramatical o no y luego, asume que el otro hablante del idioma realizará los mismos juicios recurriendo a las mismas razones. La respuesta que un "teórico de la TT" tiene a la mano implica, según Harris, proponer la posesión de dos representaciones tácitas de la gramática, una representación de primer orden que se utiliza para realizar los juicios propios y una representación de las representaciones de las otras personas, diseñada para producir

juicios equivalentes, que se utiliza para predecir los juicios que realizan los otros (Harris 1995: 210-211). Dado que esta última explicación parece complicada y poco parsimoniosa en comparación con la primera, Harris (1995) concluye que la primera debe preferirse. Este argumento por la simplicidad de Harris, que también ha sido adelantado en el capítulo 2 (sección 3.2) es cuestionable y me ocuparé con detalle del mismo en el capítulo 4. No obstante, para muchos se trata de un caso convincente de utilización de un mecanismo propio para predecir los juicios gramaticales de otras personas y consideran que es plausible extender este argumento a otros dominios tales como la predicción de inferencias (Nichols & Stich 2003: 105; Stich & Nichols 2003: 12; Goldman 2006: 182).

3. Requisitos para un enfoque híbrido de *mindreading*

Se asume que *mindreading*, en tanto la capacidad de pensar acerca de lo que las otras personas perciben, conocen, creen, desean y demás, constituye el núcleo de los procesos cognitivos claves para la interacción social y la comunicación[33]. La literatura empírica y teórica de *mindreading* se ha desarrollado extensamente en los últimos 35 años, con un gran volumen de experimentos, que además suelen presentar resultados dispares al punto de que no creo que nadie es capaz de retener en la mente todos los detalles a la vez, y darles sentido. De modo que, cualquier selección de hallazgos resulta arbitraria. No obstante, recientemente esta literatura se ha revitalizado, por un lado, con el hallazgo de que la comprensión mentalista de la conducta es temprana si las tareas de *mindreading* son adaptadas adecuadamente (Onishi & Baillargeon 2005; Southgate *et al.* 2007; Surian *et al.* 2007; Scott & Baillargeon 2009; Baillargeon *et al*. 2013) y, por el otro, en virtud de la plétora de estudios sobre las bases neuronales de *mindreading* que forman parte del programa de

[33] No obstante, se ha sugerido que esta capacidad puede estar sobrestimada y que, en numerosas ocasiones, las conductas sociales pueden coordinarse sin necesidad de inferir estados mentales sino que basta con guiarse por señales comportamentales (Perner & Ruffman 2005; Apperly 2011; Perner & Roesler 2012).

investigación de la "neurociencia social". Intentaré determinar algunas características respecto de las cuales parece existir un acuerdo considerable y que, en este sentido, pueden funcionar como *desiderata* de mínima para un enfoque de *mindreading*.

3.1. Atribución de estados mentales a los otros

En principio, todos estarían dispuestos a asumir que los adultos normales atribuyen estados mentales a las otras personas y, en la medida en que estos se consideran las causas del comportamiento, los estados mentales pueden inferirse a partir de lo que las personas hacen y dicen. Asumir esto no es controversial y se acepta que es algo que hacemos regularmente. En base a estos supuestos, *mindreading* ha sido estudiada extensamente por psicólogos del desarrollo, de la psicología comparada y, más recientemente, por neurocientíficos y psicólogos cognitivos.

Ahora bien, la cuestión de cómo es que *mindreading* se lleva a cabo genera una serie de preguntas teóricas y empíricas que han sido tratadas parcialmente en la literatura. La pregunta que ha conducido primordialmente la investigación empírica consiste en cuándo se adquieren los conceptos de *mindreading*. En este sentido, se asume que hay un criterio claro para establecer si tales conceptos se poseen o no, a saber, mostrar un desempeño exitoso en la TFC (ver Introducción, sección 1.1). De modo que, si se poseen los conceptos de *mindreading*, se ha adquirido la habilidad y no parece quedar mucho por explicar. No obstante, algunos han sugerido que quizás la habilidad de *mindreading* implique algo más que la mera posesión de conceptos mentales. Los estudios de *mindreading* en adultos muestran ciertas dificultades en el desempeño, justamente en aquella población sobre la que no hay dudas respecto de la posesión de los conceptos de *mindreading*, y esto sugiere que no basta con los mismos para dar cuenta de la habilidad madura (Apperly 2011).

La mayor parte de la investigación empírica en *mindreading* se ha concentrado en el rango de edad de los 3 a los 5 años, prestando poca atención a niños más pequeños, niños mayores y adultos. De esta literatura surgen una serie de hallazgos

sobre los que hay gran acuerdo. A saber, a la edad de 2 años, los niños claramente advierten la diferencia entre los pensamientos en la cabeza y las cosas en el mundo. En el juego de ficción (por ejemplo, finjo que un bloque es un auto), los niños muestran que pueden distinguir entre un objeto, *i.e.* el bloque, y los pensamientos acerca del objeto, *i.e.* el bloque como objeto (Kavanaugh 2006). A su vez, comprenden que las personas se sienten contentas si consiguen lo que desean y tristes si no (Wellman & Banerjee 1991). A esta misma edad, los niños son capaces de advertir que puede existir una diferencia entre lo que ellos desean y lo que otros desean (Meltzoff, Gopnik & Repacholi 1999). De modo que muestran una comprensión amplia de los deseos. Este desarrollo de la comprensión puede verse reflejado en el habla. Los niños de dos años hablan acerca de lo que les gusta, desean y sienten, y también de esto mismo en relación a otras personas. A la edad de 3 años, también pueden hablar acerca de lo que los otros piensan y conocen (Bartsch & Wellman 1995).

Alrededor de los 4 años, tiene lugar un hito del desarrollo de *mindreading*, los niños se vuelven capaces de advertir que los pensamientos en la mente pueden no ser verdaderos. Por ejemplo, a partir de ese momento los niños se pueden percatar de que una caja de confites puede, en realidad, contener lápices. Si luego se les pregunta qué pensará un amigo que hay en el recipiente cuando mire dentro, los niños de 4 años advierten que el amigo será engañado, tal como lo fueron ellos, mientras que los niños de 3 años asumen que el amigo sabrá que hay lápices adentro, tal como ellos lo saben ahora (Perner, Leekman & Wimmer 1987). Los niños de 3 años tampoco pueden reportar que su propia creencia ha cambiado (Gopnik & Astington 1988). Si los lápices son guardados en la caja de confites, y luego se les pregunta qué pensaban que había en la caja antes de abrirla, estos dirán “lápices”, no “confites”. Los niños de 4 años, en cambio, recuerdan que su pensamiento fue “confites”.

Así, alrededor de los 4 o 5 años, los niños comienzan a entender que las personas actúan y hablan en base a lo que piensan acerca del mundo, aún cuando sus pensamientos no reflejen la situación real, y de este modo no se sorprenden cuando su amigo espera encontrar confites en la caja en la que ellos saben que hay lápices. De

modo que, los niños de 3 años saben que a diferentes personas pueden gustarles diferentes cosas, también advierten que los deseos y sentimientos pueden ser diferentes. A la edad de 4 o 5 años, los niños saben que las personas piensan cosas diferentes. Pueden entender que a veces las personas pueden creer algo que no es verdadero y, en tal caso, hacer o decir algo basadas en una creencia falsa. En líneas generales, se asume que el cambio más crítico en relación al desarrollo de *mindreading* tiene lugar alrededor de los 4 años y que luego sólo se complejiza con el aprendizaje y la experiencia de navegación en el mundo social.

En el ámbito del desarrollo se han propuesto una variedad de enfoques sobre cómo la capacidad de *mindreading* infantil emerge de procesos más tempranos y simples. Sin embargo, no voy a requerir que una propuesta de *mindreading* de cuenta de esto, en virtud de que los enfoques filosóficos hacen hincapié, más bien, en el sistema de *mindreading* de los adultos normales, un sistema maduro. No obstante, evaluaré las propuestas filosóficas de tendencia científico-cognitiva en relación a si pueden dar cuenta mínimamente de que existe un salto en el desarrollo y cierta asimetría entre la atribución de deseos y la atribución de creencias (la comprensión de los deseos parece preceder a la de las creencias). Considero adecuado asumir esto como un requisito en virtud del amplio consenso existente. Sin embargo, conservo cierta reserva en la medida en que este esquema de desarrollo ha sido desafiado por hallazgos recientes sobre la comprensión temprana de las creencias (Onishi & Baillargeon 2005; Southgate *et al.* 2007; Surian *et al.* 2007; Scott & Baillargeon 2009; Baillargeon *et al*. 2013), aunque aún no hay consenso respecto de cómo tienen que interpretarse estos resultados (Apperly 2011; Carruthers 2013).

Por su parte, hay evidencia significativa de que cuando los adultos llevan a cabo juicios sobre lo que otros sienten (por ejemplo, Van Boven & Lowestein 2003), creen o conocen (por ejemplo, Nickerson 1999), a menudo comienzan con sus propias sensaciones, creencias y conocimiento que luego ajustan a las del blanco, aunque con esfuerzo. Se trata de los sesgos egocéntricos en el razonamiento. Los adultos normales tienden a no cometer fallas en tareas de *mindreading* simples, tales como las que se

utilizan con niños y primates no humanos. Sin embargo, si estas tareas se modifican de modo que los juicios que tengan que realizar los adultos sean juicios probabilísticos o de incertidumbre, puede observarse la aparición de sesgos en el razonamiento.

En un estudio de Mitchell, Robinson, Isaacs & Nye (1996), los participantes observan videos donde *Sally* recibe un mensaje que contradice su creencia acerca del contenido del recipiente (por ejemplo, ella piensa que la taza tiene leche pero se le dice que contiene jugo). La tarea de los participantes consiste en juzgar si *Sally* cambiará de opinión respecto de lo que ella creía que era el contenido de la taza, en base al mensaje. Según este estudio, cuando los participantes saben que la creencia de *Sally* es verdadera juzgan que es menos probable que ella cambie de opinión, en comparación con la condición en la que saben que la creencia es falsa. Esto sugiere que los participantes permitieron que su conocimiento contaminara el juicio de *mindreading* acerca de la decisión de *Sally*. Este efecto es denominado "sesgo de realidad".

En otro estudio se arribó a conclusiones similares. Birch & Bloom (2007) adaptaron una tarea de falsa creencia de cambio de lugar de modo tal que los participantes tuvieran que juzgar la probabilidad con la que *Sally* buscaría en diferentes lugares. En la condición crítica, *Sally* siempre tiene una creencia falsa acerca de que el objeto se encuentra en determinado lugar, pero en algunos ensayos los participantes conocen el lugar particular en que está el objeto, mientras que en otros ensayos sólo saben que éste se encuentra en algún otro lugar. Llamativamente, cuando los sujetos conocen la ubicación específica juzgan que es menos probable que *Sally* busque incorrectamente en comparación con la condición en la cual los participantes no tienen certeza acerca de la ubicación del mismo. Los investigadores llaman a este efecto la "maldición del conocimiento".

Este y otros estudios con adultos (Nickerson 1999; Royzman, Cassidy & Baron 2003) sugieren la existencia de sesgos egocéntricos en *mindreading*. Indican que cuando *mindreading* se lleva a cabo bajo condiciones de incertidumbre las personas a menudo usan el atajo de asumir que "el otro es como uno" como punto de partida

para realizar ajustes. Por supuesto, esto no sugiere nada respecto de qué proceso subyace a tales ajustes, sólo indica que, en concordancia con una tendencia general bien documentada sobre sesgos egocéntricos en el razonamiento, estos también parecen estar presentes en el razonamiento de *mindreading*.

En base a la evidencia presentada considero que una teoría de *mindreading* de los estados mentales de otras personas (HA, por hetero-atribución) tiene que dar cuenta de que:

(HA1) Los adultos normales atribuyen percepciones, conocimiento, creencias, deseos, intenciones, decisiones, razonamientos a las otras personas.

(HA2) En los niños con desarrollo típico, hay un salto significativo en las habilidades de *mindreading* con posterioridad a los tres años.

(HA3) La comprensión de los deseos precede a la comprensión de las creencias en los niños pequeños.

(HA4) Los adultos normales llevan a cabo juicios de *mindreading* proyectando sus propias sensaciones, creencias y conocimientos.

3.2. Autoatribución de estados mentales

A diferencia de lo que ocurre en la investigación empírica, en filosofía se considera a la autoatribución y a las atribuciones a otras personas, como dos caras de una misma moneda. Dado que ningún teórico de la simulación ha sostenido que la simulación pueda intervenir en la autoatribución de estados mentales ocurrentes, aunque quizás pueda utilizarse en autoatribuciones de estados mentales pasados, hipotéticos o futuros (Goldman 2006: 24, 223), ninguna propuesta simulacional podrá dar cuenta de *mindreading* de primera persona. A su vez, las propuestas de teoría y simulación son, en rigor, respecto de *mindreading* de tercera persona. No obstante, al menos en el ámbito de la filosofía y tal como adelanté en la introducción, una teoría adecuada de *mindreading* debe poder dar cuenta de *mindreading* de primera persona.

En concordancia con esto, los enfoques híbridos de teoría (Nichols & Stich 2003) y simulación (Goldman 2006) proponen abordajes de *mindreading* de primera persona, que serán analizados y evaluados en los capítulos 4 y 5 respectivamente. Llamativamente, ambos enfoques proponen un abordaje de método dual, de introspección e interpretación, y en este sentido, considero que también pueden evaluarse en relación a si proporcionan o no un criterio satisfactorio para decidir en qué circunstancia se utiliza cada uno de los métodos propuestos para llevar a cabo las autoatribuciones. Antes de establecer las cuestiones mínimas sobre la autoatribución sobre las que todo enfoque de *mindreading* tiene que dar cuenta, me ocuparé sucintamente de la cuestión de la autoatribución, de la introspección y de los enfoques sobre la misma que son de interés para dar cuenta de *mindreading*.

En general, el sentido común nos provee de la intuición de que tenemos un acceso directo a nuestros estados mentales (Gopnik 1993; Nichols & Stich 2003; Goldman 2006; Carruthers 2009). Este acceso especial es diferente del acceso a los estados mentales de las otras personas en la medida en que no es preciso involucrarse en ningún tipo de interpretación. Contrariamente, al atribuir estados mentales a los otros parece necesario interpretarlos a la luz de sus circunstancias y comportamiento. Este acceso privilegiado, inmediato y no interpretativo a los estados mentales propios usualmente se denomina introspección. Desde Descartes (1641) en adelante, los filósofos han asumido este aspecto de las autoatribuciones de estados mentales y han intentado dar cuenta del mismo. En la filosofía se han planteado principalmente dos cuestiones ¿Cómo conocemos los estados mentales propios? ¿Qué tan bien los conocemos? (Schwitzgebel 2011). En principio, los enfoques filosóficos de *mindreading* de tendencia científico-cognitiva están particularmente interesados en la cuestión de si existe una facultad psicológica que permite el acceso directo a los estados mentales propios, más que respecto de la cuestión epistemológica de la confiabilidad de tal auto-conocimiento[34].

[34] En principio, considero que no es un requisito para una propuesta de *mindreading* en relación a la autoatribución que ésta aborde los debates filosóficos surgidos en torno al auto-conocimiento y al método especial que da cuenta del acceso privilegiado a los estados mentales propios. Recientemente,

Tradicionalmente, no todos los eventos mentales se consideran introspectibles. Las dos clases de estados que usualmente se consideran blancos de la introspección son, por un lado, las experiencias conscientes asociadas a las experiencias sensoriales, las imaginaciones y las emociones. Por el otro, las actitudes proposicionales tales como las creencias, los deseos, las intenciones, los juicios, las decisiones y demás (Schwitzgebel 2010; Carruthers 2009). Algunos enfoques de la introspección hacen hincapié en las actitudes proposicionales (Nichols & Stich 2003), mientras que otros en las experiencias conscientes (Goldman 2006), y no es claro que la introspección, si existe, constituya un proceso unitario en el sentido de que el mismo tipo de proceso esté implicado en todo lo que es introspectible.

El enfoque estándar del auto-conocimiento asume que se trata de un conocimiento infalible, acerca del que no podemos estar equivocados, y en tanto tal se auto-presenta (Descartes 1641). Más recientemente, los filósofos han sostenido que el conocimiento acerca de un conjunto de pensamientos es autorizado, en el sentido de que no puede ser desafiado por otros, y privilegiado, en el sentido de que se adquiere de un modo especial que no está disponible para conocer los estados mentales ajenos. No obstante, actualmente, la mayoría de los filósofos acepta que hay estados mentales que no son accesibles, tales como ciertos estados inconscientes, y que las autoatribuciones son falibles en el sentido de que podemos cometer errores al estilo de autoatribuirnos pensamientos ocurrentes manifiestamente falsos pero sin conciencia aparente de la falsedad de los mismos (confabular).

---

ha surgido la cuestión del vínculo entre el externalismo de contenido y la posibilidad de que las personas tengan un acceso privilegiado, o un conocimiento privilegiado de, sus actitudes proposicionales. El externalismo de contenido es la tesis de que la individuación de los contenidos de las actitudes proposicionales reside, en parte, en factores ambientales y, en este sentido, no proviene de las propiedades intrínsecas del individuo (Putnam 1975; Burge 1979). Esto constituye un problema para el acceso privilegiado en la medida en que un sujeto no podría conocer el contenido de sus pensamientos puesto que el mundo externo es inaccesible a la introspección. No obstante, algunos han defendido la compatibilidad del externalismo de contenido y el acceso privilegiado (McLaughlin & Tye 1998). Sin embargo, estas discusiones resultan ortogonales a la cuestión de los procesos subyacentes a *mindreading* (Nichols & Stich 2003: 152; Goldman 2006: 228). La pregunta relevante, en este caso, es por la existencia de un método especial y su naturaleza, o en términos más cognitivos, la cuestión de si existe una facultad en la psicología humana que permite el acceso directo, no interpretativo, a las propias experiencias, juicios y demás eventos mentales (Engelbert & Carruthers 2011).

No obstante, se ha cuestionado la existencia de la introspección y esta cuestión tiene especial importancia para los enfoques de *mindreading*. Entre los psicólogos y filósofos de tendencia científico-cognitiva existen varias posturas. Algunos enfoques de la autoatribución asumen la existencia de la introspección y proponen enfoques psicológicamente plausibles del auto-conocimiento (Nichols & Stich 2003; Goldman 2006), otros enfoques niegan que exista la introspección y asumen que se usan los mismos recursos que en la atribución de tercera persona (Gopnik 1993; Gopnik & Meltzoff 1994; Carruthers 2009). Finalmente, hay posturas intermedias, sostenidas principalmente por científicos, que postulan la existencia de un mecanismo de auto-conocimiento aunque estas autoatribuciones se consideran sistemáticamente no confiables (Gazzaniga 1995; Wegner 2002; Wilson 2002). En relación a las posturas mencionadas, en éstas pueden reconocerse dos enfoques sobre la introspección, los enfoques del auto-monitoreo y de la teoría de la teoría[35].

Básicamente, en el enfoque del auto-monitoreo se postula que los sistemas cognitivos poseen un escáner cuya función es detectar la presencia de creencias y producir como *output*, creencias, juicios o representaciones de tales creencias. Si bien este enfoque admite una serie de variantes, se asume que, por ejemplo, uno posee una "caja de creencias" en la mente, de modo que creer que *P* implica tener una representación con el contenido *P* en la "caja de creencias". Normalmente, el proceso de monitoreo hace una copia de la creencia alojada en la "caja de creencias", le agrega

[35] En la filosofía, la introspección se ha concebido de distintas maneras. Además de los enfoques de auto-monitoreo y de "teoría de la teoría", Schwitzgebel (2011) menciona los enfoques de la transparencia y de constitución parcial. El enfoque de la transparencia sostiene que no es necesario escanearse a uno mismo para descubrir la presencia o ausencia de un estado interior pre-existente de creencia de que *P*, más bien pienso o considero los hechos relevantes en relación a *P* en el mundo externo, y considero si *P* es verdadero. Este enfoque tiene variantes, entre ellas, las rutinas de ascenso (Evans 1982) mencionadas en el capítulo 2, y la propuesta de Gordon (1995a, 2007), entre otros. El enfoque de constitución parcial afirma que parte de creer que *P* es justamente estar dispuesto a auto-adscribirse la creencia de que *P*. De modo que la creencia de que *P* y la disposición a auto-adscribirse la creencia de que *P* no son ontológicamente distintas, tal como parecen asumir otros enfoques (Shoemaker 1996, 2009; Schwitzgebel 2009, 2011). En principio, el enfoque disposicional no se pronuncia respecto de los mecanismos cognitivos subyacentes de modo que sería compatible con distintos procesos (Schwitzgebel 2011: 48). Por su parte, el enfoque de la transparencia no parece ser compatible con un proceso de tipo teórico (en el que medien inferencias a partir de una teoría psicológica y de información acerca del sujeto y su circunstancia), de hecho, Gordon (1995a, 2007) propone un proceso subyacente de tipo simulacional.

el prefijo actitudinal "yo creo que" y genera como *output* la creencia de segundo orden "yo creo que *P*", que coloca nuevamente en la "caja de creencias" y así, permite saber qué creencias se poseen (Nichols & Stich 2003).

En el enfoque de la "teoría de la teoría" se asume que el acceso a la propia mente depende de los mismos procesos y recursos cognitivos que subyacen a la capacidad de atribuir estados mentales a otras personas. Entre estos recursos se cuentan una teoría o cuerpo de conocimiento sobre la mente, información disponible perceptualmente respecto del blanco y su circunstancia, e información almacenada en la memoria sobre el blanco y su circunstancia. De modo que, a pesar de que la autoatribución parece ser una cuestión tan simple y directa, es pura ilusión. En realidad, la autoatribución es consecuencia de un proceso de interpretación o inferencia y no hay un acceso directo a los estados mentales propios (Gopnik 1993; Gopnik & Wellman 1994; Gopnik & Meltzoff 1994; Carruthers 1996, 2009, 2011). Este enfoque es tradicional en la psicología del desarrollo y ha encontrado apoyo principalmente en dos tipos de evidencia. Por un lado, en evidencia que sugiere un desempeño infantil paralelo en tareas de *mindreading* y de metacognición (Gopnik 1993; Gopnik & Wellman 1994) y, por otro lado, por la ausencia de evidencia contundente sobre la existencia de una asimetría yo-otro en el desarrollo (Wellman *et al.* 2001)[36]. En base a esta evidencia, el enfoque de la "teoría de la teoría" niega rotundamente la existencia de la introspección a excepción de cierto "zumbido cartesiano en la cabeza" que permitiría detectar cierta actividad cognitiva interna, aunque no queda claro en qué pueda consistir esto.

En virtud de la negación de la existencia de la introspección, este enfoque carga con el peso de tener que dar cuenta de la intuición de sentido común del conocimiento directo de los estados mentales propios. Con este propósito, Gopnik (1993) recurre a la metáfora del experto. La idea es que, en la adultez, se ha acumulado experiencia en

[36] En psicología el término "metacognición" refiere a la capacidad de saber acerca de los procesos cognitivos. Por ejemplo, la metamemoria consiste en el conocimiento que podemos tener de las estrategias que usamos para memorizar. "Metacognición" también hace referencia al conocimiento que podemos tener de las estrategias que utilizamos para resolver problemas o aprender. Las medidas de metacognición suelen utilizarse como indicadores de la autoreflexión.

relación a la detección de actividad cognitiva interna (zumbido cartesiano) y evidencia contextual y comportamental apropiada de modo que esto convierte al adulto en un experto. Como consecuencia, posiblemente el adulto aplica el aparato teórico de "Teoría de la Mente" y realiza inferencias sobre sus estados mentales con muy poca conciencia de las mismas, e interpreta tales experiencias cargadas de teoría como percepciones directas de sus estados psicológicos, tal como el ajedrecista experto "ve" la posición débil de la reina. No obstante, los defensores de la introspección señalan que aún concediendo esto, hay otras intuiciones que este enfoque no puede explicar. A saber, nos parece obvio que uno puede estar tranquilamente sentado, sin exhibir comportamiento manifiesto relevante y, no obstante, ser capaz de reportar sus propios pensamientos (Nichols & Sitch 2003)[37].

Hasta aquí, hemos visto un panorama sucinto de las cuestiones que se debaten, en el ámbito de *mindreading*, respecto de la autoatribución. A continuación, me ocuparé de aquellas cuestiones en las que hay un acuerdo considerable respecto a ciertos rasgos asociados a la autoatribución. Si bien considero que no es un requisito de una propuesta de *mindreading* sobre la autoadscripción intentar dar cuenta de todo

[37] Carruthers (2009) niega la existencia de un método especial para acceder a las actitudes proposicionales, pero tiene una respuesta para la objeción a Gopnik (1993) señalada por Nichols & Stich (2003). Si bien no hay introspección de actitudes proposicionales, hay conocimiento directo de los estados mentales imaginativos, perceptuales y somato-sensoriales. Esta propuesta de introspección para los estados perceptuales está basada en la noción de que la información perceptual y cuasi-perceptual (*i.e.* imaginativa) que es atendida se "transmite globalmente" a los sistemas de toma de decisión y de formación de creencias (Baars 1988, 1997). Si concedemos esto a Carruthers (2009), el sistema de *mindreading* tendría acceso a esta información y podría utilizar directamente los estados perceptuales que recibe como *inputs* para producir atribuciones del tipo "yo veo rojo", "yo tengo dolor", "yo tengo hambre". Contrariamente, el sistema de *mindreading* no puede dar cuenta de las autoatribuciones de actitudes proposicionales porque los *outputs* de los sistemas formadores de creencias y de toma de decisiones no se transmiten globalmente. En su lugar, el sistema de *mindreading* debe acudir a los datos perceptuales disponibles para inferir las actitudes proposicionales ocurrentes (tal como en el caso de *mindreading* de tercera persona). En este sentido, Carruthers (2009) considera que puede explicar lo que Gopnik (1993) no puede. Somos capaces de atribuirnos estados mentales a nosotros mismos aunque no estemos involucrados en comportamiento manifiesto, porque alguien sentado e inmóvil aún tiene información abundante acerca de su situación ocurrente bajo la forma de datos sensoriales, imaginativos y somato-sensoriales. Por esto, es capaz de autoatribuirse estados mentales. No obstante, no queda claro que esta propuesta pueda dar cuenta de todos los casos de autoatribuciones, particularmente de los auto-reportes que sugieren que existe el pensamiento sin símbolos o el pensamiento proposicional que no es acompañado por ningún tipo de imagen mental (Hulburt 1990; Hulburt & Akhter 2008).

el problema del auto-conocimiento (ver nota 2), no obstante, las propuestas de autoatribución, mínimamente, tienen que dar cuenta de dos cuestiones sobre las que existe un amplio acuerdo. En primer lugar, de la intuición de sentido común de conocimiento directo de los estados mentales propios mencionada anteriormente, basada en la observación de que los adultos normales a menudo llevan a cabo la autoatribución fácilmente, aunque no tienen una idea clara de cómo lo hacen. En principio, esto no es un problema para aquellos enfoques que postulan la existencia de un método especial de conocimiento de los estados mentales propios, sino para aquellos enfoques que lo niegan. Tal como mencioné, estos llevan la carga de tener que explicar la intuición de sentido común.

En segundo lugar, los enfoques de *mindreading* sobre la autoatribución tienen que dar cuenta de la confabulación. Se ha acumulado evidencia, principalmente de psicología social, que sugiere que las personas a menudo confabulan respecto de sus pensamientos ocurrentes y del pasado reciente. Esto es, reportan pensamientos manifiestamente falsos pero sin conciencia aparente de la falsedad de los mismos. En este sentido, hay acuerdo sobre el fenómeno de la confabulación y de que éste surge de la auto-interpretación. De modo que esto será, en principio, un problema para aquellos enfoques que postulan el acceso directo a los estados mentales propios y niegan la interpretación de los mismos.

La literatura de psicología social sobre la “disonancia” muestra numerosos ejemplos de confabulación en relación a los juicios. En el estudio de Linder *et al*. (1967), los participantes son inducidos a escribir un ensayo defendiendo una posición que no comparten. Este estudio muestra que, al finalizar el experimento, los participantes se mostraron más compasivos respecto de la posición a la que no adherían, en dos condiciones. A saber, cuando los sujetos recibieron un pago escaso en comparación con los participantes que recibieron un incentivo razonable, y cuando decidieron libremente participar del experimento en comparación con los participantes que lo hicieron por obligación. Este hallazgo se interpretó de la siguiente manera. Si una persona es inducida a comportarse públicamente de una manera que no coincide

con sus actitudes privadas, experimentará disonancia cognitiva. De modo que, al no poder apelar a razones de peso tales como el incentivo o la obligación, las personas cambiaron sus actitudes respecto de la posición que no compartían para reducir la magnitud de la disonancia cognitiva. En términos de autoatribución, estos datos sugieren que los sujetos intentaban explicar su comportamiento disonante (escribir un ensayo a favor de una posición que no sostenían) y al no poder recurrir a razones significativas para su interpretación moderaron su actitud hacia la postura antipática. Estos resultados asociados a la disonancia cognitiva se han replicado numerosas veces.

Otra serie de estudios sugiere que hay auto-interpretación de los juicios. El estudio de Wells & Petty (1980) muestra que los sujetos que mueven su cabeza con un movimiento arriba-abajo (asentir) mientras escuchan un mensaje persuasivo lo consideran más convincente que aquellos que lo escuchan sacudiendo la cabeza (negar). En un estudio posterior se complejizó el hallazgo (Briñol & Petty 2003). En este caso, algunos sujetos escuchan argumentos convincentes para una proposición mientras que otros escuchan argumentos irrelevantes. En concordancia con el estudio anterior, los sujetos que escucharon argumentos convincentes los encuentran más convincentes si mueven la cabeza asintiendo que si la sacuden negando. La novedad reside en que los sujetos que escuchan los argumentos débiles, los creen más al sacudir la cabeza (negar) que al moverla en sentido arriba-abajo (asentir). La explicación propuesta postula la interpretación de los pensamientos. Al parecer, al percibir sus pensamientos los sujetos interpretan sus movimientos manifiestos de cabeza como expresando acuerdo o desacuerdo y esto afecta la confianza en la validez del propio pensamiento y no la confianza en la validez del mensaje persuasivo. El movimiento de la cabeza fue interpretado como acuerdo/desacuerdo con el mensaje. De modo que, si el mensaje era convincente y se asentía estaban más de acuerdo que si se negaba, y si el mensaje no era convincente y se negaba con la cabeza, se lo creía más que si se asentía. Esto sugiere un efecto de interpretación de los propios pensamientos en relación con movimientos manifiestos propios.

Los efectos de la confabulación también se han observado en relación a las decisiones. En el estudio de Brasil-Neto *et al.* (1992) se pidió a los sujetos que decidieran levantar el dedo índice derecho o izquierdo al escuchar el sonido de un *click*. Sin que lo supieran los participantes, el sonido era causado por el aparato de estimulación magnética transcraneana, que los experimentadores utilizaron para causar los movimientos de dedos de los sujetos mediante estimulación de áreas motoras de la corteza. A pesar de que los movimientos eran causados por este tipo de estimulación, los participantes reportaron haber tomado la decisión de mover el dedo. En este sentido, los datos sugieren que los sujetos interpretan (falsamente) sus decisiones a partir información perceptual asociada a su comportamiento manifiesto.

Otros estudios que proveen evidencia de la confabulación son los estudios con pacientes con cerebro dividido (Gazzaniga 1995; Baynes & Gazzaniga 2000). Como tratamiento de la epilepsia, estos pacientes han sido sometidos a una cirugía que corta el cuerpo calloso (comisurotomía) y separa los hemisferios cerebrales, dejándolos incomunicados. De modo que los estímulos pueden ser presentados a los hemisferios por separado. Cuando un estímulo se presenta al hemisferio derecho, éste solo puede producir respuestas manuales, pero no puede responder verbalmente, dado que usualmente el hemisferio izquierdo controla la producción del habla. Por su parte, el hemisferio izquierdo es ciego a los estímulos presentados al derecho. En determinado estudio, distintos estímulos visuales fueron presentados a cada hemisferio. Una garra de pollo al izquierdo y una escena nevada al derecho. Posteriormente, se presentaron imágenes ante la vista de ambos hemisferios y se pidió al paciente que eligiera un ítem asociado. El paciente eligió con la mano izquierda (conectada con el hemisferio derecho) la imagen de una pala y con la mano derecha (conectada con el hemisferio izquierdo) la imagen de un pollo. De modo que figuradamente puede decirse que el hemisferio izquierdo emparejó la garra de pollo y el hemisferio derecho la escena nevada. No obstante, cuando el paciente explicó por qué había elegido tales ítems, sostuvo: “Es simple, la garra de pollo va con el pollo y la pala se precisa para limpiar el gallinero” (Gazzaniga 1995). Según Gazzaniga, puesto que el hemisferio izquierdo no

tenía información respecto de la escena nevada, generó una historia que conectara la pala con el pollo. Este tipo de hallazgos han sugerido a los investigadores la existencia de un "intérprete" en el hemisferio izquierdo cuya función es generar explicaciones del comportamiento del agente, generando teoría plausibles, aunque tales interpretaciones pueden ser erróneas (confabulaciones).

Por su parte, Wegner & Wheatley (1999) pidieron a los participantes que reportaran sus intenciones en un experimento en el que fueron inducidos a creer que controlaban el cursor en una pantalla de computadora, conjuntamente con otro sujeto (que de hecho era un cómplice en el experimento). Inmediatamente después de cada ensayo los participantes debían registrar en qué medida habían tenido la intención de colocar el cursor en la posición final (en una escala del 1 "dejé que parara" al 100 "quise que parara"). En una de las condiciones, el cómplice no tenía que participar del movimiento del cursor, otorgándoles a los participantes el control total. Llamativamente, en esta condición, el promedio de los participantes midió su grado de intención en 56, sólo 6 puntos por encima del punto medio. Probablemente, esto pueda explicarse apelando a la posibilidad de que los sujetos hayan supuesto razonablemente que, en virtud de que el control es compartido, éste se ancla en el punto medio de la escala y sólo se ajusta levemente hacia arriba en ocasiones de control completo. Sin embargo, esta explicación apela a cierta interpretación basada en teoría. Si, en cambio, los sujetos hubieran tenido un acceso directo a sus decisiones, es esperable que pudieran evaluar el control total de la causalidad en tales circunstancias. Es sabido que los eventos contiguos en el tiempo otorgan a las personas un sentido fuerte de causación. De modo que, la ausencia de tal efecto, en este caso, sugiere que el acceso a las intenciones propias es interpretativo y no directo.

En conclusión, en base a ciertas intuiciones sobre las que hay un amplio consenso y a la evidencia presentada en relación a la confabulación, considero que una teoría de *mindreading* sobre la autoatribución (AA) tiene que, mínimamente, dar cuenta de:

(AA1) La intuición de sentido común del conocimiento directo de los estados mentales propios.

(AA2) A menudo, las personas interpretan (falsamente) sus pensamientos ocurrentes y del pasado reciente. Esto es, reportan pensamientos manifiestamente falsos pero sin conciencia aparente de la falsedad de los mismos.

3.3. Un criterio para determinar el tipo de proceso subyacente

Como ya he mencionado, desde la perspectiva de la investigación empírica existe un amplio acuerdo respecto de que son varios procesos o mecanismos cognitivos que subyacen a *mindreading*. En concordancia con esto, los enfoques híbridos de teoría y simulación postulan dos tipos de procesos subyacentes a *mindreading*. No obstante, es preciso que estos enfoques híbridos puedan decir algo más. Puesto que un enfoque que simplemente afirma que a veces la simulación subyace a *mindreading* y que a veces la teoría es el proceso subyacente, no permite realizar predicciones de ningún tipo, por ejemplo, sobre cuándo ocurren errores sistemáticos en *mindreading* (que al parecer sólo se pueden explicar por teoría). En este sentido, se requieren ciertas especificaciones sobre las circunstancias en las que cada tipo de proceso está implicado. Esto es, es preciso un criterio para determinar el tipo de proceso subyacente a *mindreading* de tercera persona.

En principio, los requisitos anteriores de hetero y autoatribución pueden aplicarse a cualquier tipo de enfoque de *mindreading*. En cambio, este requisito del criterio de distinción de procesos es particular para aquellos enfoques que postulan la existencia de varios tipos de procesos o mecanismos subyacentes a *mindreading*, tal como es el caso de los enfoques híbridos de teoría y simulación. Considero que este criterio es pertinente en relación a los enfoques híbridos que voy a evaluar en esta tesis (Nichols & Stich 2003; Goldman 2006) en la medida en que ambos no sólo postulan dos procesos, teoría y simulación, subyacentes a *mindreading* de tercera

persona sino que, a su vez, proponen un proceso dual de introspección e interpretación para *mindreading* de primera persona. En este último caso, el proceso de interpretación viene a dar cuenta de datos empíricos que sugieren que las personas a menudo confabulan respecto de sus pensamientos ocurrentes y del pasado reciente. Esto es, reportan pensamientos manifiestamente falsos pero sin conciencia aparente de la falsedad de los mismos. La idea de incluir este requisito se basa en la convicción de que proponer una teoría híbrida no consiste meramente en juntar dos cosas que funcionan bien (o no tanto).

Considero que los requisitos mencionados son fundamentales. Los requisitos de hetero y autoatribución responden a la cuestión de que las atribuciones de primera y de tercera persona son dos aspectos de una misma competencia, a saber, la atribución mentalista. El requisito sobre el criterio para distinguir entre procesos subyacentes resalta la necesidad de que los enfoques híbridos puedan generar predicciones adecuadas. Para esto es necesario que provean un criterio claro para determinar qué tipo de proceso postulado, entre los postulados, subyace a una instancia de *mindreading*.

4. Conclusión

En este capítulo he considerado a *mindreading* como el conjunto de capacidades cognitivas de atribución de estados mentales, propios y ajenos, y de explicación y predicción de estados mentales y comportamiento en términos mentalistas. Además, he considerado que la cuestión principal en relación a *mindreading* es la determinar cuáles son los procesos cognitivos subyacentes. En este sentido, se han visto las motivaciones para los enfoques híbridos de teoría y simulación para *mindreading* de tercera persona, y los enfoques duales de interpretación e introspección para *mindreading* de primera persona (estos últimos pueden considerarse híbridos en la medida en que postulan más de un proceso subyacente).

Además, he propuesto tres requisitos fundamentales o de mínima con el propósito de evaluar enfoques híbridos de teoría y simulación particulares (Nichols &

Stich 2003, Goldman 2006). En base a la evidencia mencionada en relación con *mindreading* de los estados mentales de otras personas (sección 3.1), considero que un enfoque sobre *mindreading* tiene que dar cuenta, al menos, de que (HA1) los adultos normales atribuyen estados mentales a las otras personas, (HA2) alrededor de los 4 años hay un salto significativo en el desarrollo de *mindreading*, (HA3) la comprensión de los deseos precede a la comprensión de las creencias, (HA4) las personas llevan a cabo juicios de *mindreading* proyectando sus propias sensaciones, creencias y conocimientos.

Asimismo, en base a la evidencia presentada en relación a la confabulación y a ciertas intuiciones sobre las que hay un amplio consenso, considero que una teoría de *mindreading* sobre la autoatribución tiene que, mínimamente, dar cuenta de: (AA1) la intuición del conocimiento directo de los estados mentales propios y (AA2) de que a menudo, las personas interpretan (falsamente) sus pensamientos ocurrentes. Esto es, reportan pensamientos manifiestamente falsos pero sin conciencia aparente de la falsedad de los mismos. Paralelamente, tal como mencioné, la cuestión principal en *mindreading*, para un enfoque híbrido, es determinar cuáles son los procesos cognitivos subyacentes; de manera de ofrecer una explicación que permita realizar predicciones adecuadas (el tercer requisito).

CAPÍTULO 4. UN ENFOQUE HÍBRIDO DE *MINDREADING* CON ÉNFASIS EN LA TEORÍA

1. Introducción

Nichols & Stich (2003) definen *mindreading* como la capacidad del común de la gente para entender la mente y sostienen que las distintas teorías de *mindreading* dicen muy poco sobre los mecanismos mentales que éstas presuponen y sobre cómo se supone que estos mecanismos explican los aspectos importantes de *mindreading.* En virtud de esto, su propósito es caracterizar las habilidades complejas y variadas que constituyen la capacidad de *mindreading* y comenzar el trabajo de explicar cómo estas habilidades se llevan a cabo. En este sentido, los autores brindan una explicación de *mindreading* postulando mecanismos subyacentes, caracterizados funcionalmente, con capacidades que son más simples que la capacidad que se trata de explicar. Para esto, postulan "cajas" que aluden a mecanismos de procesamiento caracterizados funcionalmente o bien a un conjunto de estados mentales caracterizados funcionalmente, sin establecer compromiso alguno con una ubicación espacial en el cerebro para estas cajas[38].

En este capítulo presentaré el enfoque híbrido de Nichols & Stich (2003) para *mindreading* de tercera persona, en el cual algunos aspectos de *mindreading* se explican por simulación y otros por procesos ricos en información (Nichols & Stich 2003: 132). A diferencia del carácter híbrido de la propuesta de tercera persona, en el caso de *mindreading* de primera persona, postulan un mecanismo dual de introspección e interpretación. A la base de ambas propuestas de *mindreading* se supone una arquitectura cognitiva básica y un enfoque representacional de los estados mentales. La arquitectura cognitiva básica consta de dos tipos de estados representacionales, creencias y deseos, que difieren funcionalmente entre sí en tanto

[38] Según los autores, los hallazgos en la estructura y el funcionamiento del cerebro pueden y, en última instancia lo harán, imponer restricciones fuertes sobre teorías de *mindreading* de este tipo, sin embargo, consideran que este tipo de investigaciones no han hecho aún aportes directos a las preguntas que están considerando (Nichols & Stich 2003).

son causados de diferentes maneras e interactúan de distinto modo con los otros componentes de la mente. Las creencias, los deseos y otras actitudes proposicionales son considerados estados relacionales. De modo que tener una creencia o un deseo con un contenido particular es tener una instancia de representación con un contenido almacenado funcionalmente de modo apropiado en la mente. Esto es, tener la creencia de que "Sócrates es mortal" es tener una instancia de representación cuyo contenido es "Sócrates es mortal" almacenado en la "caja de creencias". Los procesos perceptuales y los mecanismos inferenciales generan creencias, mientras que los deseos son generados por sistemas de monitoreo de estados corporales o por el sistema de toma de decisiones, en tanto metas o deseos intermedios. Además, se asume la existencia de otros mecanismos generadores de deseos que aún no se comprenden.

En la sección 2, presentaré la propuesta de los mecanismos subyacentes a *mindreading* de tercera persona que están repartidos en un sistema temprano y otro tardío. Enumeraré los mecanismos componentes de cada sistema, describiré su función y cómo se supone que interactúan entre sí para llevar a cabo *mindreading*. Además, en este esquema quedará claro el aporte del surgimiento de un sistema tardío, que se construye sobre el temprano. En la sección 2.1, señalaré qué aspectos de la propuesta se consideran de tipo simulacional y cuáles de tipo ricos en información con el propósito de aclarar el carácter híbrido de la misma, que no se aprecia claramente en la descripción de los mecanismos. En la sección 2.2, analizaré la propuesta según los requisitos para *mindreading* de las otras personas (HA1-HA4) establecidos en el capítulo 3 (sección 3.1). Intentaré mostrar que los requisitos (HA1) y (HA2) se satisfacen, mientras que el requisito (HA3) no se satisface y el (HA4) se satisface parcialmente. En la sección 2.3, analizaré el criterio de corrección propuesto por Stich & Nichols (2003) y Nichols & Stich (2003) para distinguir el tipo de proceso subyacente a un caso de *mindreading*. Según este criterio, si el resultado de *mindreading* es correcto es probable que el proceso subyacente sea de tipo simulacional, si el resultado de *mindreading* es incorrecto es probable que el proceso subyacente sea de

tipo rico en información. En las secciones 2.3.1 y 2.3.2 discutiré este criterio y concluiré que no es satisfactorio, dado que básicamente los argumentos presentados no permiten descartar, en ninguno de los dos casos, explicaciones basadas en el proceso contrario.

En la sección 3 presentaré la propuesta de un método dual de introspección e interpretación para *mindreading* de primera persona. En la sección 3.1, analizaré la propuesta según los requisitos para la autoatribución de estados mentales y el criterio para distinguir entre procesos subyacentes, establecidos en el capítulo 3 (secciones 3.2 y 3.3), respectivamente. Intentaré mostrar que el requisito (AA1) para la autoatribución se satisface pero el requisito (AA2) no. Además, la evidencia sobre la confabulación sugiere, en contra Nichols & Stich (2003), que hay interpretación de estado mentales ocurrentes. Si esto es así, el criterio propuesto para distinguir entre introspección e interpretación no parece funcionar porque no da cuenta de todos los casos de confabulación, y el requisito de proponer un criterio claro para distinguir entre los procesos subyacentes a *mindreading* no se satisface.

2. Los mecanismos subyacentes a *mindreading* de tercera persona

En esta sección reconstruiré la propuesta de *mindreading* de tercera persona de Nichols & Stich (2003) que, a mi entender, está claramente motivada por los casos a favor para los enfoques de teoría y simulación mencionados en el capítulo 3 (sección 2.2), más que por cierta complementación entre teoría y simulación. Es más, en la sección 2.3, mostraré cómo en la propuesta de Nichols & Stich (2003) estos casos favorables configuran el criterio para determinar el tipo de proceso subyacente a un caso de *mindreading* y cuestionaré este criterio. Antes, en la sección 2.1, señalaré qué aspectos en la propuesta se consideran procesos de tipo simulacional o ricos en información para aclarar el carácter híbrido de la misma. En la sección 2.2, evaluaré la propuesta en relación a los requisitos para *mindreading* de los estados mentales de otras personas, propuestos en el capítulo 3 (sección 3.1).

La capacidad de *mindreading* de tercera persona está compuesta por dos mecanismos subyacentes que, según Nichols & Stich (2003), son diferenciables filogenética y ontogenéticamente. El sistema temprano, y más antiguo, permite predecir el comportamiento de los otros atribuyendo metas o deseos y definiendo la mejor estrategia para alcanzarlos. El sistema tardío utiliza los mecanismos subyacentes al sistema temprano pero con la particularidad de reclutar, además, sistemas componentes de la capacidad cognitiva de ficción. Esto le otorga a *mindreading* un mayor poder predictivo en la medida en que es posible generar un modelo de las creencias del blanco para realizar atribuciones mentalistas, tal como veremos a continuación.

El sistema temprano está compuesto por tres tipos de mecanismos: los mecanismos detectores de deseos, el mecanismo planificador y el mecanismo coordinador de *mindreading*. Los mecanismos detectores de deseos, o metas (en adelante MDD) son un conjunto de estrategias que utilizan señales variadas para determinar las metas del blanco. Un primer grupo de estrategias utilizan señales de comportamiento no verbal (por ejemplo, la dirección de la mirada, los ojos abiertos o cerrados). Posiblemente, se trate de estrategias innatas que, a su vez, pueden enriquecerse con el aprendizaje de señales conductuales asociadas a los distintos tipos de metas. Otras estrategias recurren a las expresiones faciales como señales para determinar qué quieren los otros. Quizás, en este caso, se trata del sistema innato que forma parte del mecanismo que subyace a la atribución de emociones basada en expresiones faciales (Ekman 1992). Asimismo, hay estrategias basadas en señales verbales. Por un lado, lo que el propio blanco dice que son sus deseos y, por otro lado, lo que los otros dicen sobre los deseos del blanco. Otra estrategia posible consiste en generalizar sobre el propio caso y asumir que los otros tienen los mismos deseos que uno.

Ahora bien, una vez que los MDD generan una creencia acerca de los deseos del blanco, entra en acción el planificador. Éste es un submecanismo de los mecanismos inferenciales. Tiene una función general en la cognición de permitir a los organismos

decidir entre los cursos de acción. En el caso específico de *mindreading*, el planificador es convocado para determinar cómo las metas o deseos del blanco pueden ser satisfechos. No obstante, el planificador típicamente determinará un plan desde el punto de vista del *mindreader* puesto que, en el sistema temprano, el planificador aún no es sensible al hecho de que el blanco puede tener creencias discrepantes (Nichols & Stich 2003: 81). Esta situación puede conducir a predicciones incorrectas.

El tercer mecanismo, el coordinador de *mindreading*, cumple una función que no realizan los dos sistemas antes mencionados. Éste genera las predicciones sobre el comportamiento futuro del blanco. Además, lleva a cabo la tarea de coordinar el proceso. Primero, recopila información sobre los deseos y metas del blanco, y pone en acción a aquellos MDD que no se activen automáticamente. Una vez que se generan creencias acerca de las metas del blanco, las recolecta de la caja de creencias y las envía al planificador, encargado de generar el mejor curso de acción para alcanzar la meta según la situación del blanco. Una vez que el planificador reporta el plan de acción, el coordinador genera la creencia de que el blanco actuará según el plan y esta creencia se usará para predecir el comportamiento del blanco (Nichols & Stich 2003: 81). Además, el coordinador tiene un segundo rol, que consiste en generar creencias acerca de deseos instrumentales. Cuando el planificador presenta al coordinador un plan que no sólo refiere a la meta sino que incluye pasos intermedios, éste genera creencias sobre tales deseos o metas, de modo tal que pueda considerarse su estatus intermedio o instrumental.

Hasta aquí el sistema temprano presente en niños muy pequeños (antes de los 3 años), pone a disposición una estrategia para predecir el comportamiento[39]. Esta estrategia está basada en la detección de deseos o metas y en establecer un plan, el mejor, para satisfacerlas. Es preciso mencionar que se asume que los mecanismos intervinientes son ricos en información. La capacidad de detectar o reconocer metas

[39] Según Nichols & Stich (2003) cierta evidencia empírica brinda apoyo a su propuesta acerca del sistema temprano de *mindreading*. Se ha reportado la atribución de deseos a partir de gestos a los 18 meses (Repacholi & Gopnik 1997), la atribución de metas a partir de acciones incompletas a los 18 meses (Meltzoff 1995) y la atribución de metas a los 12 meses (Gegerly, Csibra *et al*. 1995).

implica el concepto de meta y cierta teoría sobre las mismas. A saber, que las metas son cosas que los organismos tienen y que cada meta está vinculada con una acción o estado de cosas que, cuando ocurren, satisfacen la meta. Además, que cuando el organismo tiene una meta puede exhibir una serie de patrones de comportamiento que permiten alcanzar la meta (Nichols & Stich 2003: 62)[40]. Asimismo, según Nichols & Stich (2003: 136), los MDD son procesos ricos en información en la medida en que los seres humanos cometemos fallas sistemáticas en la atribución y predicción de deseos y preferencias, y la mejor explicación disponible de las fallas sistemáticas implica considerarlas como predicciones erróneas generadas por teorías equivocadas o incompletas. Nótese que este es el argumento que reiteradamente van a esgrimir para sostener que un mecanismo subyacente es rico en información.

Como mencioné, el sistema temprano puede conducir a predicciones erróneas en la medida en que el plan para satisfacer los deseos se genera desde el punto de vista del *mindreader*. Básicamente, el sistema planificador no es sensible a la presencia de creencias discrepantes en el blanco. Ahora bien, esta dificultad se irá remediando a medida que se vaya complejizando la información sobre el blanco y su situación, de modo de poder generar un plan para alcanzar una meta desde el punto de vista del blanco. En esto consistirá la mejora que aporta el sistema tardío. Éste está constituido por el sistema temprano con el agregado de poder reclutar a la "caja de mundos posibles" (en adelante CMP) para llevar a cabo *mindreading*. Nichols & Stich (2003) especulan que la CMP ha surgido en la cognición para llevar a cabo el razonamiento hipotético. Según los autores, un organismo debe estar dotado con un mecanismo que le permita pensar qué pasaría si lleva a cabo determinada acción. En esto consiste la función evolutiva original de la CMP y el razonamiento hipotético se lleva a cabo de la siguiente manera. En la CMP se coloca una representación del tipo "yo hago A". Luego, todas las creencias alojadas en la caja de creencias se vuelcan a la CMP y el

[40] Como es sabido, la capacidad de anticipar conducta no siempre implica atribuir estados internos al organismo, ni una teoría acerca de los mismos. En este sentido, es posible asociar señales estereotípicas con patrones de conducta. Por ejemplo, si el contrincante muestra los dientes posiblemente de pelea. Esta anticipación de comportamiento no implica atribuir meta alguna, simplemente se basa en la asociación de una señal con un comportamiento.

actualizador las modifica según la representación "yo hago A". De modo que de este proceso resulta una descripción del mundo que se utiliza para saber qué pasa si "yo hago A". Estas representaciones quedarán alojadas en al CMP mientras dure la tarea. Así, la CMP elabora descripciones del mundo que son diferentes de las descripciones en la caja de creencias. En este sentido, la CMP es concebida como un espacio de trabajo donde nuestro sistema cognitivo construye y almacena temporalmente representaciones de uno u otro mundo posible.

Es preciso notar que la CMP contiene instancias de representaciones, tal como sucede con las cajas de creencias y deseos, aunque estas representaciones son de un tipo especial. Sus roles funcionales y el patrón de interacción con el resto de los componentes de la mente difiere de los roles funcionales de las creencias y los deseos. En este caso, las representaciones tienen la función de representar el mundo en base a un conjunto de presupuestos que pueden no creerse como verdaderos, ni querer que sean verdaderos. De modo que la CMP no representa el mundo tal como es, ni tal como quisiéramos que fuera[41]. Como ya he mencionado, la CMP junto al planificador construye planes que pueden llevarse a cabo en una variedad de situaciones posibles. Particularmente, en la ficción, la CMP es un espacio de trabajo que aloja las representaciones que especifican lo que sucede en un episodio de ficción. De los tres mecanismos que se postulan como subyacentes a la ficción o la imaginación sólo dos intervendrán en *mindreading*, a saber, la caja de los mundos posibles y el mecanismo actualizador, que forma parte del mecanismo inferencial postulado en la arquitectura cognitiva básica. El tercer mecanismo postulado es el elaborador de guiones. Nichols & Stich (2003) postulan este mecanismo para dar cuenta de ciertos rasgos de la ficción que no quedan explicados por otros mecanismos o factores. En *mindreading*, el rol de la CMP consiste en construir un modelo del blanco más preciso y flexible, que permita generar el mejor plan para satisfacer un deseo desde el punto de vista del blanco. Este modelo se elabora en la CMP con un procedimiento similar al descripto anteriormente para el razonamiento hipotético. Así, se generan representaciones de las creencias del

[41] Estas representaciones en la CMP son representaciones "opacas", en las que las relaciones de referencia, verdad y existencia están suspendidas.

blanco según diferentes estrategias, por ejemplo según la situación perceptual[42]. Éstas se alojan en la CMP y luego se vuelca la caja de creencias del *mindreader* a la CMP. A continuación opera el mecanismo inferencial y el actualizador que particularmente remueve aquellas representaciones incompatibles[43]. De este modo, se logra una descripción de la caja de creencias del blanco que se utilizará para determinar el mejor plan para satisfacer un deseo del blanco. En la medida en que se dispone de un modelo de la caja de creencias del blanco, se asume que la estrategia formulada es óptima y que la predicción resultante es más acertada que la que se genera con el sistema temprano.

2.1. Acerca del carácter híbrido de teoría y simulación

Como he mencionado, esta es una propuesta híbrida en tanto algunos aspectos de *mindreading* son explicados por simulación y otros por procesos ricos en información. Esta sección está dedicada a señalar aquellos aspectos de la propuesta que son de tipo simulacional o ricos en información. Llamativamente, el libro de Nichols & Stich (2003) comienza con una propuesta de los mecanismos subyacentes a la capacidad cognitiva de la ficción. Sin embargo, esto no resulta ser tan llamativo si se

[42] Además de la información acerca de la situación perceptual se postulan otros modos de atribuir deseos y creencias discrepantes al blanco. Si se conocen sus metas se puede saber qué cree el blanco a partir de lo que está haciendo. Por ejemplo, si la meta del blanco es agarrar las llaves y se lo ve revolviendo un cajón podemos saber que está buscando las llaves. Además, como señale más arriba, el pensamiento hipotético supone una atribución por *default* de todas las creencias del *mindreader*, al volcar la caja de creencias en la caja de mundos posibles. Sin embargo, esta atribución tiene límites. Si bien hay creencias que podemos volcar, por ejemplo el fuego quema, hay otras que no, por ejemplo si el blanco es un vecino nuevo no puedo atribuirle mis creencias religiosas. Estos límites en la atribución por *default* van a depender del individuo y la cultura. Nichols & Stich (2003) no proponen ningún mecanismo, sólo sugieren que puede tratarse de una adquisición tardía.

[43] El actualizador forma parte de los mecanismos inferenciales y es formulado por Stich & Nichols (2003) para dar cuenta de una función general en la cognición. A saber, cuando el sistema cognitivo lleva a cabo sus tareas normales, el agente cognitivo está enterándose constantemente acerca de los hechos por medio de la percepción, de las inferencias o del reporte de otra persona. Así, nuevas creencias se agregan constantemente al sistema cognitivo y éste está actualizando su sistema de creencias a medida que ingresa nueva información. Si bien no se sabe cómo el sistema cognitivo lleva a cabo la tarea de actualizar numerosas creencias durante su funcionamiento, se supone que lo hace todo el tiempo, de manera adecuada y no costosa. Así, el mecanismo actualizador se postula para dar cuenta de la ausencia de caos inferencial.

advierte que con esta propuesta se está formulando y fundamentando la arquitectura cognitiva que posibilita la simulación. Nótese que, más allá de la importancia de la capacidad cognitiva de la ficción, que parece estar implicada en múltiples capacidades cognitivas, se asume que se trata de la capacidad que, en general, subyace a la imaginación (Nichols & Stich 2003: 17). Es más, los autores sostienen que en tanto que no brindan una explicación para el comportamiento de ficción, su propuesta constituye, más bien, una propuesta sobre la imaginación (Nichols & Stich 2003: 37). Tradicionalmente, la simulación y la imaginación están estrechamente relacionadas (Goldman 1995a; Gordon 1995a; Harris 1995; Heal 1995a; Davies & Stone 1998).

En principio, el elemento simulacional se puede asociar al reclutamiento de la CMP para "asistir" a *mindreading*, esto es, el sistema tardío. Si bien Nichols & Stich (2003) han tratado de ser claros al respecto, la cuestión del significado de "simulación" es siempre engorrosa debido a los múltiples y diferentes procesos a los que se ha denominado "simulación"[44]. No obstante, los autores establecen un prototipo respecto del cual comparan y diferencian su propuesta, la simulación *off-line*. Recordemos que la simulación *off-line* se caracteriza por reclutar un sistema cognitivo que funciona de manera no estándar, desconectado de los mecanismos controladores de la acción, y es alimentado con *inputs* no estándar o ficticios. En principio, en la propuesta de Nichols & Stich (2003) hay dos procesos que se asemejan a este tipo de simulación. La predicción de inferencias se logra permitiendo que el mecanismo inferencial funcione sobre las representaciones en la CMP. El uso del planificador permite generar predicciones sobre cómo satisfacer deseos y para determinar los deseos instrumentales. En ambos casos, los sistemas funcionan de manera no estándar, desconectados de los mecanismos controladores de la acción. Esto, en la medida que generan creencias que se utilizarán para formular predicciones sobre inferencias y comportamiento. No obstante, según Nichols & Stich, los mecanismos mencionados se diferencian de la simulación *off-line* en tanto que las representaciones en la CMP no implican la existencia de un generador de *inputs* ficticios.

[44] Una lista puede encontrarse en Nichols & Stich (2003): 132-233.

Más allá de las similitudes y diferencias con la simulación *off-line*, hay otros dos procesos en esta propuesta que pueden asimilarse a la simulación, pero de otro tipo. La atribución por *default* de creencias y el uso del actualizador para ajustar el modelo del blanco en la CMP recuerdan a aquellas propuestas de simulación según las cuales es posible ponerse imaginariamente en el lugar del otro y usar las propias inferencias, los propios juicios, o incluso, asumir las mismas necesidades básicas para hacer luego los ajustes necesarios del caso (Goldman 1995a, Gordon 1995a, Harris 1995). En estos procesos nada corresponde a un *input* ficticio, ni hay ningún sistema funcionando *off-line*. Además, otra diferencia de la propuesta de Nichols & Stich (2003) respecto de otras formas de simulación reside en que los deseos no se generan de manera vicaria, sólo se utilizan creencias acerca de los deseos del blanco. Como he mencionado, los MDD identifican las metas y deseos, y generan creencias acerca de los mismos que son alojadas en la caja de creencias del *mindreader* y luego, volcadas en la CMP para ser utilizadas en *mindreading*.

Por su parte, el sistema temprano consta de mecanismos subyacentes de tipo ricos en información. Los MDD, así como los mecanismos detectores de percepción y algunos mecanismos para generar creencias discrepantes suponen conocimiento sobre las metas, la percepción y demás. Particularmente, lo que lleva a Nichols & Stich (2003) a sostener que se trata de procesos ricos en información es la presencia de errores sistemáticos ligados a los mismos. Por ejemplo, los mecanismos detectores de percepciones no son capaces de detectar ilusiones perceptuales y las predicciones al respecto son sistemáticamente erróneas. No obstante, a mi entender, en esta propuesta el sistema temprano ya resulta suficiente para predecir la conducta en términos mentalistas, aunque con ciertos límites, mientras que el sistema tardío viene a agregar flexibilidad, en la medida en que permite generar un modelo más acabado del blanco con el auxilio de la CMP. Sin embargo, los errores a los que conducen los mecanismos ricos en información del sistema temprano no se rectifican con este agregado. En este sentido, si las fallas asociadas a los mecanismos ricos en información no se rectifican con la asistencia de la simulación, en mi opinión, este enfoque híbrido

no sugiere un complemento entre teoría y simulación. Más bien, la asistencia de la simulación consiste en contribuir con aquello para lo que este proceso parece tener un caso a favor, la predicción de las inferencias y, en menor medida, la atribución por *default* de creencias. Como mencioné en el capítulo 3 (sección2.2), en la literatura de *mindreading* se acepta que un proceso de tipo simulacional subyace a la predicción de las inferencias ajenas (Harris 1995, Nichols & Stich 2003; Stich & Nichols 2003, Goldman 2006, Apperly 2008). Por ejemplo, para saber cuál dirá *Sally* que es el resultado de la suma "2 +2", basta con realizar uno mismo la suma y atribuir el resultado de la misma a *Sally*. En este sentido se asume que *recurrimos* a los propios procesos inferenciales para predecir las inferencias ajenas. Si recurrimos a los propios mecanismos inferenciales y no a una teoría sobre cómo las personas realizan inferencias, la TS tiene un caso a favor. Si bien no hay un complemento entre teoría y simulación, la propuesta de los sistemas temprano y tardío para *mindreading* de tercera persona puede considerarse como un enfoque híbrido de teoría y simulación con énfasis en la teoría, en la medida en que los mecanismos del sistema temprano son en su mayoría ricos en información y el sistema tardío aporta el elemento simulacional.

2.2. Los requisitos para mindreading de otras personas

En esta sección me ocuparé de evaluar si la propuesta de Nichols & Stich (2003) satisface los requisitos de heteroatribución (HA1)-(HA4) para *mindreading*, establecidos en el capítulo 3 (sección 3.1). A mi entender, el requisito de que (HA1) los adultos normales atribuyen percepciones, conocimientos, creencias, deseos, intenciones, decisiones y razonamientos a otras personas queda satisfecho con la propuesta del sistema de *mindreading* tardío, que puede dar cuenta de todo el rango de atribuciones con los mecanismos postulados. Básicamente, este enfoque explica cómo se generan creencias acerca de los deseos y las percepciones ajenas con la propuesta de los MDD y los mecanismos detectores de percepciones. También da cuenta de las creencias, conocimiento e intenciones ajenas con la propuesta de la CMP,

que junto con el actualizador permiten generar un modelo de (la caja de) las creencias del blanco. Los mecanismos inferenciales, incluidos el planificador y el actualizador, operan sobre la CMP del mismo modo como operan sobre las cajas de creencias y de deseos. En este sentido, al aplicarse los mecanismos inferenciales sobre el modelo de las creencias del blanco, alojado en la CMP del *mindreader*, se simulan los razonamientos e inferencias del blanco, que luego se le atribuyen.

El sistema temprano parece dar cuenta especialmente de los requisitos (HA2) y (HA3), aunque con ciertas reservas. El requisito (HA2) sobre el salto significativo en las habilidades de *mindreading* alrededor de los 3 años, queda satisfecho con la propuesta del paso del sistema temprano al sistema tardío cuando comienza a estar disponible la posibilidad de utilizar la CMP, y el actualizador, para llevar a cabo *mindreading*. Según esta propuesta, los estudios sobre el juego de ficción sugieren que la CMP está activa tempranamente a la edad de dos años. Los niños de dos años y medio ya pueden involucrarse en el juego de ficción y en razonamientos hipotéticos, dos capacidades que utilizan el recurso cognitivo de la CMP, pero aún no pueden pasar con éxito las tareas de falsa creencia. Nichols & Stich (2003) interpretan que en este estadio los niños aún no pueden utilizar la CMP para generar modelos de creencias del blanco. Recién a los tres años, la CMP comienza a asistir a *mindreading* cuando se produce el salto en el desarrollo en relación a las habilidades mentalistas, sugerido por el desempeño exitoso en tareas de falsa creencia. Así, el salto se genera por la posibilidad de construir un modelo más acabado de las creencias del blanco y esto permite generar predicciones óptimas sobre el comportamiento y las inferencias ajenas. Antes de esto, los niños sólo pueden dar cuenta del comportamiento ajeno atribuyendo metas y el mejor plan para alcanzarlas. Sin embargo, esta estrategia es limitada en tanto que siempre se construye desde las creencias del *mindreader*. En la medida en que los niños pequeños no pueden construirse un modelo del blanco con creencias discrepantes, fallan en la tarea de falsa creencia.

La cuestión de que la comprensión de los deseos precede a la comprensión de creencias en los niños pequeños (el requisito HA3) queda parcialmente satisfecha con

la postulación del sistema temprano. Según esta propuesta, las estrategias de detección de deseos y metas están disponibles de manera temprana en el desarrollo. Esto sugiere una comprensión temprana de los deseos. Sin embargo, a mi entender, esta propuesta no da cuenta de cierto rasgo llamativo en la comprensión temprana de los deseos. Como mencioné en el capítulo 3 (sección 3.1), la comprensión temprana de los deseos implica que los niños de 2 años son capaces de advertir que puede existir una diferencia entre lo que ellos desean y lo que otros desean (Meltzoff, Gopnik & Repacholi 1999). Esto es, los niños pueden advertir la variabilidad de los deseos (a Sally le gusta el brócoli y a mí no) antes que la variabilidad de las creencias (Sally tiene una creencia distinta a la mía). Considero que este rasgo particular de la comprensión temprana de los deseos no está explicado en la propuesta, aunque la propuesta puede dar cuenta de que la comprensión de los deseos precede a la de las creencias.

En la propuesta de Nichols & Stich (2003), como ya he señalado, los mecanismos subyacentes a la atribución de deseos son los MDD. Estos mecanismos dan cuenta de la detección de los deseos o metas a partir de señales variadas, tales como comportamiento no verbal y verbal, expresiones faciales y demás. A partir de estas señales, los MDD generan creencias acerca de los deseos ajenos que se colocan en la caja de creencias del *mindreader*. Por ejemplo, se genera la creencia “a Juan le gusta el brócoli”. Sin embargo, la comprensión temprana de los deseos implica también que los niños comprenden la variabilidad de los mismos, y esto es algo más que detectar un deseo. Un niño de dos años no sólo comprende que “a Juan le gusta el brócoli” sino también que “a Juan le gusta el brócoli pero a Ana no”, o bien que “a Juan le gusta el brócoli, pero a mí no”. Sin embargo, no todo esto está implicado en la mera detección de los deseos. Los MDD captan los deseos pero no la variabilidad o “discrepancia”, de modo que esta queda sin explicar. Ahora bien, si el modo de acceder a la variabilidad de los deseos implica de algún modo contemplar creencias

discrepantes, esta propuesta también está en problemas porque esta posibilidad requiere de la puesta en marcha del sistema tardío, que todavía no está disponible[45].

El requisito de que (HA4) las personas llevan a cabo juicios de *mindreading* proyectando sus propias creencias, sensaciones y conocimientos, en principio, puede considerarse satisfecho con la propuesta de la estrategia de atribución de creencias por *default*. Tal como sucede en el razonamiento hipotético y la ficción, en *mindreading* las creencias del *mindreader* son volcadas en la CMP para generar un modelo de la caja de creencias del blanco. Ahora bien, si se asume que el blanco comparte todas las creencias del *mindreader*, la CMP sería irrelevante para llevar a cabo predicciones sobre el blanco. Sólo cuando el sistema es sensible a la discrepancia de las creencias, la CMP se vuelve una herramienta relevante (Nichols & Stich 2003: 97). El modelo de la caja de creencias del blanco se configura en la CMP mediante la operación del actualizador que se encarga de remover las creencias que son incompatibles con las del blanco. Asimismo, se utilizan varios mecanismos para atribuir creencias discrepantes, como los mecanismos detectores de percepción, la explicación del comportamiento a partir del conocimiento de las metas del blanco, reportes por parte del blanco o por parte de lo que otros dicen del blanco, y demás. El otro mecanismo que sugiere la utilización de los propios recursos es la predicción de inferencias ajenas. Para llevar esto a cabo, se aplican a las representaciones en la CMP los mismos mecanismos

[45] Nichols & Stich conocen la evidencia de que los niños captan tempranamente la variabilidad de los deseos y la utilizan para discutir la propuesta de que el salto en el desarrollo es producto del paso de una teoría no representacional a una teoría representacional de los estados mentales. Según los autores, los psicólogos del desarrollo entienden que la apreciación de la variabilidad intersubjetiva de los deseos es un indicador del entendimiento representacional de los deseos. No obstante, si la apreciación de la variabilidad intersubjetiva cuenta como evidencia de que el niño tiene un entendimiento representacional de los deseos, la diferencia entre los niños pequeños y grandes se diluye puesto que, según la evidencia, el entendimiento de la variabilidad de los deseos es anterior a la adquisición de la teoría representacional. De modo que esta evidencia no puede ser usada por los psicólogos del desarrollo para sostener que el salto en el desarrollo consiste en la adquisición de una comprensión representacional (Nichols & Stich 2003: 112-113). Nichols & Stich sostienen que el salto en el desarrollo se debe a la posibilidad de usar la CMP, con la concomitante capacidad para generar un modelo del blanco que sea sensible a las creencias discrepantes. Esto implica que no se puede recurrir al uso de la CMP para dar cuenta de la comprensión de la variabilidad de los deseos porque la comprensión de la variabilidad es anterior al momento en que la CMP está disponible para ser usada en *mindreading*. En este sentido, no disponen de elementos para explicar la comprensión de la variabilidad de los deseos.

inferenciales que operan sobre las representaciones en la caja de creencias cuando se llevan a cabo inferencias.

2.3. El requisito del criterio para distinguir entre teoría y simulación

Nichols & Stich proponen una teoría híbrida de *mindreading* de tercera persona donde algunos aspectos se explican por simulación y otros por procesos ricos en información. Como propuse en el capítulo 3 (sección 3.3), considero necesario que los enfoques híbridos puedan especificar las circunstancias en que cada uno de los procesos está implicado. En este sentido, es preciso que puedan brindar un criterio para distinguir el tipo de proceso subyacente a un caso de *mindreading*. Si bien en la sección 2.1 intenté aclarar qué mecanismos se consideran de tipo simulacional y cuáles de tipo ricos en información, no quedaron especificadas claramente las circunstancias a las que subyace cada tipo de proceso. Nichols & Stich (2003: 106) y Stich & Nichols (2003: 14-15) proponen un criterio de corrección en base argumentos que voy a especificar en 2.3.1 y 2.3.3:

> En aquellos dominios donde somos particularmente buenos para predecir o atribuir estados mentales, no es probable que el proceso que subyace a *mindreading* sea de bases de información. Pero en aquellos casos donde somos malos para predecir o atribuir estados mentales no es probable que el proceso que subyace a *mindreading* sea simulacional… [P]ensamos que [esto] justifica una conjetura inicial fuerte de que a los procesos correctos de *mindreading* les subyacen procesos de tipo simulacional y que a los incorrectos no. (Stich & Nichols 2003: 14-15, traducción propia)

A mi entender, de esta cita se desprende el siguiente criterio de corrección: si el resultado de *mindreading* es correcto es probable que el proceso subyacente sea de

tipo simulacional, si el resultado de *mindreading* es incorrecto es probable que el proceso subyacente sea de tipo rico en información. En esta sección evaluaré la capacidad de este criterio de corrección para distinguir entre teoría y simulación. Para llevar a cabo esta tarea dividiré el criterio de corrección en dos aspectos.

En relación al primer aspecto del criterio, *i.e.* si somos buenos en *mindreading* es probable que el proceso subyacente sea de tipo simulacional, sostendré que no resulta satisfactorio, porque los argumentos no permiten descartar explicaciones de procesos ricos en información subyacentes a casos de *mindreading* exitoso, condición necesaria para constituirse como una explicación preferible. En contra del segundo aspecto del criterio, *i.e.* si sistemáticamente nos equivocamos en *mindreading* es probable que el proceso subyacente sea de tipo rico en información, sostendré que no es satisfactorio porque un proceso de tipo simulacional también puede subyacer a *mindreading* incorrecto. Concluiré, en base a un ejemplo, que el criterio de corrección propuesto no permite distinguir entre tipos de procesos subyacentes.

2.3.1. Acerca de mindreading exitoso

Stich & Nichols (2003) ofrecen dos argumentos para sostener la afirmación de que un proceso de tipo simulacional subyace a *mindreading* exitoso. En primer lugar, los autores brindan un argumento por la simplicidad que parte del hecho de que los seres humanos "somos buenos para predecir las inferencias ajenas, incluso, las inferencias no demostrativas" (Stich & Nichols 2003: 12). Según Stich & Nichols (2003), la única explicación disponible de este fenómeno por parte de los defensores de un enfoque rico en información requiere postular más información. Así, si somos buenos para predecir las inferencias ajenas, debe ser porque hemos adquirido una buena teoría sobre cómo razona la gente. Sin embargo, Stich & Nichols (2003) consideran que este enfoque es derrochador:

> Para entender el punto considérese la analogía entre predecir inferencias y predecir las intuiciones gramaticales de los hablantes de la propia lengua. Para explicar el éxito en esta última tarea, un defensor del enfoque rico en información tendrá que postular que poseemos una teoría sobre los procesos que subyacen a la producción de intuiciones gramaticales ajenas. Pero, como sugirió Harris (1992), esto es poco probable. Una hipótesis más simple es que confiamos en nuestros propios mecanismos para generar intuiciones lingüísticas, y habiendo determinado nuestras propias intuiciones sobre una oración particular, se las atribuimos al blanco. (Stich & Nichols 2003: 12).

Así, según este argumento, un teórico de la teoría podría sostener que los juicios de gramaticalidad se llevan a cabo mediante una (teoría) gramática subyacente. Si esto se asume, la gramática no resulta una teoría psicológica de sentido común. Su tema es la sintaxis, no las creencias acerca de la sintaxis. La gramática y las creencias acerca de la gramática son dominios diferentes y deben tener teorías diferentes (Goldman 2006: 182). Algo similar se asume en el caso de las inferencias. Poseer un conjunto de principios que guían las inferencias no es poseer unos principios sobre cómo las personas llevan a cabo inferencias en general. Sólo indican qué conclusiones se pueden inferir a partir de ciertas premisas, y no qué es lo que infieren las personas a partir de esas premisas (Goldman 2006: 181). De modo que según este argumento, un teórico de la teoría se vería obligado en estos casos a postular una reduplicación de teorías para dar cuenta de los juicios de gramaticalidad y las predicciones de inferencias ajenas. Frente a esto, resulta una explicación más simple postular que se utilizan los propios principios en la tarea de predecir las intuiciones gramaticales y las inferencias ajenas. Esta aplicación de los propios principios implica un proceso de tipo simulacional.

Ahora bien, si bien se puede conceder la plausibilidad de estos casos, resulta controversial la posibilidad de poder extender esto a todos los casos exitosos de

mindreading, tal como sugiere el criterio. Por ejemplo, en el caso de la predicción de una creencia cualquiera no parece ser el caso que un teórico de la teoría tenga que postular la reduplicación de teorías. Más bien, la posesión de creencias y la predicción de creencias parecen pertenecer ambas al mismo dominio de la teoría psicológica ordinaria. Es más, como he mencionado en el capítulo 3 (sección 3.2), algunas versiones de la "teoría de la teoría" asumen que utilizamos un mismo cuerpo de información o teoría para atribuir estados mentales, y explicar y predecir el comportamiento propio y ajeno (Sellars 1956; Gopnik & Meltzoff 1992).

El segundo argumento propuesto de parsimonia parte de la observación de que los seres humanos somos muy buenos para ciertas tareas de *mindreading*, por ejemplo para predecir las inferencias ajenas, pero muy malos para otras, por ejemplo para predecir los deseos ajenos (Stich & Nichols 2003). Según los autores, esto implica el siguiente problema para un enfoque rico en información:

> …¿cómo se las arreglan los *mindreaders* comunes para llegar a una teoría tan acertada sobre cómo las personas realizan inferencias –¿una teoría que permite predicciones correctas incluso en tipos de inferencias novedosas? El problema se vuelve más grave por el hecho de que hay otros tipos de tareas de *mindreading* que las personas realizan muy mal. *¿Por qué las personas adquieren la teoría correcta sobre las inferencias y una teoría incorrecta sobre otros procesos mentales?* Contrariamente, un enfoque de simulación sobre la predicción de inferencias tiene una explicación de la corrección. Según el enfoque de simulación, estamos usando el mismo mecanismo inferencial en ambos casos, al realizar y al simular inferencias, entonces es esperable predecir que los otros llevarán a cabo nuestras mismas inferencias. (Stich & Nichols 2003: 14, énfasis agregado)

Al parecer, Stich & Nichols (2003) creen que hay buenas razones para sostener que poseemos teorías incorrectas, en la medida en que hay ciertos errores sistemáticos en *mindreading*. Un enfoque rico en información no puede ofrecer una explicación satisfactoria de *mindreading* exitoso porque tiene que dar cuenta, al mismo tiempo, de la posesión de teorías correctas subyacentes a los aspectos exitosos y de la posesión de teorías incorrectas subyacentes a los aspectos incorrectos de *mindreading*. Según los autores, esto resulta poco parsimonioso en comparación con la explicación más simple que brinda la simulación.

Ahora bien, se podría conceder a los autores que brindar una explicación de la co-existencia de teorías psicológicas correctas e incorrectas implica cierta complejidad. Sin embargo, a mi entender, esto no es suficiente para aceptar el enfoque de simulación. En primer lugar, la mera posibilidad de que la simulación pueda dar cuenta de los aspectos exitosos de *mindreading* no descarta que estos mismos puedan ser explicados por la posesión de una buena teoría sobre cierta competencia psicológica. En la medida en que esta posibilidad no es descartada, el enfoque simulacional no es preferible. La TT y la TS pueden explicar los aspectos exitosos. Es más, cierta versión de la TT postula un mismo y único cuerpo de información subyacente a *mindreading* de primera y de tercera persona. En este sentido, la propuesta es simple porque no hay reduplicación de teorías De este modo, la simulación definida como el recurso a un único mecanismo resulta una explicación más simple respecto de la postulación de varias teorías (correctas e incorrectas), sin embargo, por sí misma la simplicidad no parece suficiente para preferir la explicación.

De este modo, en la medida en que ninguno de los dos argumentos logra descartar la posibilidad de que un proceso rico en información subyazca a *mindreading* exitoso, el primer aspecto del argumento de corrección, *i.e.* que un proceso de tipo simulacional subyace a *mindreading* exitoso, no encuentra apoyo en los argumentos ofrecidos y, en consecuencia, no ofrece un criterio para distinguir entre tipos de procesos subyacentes a *mindreading*, tal como proponen Stich & Nichols (2003).

2.3.2. Acerca de mindreading incorrecto

Como he mencionado, se asume que un enfoque simulacional no podría dar cuenta de la fallas en las predicciones, ni de las fallas sistemáticas de otros aspectos de *mindreading*, porque éste niega que se utilicen cuerpos de información en su ejecución, y no parece haber otro modo de explicar las fallas sistemáticas en *mindreading* que apelando a cuerpos de información erróneos o incompletos. No obstante, aquí sostendré que el segundo aspecto del criterio de corrección, *i.e.* si sistemáticamente nos equivocamos en *mindreading* es probable que el proceso subyacente sea de tipo de bases de información, no se sostiene si es posible que un proceso de tipo simulacional subyazca a ciertas fallas en las atribuciones mentalistas. Si esto es así, las fallas sistemáticas no sólo se explican por el recurso a cuerpos de información erróneos o incompletos. A continuación argumentaré que es posible que a *mindreading* incorrecto le subyazca un proceso de tipo simulacional.

Como hemos visto en la sección 2.1, en la propuesta de Nichols & Stich (2003) hay procesos que se asemejan a la simulación *off-line*, la predicción de inferencias y el uso del planificador para generar planes para alcanzar metas, en la medida que funcionan desconectados de los mecanismos controladores de la acción. Nótese que estos generan predicciones sobre inferencias o comportamiento ajenos. Estos mecanismos parecen diferenciarse de la simulación *off-line* con respecto a la utilización de *inputs* ficticios. Según Nichols & Stich (2003), su propuesta de la CMP se diferencia de la propuesta de *inputs* ficticios en tanto no supone la existencia de un mecanismo generador de tales *inputs*. En principio, basta con la semejanza respecto del funcionamiento *off-line* para ofrecer mi argumento.

Es preciso señalar que, según la TS, cualquier sistema cognitivo puede ser reclutado para la simulación siempre y cuando éste sea alimentado por actitudes proposicionales como *input* y produzca cualquier tipo de estados mentales como *output* (Stich & Nichols 1992). No obstante, el ejemplo paradigmático de *mindreading* vía simulación es la predicción del comportamiento ajeno, que recluta al sistema de toma de decisiones (o razonamiento práctico). Éste, alimentado con *inputs* ficticios y

operando de manera *off-line*, producirá un *output* que será utilizado para predecir el comportamiento del blanco (Gordon 1986; Goldman 1992; Stich & Nichols 1992; Nichols *et al.* 1996).

Tan paradigmático resulta el caso de la toma de decisiones que los teóricos de *mindreading* se hacen eco de una característica llamativa de los procesos decisorios. A saber, a menudo las personas toman decisiones inesperadas o irracionales. En relación a esto, los teóricos de *mindreading* sostienen que no es algo preocupante. Si el sistema cognitivo en cuestión tiene alguna peculiaridad que lo lleva a comportarse, en ciertas circunstancias, de maneras inesperadas para las personas, esto no afectará la corrección de las predicciones puesto que utilizamos el *mismo* sistema para realizar inferencias y para predecir las ajenas (Stich & Nichols 1992). Estas peculiaridades son los sesgos irracionales que, según la literatura de psicología social, operan usualmente sobre la toma de decisiones, aunque también sobre las inferencias y las atribuciones de estados mentales.

Sin embargo, a mi entender, no es claro que del reclutamiento del sistema de toma de decisiones para simular una decisión ajena se siga que, si el proceso decisorio incluye la operación de algún sesgo, éste también operará en la decisión que se genera vía simulación. El problema reside en que, por un lado, no disponemos de modelos sobre en qué momento del proceso de toma de decisiones intervienen los sesgos y, en este sentido, son plausibles varias posibilidades. Esto último es importante porque si se asume el funcionamiento no estándar del sistema de toma de decisiones en la simulación, hay razones para suponer que los sesgos pueden no reclutarse.

En principio, el diagrama de flujo utilizado por los teóricos de *mindreading* para caracterizar la toma de decisiones ordinaria, sobre la cual se modela *mindreading* por simulación, es muy esquemático (Stich & Nichols 1992: 40, 1995b: 105; Gallese & Goldman 1998: 497; Goldman 2006: 27)[46]. En virtud de la generalidad con la que los

[46] Stich & Nichols son claros en este respecto cuando sostienen que: "Los diagramas son considerados esquemas rudimentarios de algunos de los mecanismos y procesos subyacentes a varias capacidades cognitivas, y debe tenerse en mente que no pretenden captar todos los mecanismos y procesos que pueden afectar el desempeño de las personas" (Stich & Nichols 1997: 305).

teóricos de *mindreading* describen el proceso de toma de decisiones y en la medida en que no se dispone de propuestas sobre en qué momento del proceso de toma de decisiones operarían los sesgos, en principio, todas las posibilidades quedan abiertas. En este sentido, lógicamente, puede ser el caso que (i) los sesgos operen a nivel del *input* del sistema de toma de decisiones, (ii) que operen a nivel del *output* del sistema de toma de decisiones, o bien, (iii) que operen al interior del sistema de toma de decisiones. Los modos (i) y (ii) pueden considerarse externos al sistema de toma de decisiones.

Ahora bien, como señalé anteriormente, la TS postula un funcionamiento no estándar del sistema cognitivo cuando es reclutado en la simulación, en la medida en que:

(A) se recluta el sistema de la competencia cognitiva que funcionará de modo *off-line*,

(B) este sistema es alimentado por *inputs* ficticios.

Sin embargo, dado que los procesos postulados por Nichols & Stich (2003) se asemejan claramente a la simulación *off-line* en el aspecto (A), pero no es claro que se asemejen en el aspecto (B), voy a desarrollar sólo los casos (ii) y (iii) que son suficientes para presentar mi argumento.

Si es el caso que (iii) los sesgos operan al interior del sistema cognitivo, puesto que en la simulación mental (A) se recluta el sistema de la competencia cognitiva, entonces los sesgos también serán reclutados e intervendrán en la simulación. Esto conducirá a la concordancia entre el *output* de *mindreading* y el *output* del blanco. De este modo, puede asumirse que la simulación conducirá a *mindreading* correcto, en concordancia con la afirmación de Stich & Nichols (1992).

Sin embargo, el problema surge si es el caso que (ii) los sesgos operan a nivel del *output*. Dado que (A) el sistema de la competencia cognitiva reclutado para la simulación funciona de manera *off-line*, es decir, desconectado de los sistemas controladores de la acción, el funcionamiento del sistema es no estándar. Ahora bien, en la simulación los *outputs* no son estándar, aunque no es claro qué significa esto más

allá de que se usan para generar predicciones y no para generar comportamiento. El punto es que si el sesgo funciona a nivel del *output*, quiere decir que funciona después que el sistema de toma de decisiones hizo lo propio y quizás hasta opere justo antes de que este *output* alimente como *input* a los sistemas controladores de la acción. Si este es el caso, puesto que en la simulación *off-line* el *output* de sistema de toma de decisiones se desacopla de los controladores de la acción, no hay razones para afirmar que los sesgos también operan sobre los *outputs* no estándar. Así, la corrección de las predicciones no queda garantizada en virtud del recurso a un sistema relevantemente similar al del blanco para realizarlas. Dado el funcionamiento no estándar, puede ser el caso que el sesgo no sea reclutado para *mindreading*, de modo que el *output* de *mindreading* y el *output* del blanco no concuerden por esta razón, y *mindreading* no sea correcto.

Así, la posibilidad de ejecutar *mindreading* mediante un proceso de tipo simulacional que sólo recluta el mecanismo cognitivo sin la intervención de los sesgos implica que el *output* de *mindreading* no incluirá el aporte de los mismos. De esta manera, el *output* de la simulación no coincidirá con el *output* del sistema cognitivo del blanco cuando éste esté sometido a sesgos. En consecuencia, no habrá concordancia entre el *output* de *mindreading* y el *output* de la competencia cognitiva. En este sentido, un proceso de tipo simulacional puede conducir a *mindreading* defectuoso, así como también lo hacen las teorías internalizadas sobre el razonamiento teórico y práctico que pueden incluir información errónea sobre los sesgos y conducir a predicciones erróneas. En este sentido, considero que es una cuestión crucial determinar si en la simulación mental se recluta únicamente al sistema cognitivo o al mismo junto con los sesgos. Mientras tanto no queda descartada la posibilidad de que la simulación pueda conducir a *mindreading* incorrecto.

De modo que ninguno de los aspectos del criterio de corrección es satisfactorio. Por un lado, la afirmación de que la simulación subyace a *mindreading* exitoso no se sostiene en tanto no se han descartado explicaciones basadas en un proceso de tipo rico en información. Por otro lado, la afirmación de que un proceso rico en información

subyace a *mindreading* incorrecto no se sostiene en la medida en que no queda descartada la posibilidad de que la simulación pueda conducir a errores sistemáticos en *mindreading*. En este sentido, concluyo que el criterio de corrección propuesto por Stich & Nichols (2003) no es satisfactorio para distinguir entre tipos de procesos subyacentes a *mindreading*.

3. Los mecanismos subyacentes a *mindreading* de primera persona

En esta sección presentaré la propuesta de Nichols & Stich (2003) para *mindreading* de primera persona. Ésta propone un método dual de introspección e interpretación, de modo que al autoatribuirse estados mentales en algunas ocasiones las personas tienen acceso directo a los mismos y en otras ocasiones utilizan los mecanismos interpretativos que emplean para atribuir estados mentales a las otras personas. Este tipo de propuesta permite conservar un enfoque tradicional sobre el autoconocimiento y, a la vez, dar cuenta de la evidencia que sugiere que las personas se autoatribuyen frecuentemente estados mentales mediante interpretación, sin percatarse de esto. En la sección 3.1, analizaré la propuesta en relación a los requisitos establecidos en el capítulo 3 (secciones 3.2 y 3.3) que, en mi opinión, se satisfacen parcialmente. Según el criterio propuesto para distinguir entre introspección e interpretación, la interpretación sólo se aplica a casos en los que se les pregunta a las personas por las causas de su comportamiento, mientras que la introspección subyace al resto de las autoatribuciones. Ahora bien, no toda la evidencia en torno a la confabulación puede explicarse como casos de este tipo, contra lo que sostienen Nichols & Stich. Como consecuencia, el criterio propuesto para distinguir entre los procesos subyacentes no parece apropiado.

Con el propósito de dar cuenta de que un adulto que cree que *p* se forma la creencia de que "yo creo que *p*", Nichols & Stich (2003) proponen un mecanismo de monitoreo subyacente a la autoatribución de actitudes proposicionales. Según los autores, este mecanismo no necesita utilizar un cuerpo de información sobre el funcionamiento de la mente. Más bien, cuando el mecanismo de monitoreo (en

adelante MM) es activado, toma una representación de la caja de creencias como *input*, por ejemplo *p*, y se forma la representación "yo creo que *p*" como *output*. Esto se implementa de la siguiente manera. El MM tiene que copiar la instancia de representación alojada en la caja de creencias e incrustarla en un esquema de representación con la forma "yo creo que..." y luego, volver a colocar la representación nueva en la caja de creencias. Según Nichols & Stich (2003: 161), el mismo mecanismo, o quizás uno paralelo y distinto, operará de la misma manera para producir representaciones de deseos, intenciones e imaginaciones.

A la base de la propuesta para la autoatribución se asume una distinción que, según los autores, ya está implícita en la propuesta de *mindreading* para la tercera persona. Es preciso distinguir entre "detectar" o la capacidad de atribuir a alguien, uno mismo u otra persona, estados mentales ocurrentes, y "razonar" o la capacidad de usar la información sobre los estados mentales de alguien, típicamente junto con otro tipo de información, para realizar predicciones sobre los estados mentales pasados, futuros, su comportamiento y su ambiente (Nichols & Stich 2003: 151). Según los autores, el MM da cuenta de la capacidad de detectar estados mentales propios, mientras que el razonamiento sobre los mismos está mediado por inferencias. En este último caso, es el mismo mecanismo que se utiliza para detectar y razonar sobre los estados mentales de otras personas. De este modo, se postulan mecanismos distintos para detectar estados mentales propios y para razonar sobre los mismos. No obstante, ambos mecanismos pueden dar lugar a representaciones de tipo "yo creo". Es más, estos mecanismos pueden generar representaciones contradictorias, pero no se propone un modo de resolver este conflicto. Esta distinción entre detectar y razonar parece ser un criterio para distinguir instancias de introspección y de interpretación, respectivamente.

Hasta aquí la propuesta abarca las actitudes proposicionales. No obstante Nichols & Stich (2003) sostienen que es probable que algún tipo de proceso de monitoreo subyazca también a la percatación de los estados perceptuales, el otro tipo

de estados mentales susceptibles de introspección[47]. Según los autores, la arquitectura que subyace al autoconocimiento de los estados perceptuales involucra una "caja perceptual" que aloja los perceptos captados por los sistemas de procesamiento perceptual. La caja perceptual alimenta a la caja de creencias, al menos, de dos maneras. Primero, permite formar, a partir de perceptos como *input,* creencias perceptuales como *output*. Por ejemplo, si un adulto normal mira en una cantera, su sistema perceptual genera perceptos que conducen, *ceteris paribus*, a la creencia de que "allí hay piedras". Segundo, se postula la existencia de un conjunto de mecanismos de monitoreo de perceptos que tienen como *input* los perceptos y producen como *output* creencias acerca de los mismos. De modo que se forman creencias acerca de la experiencia perceptual misma. Por ejemplo, si miro en una cantera puedo formarme la creencia de que "yo veo piedras".

De modo que Nichols & Stich (2003) sostienen que hay al menos dos (o más) mecanismos de monitoreo. Uno para detectar y proveer autoconocimiento de estados perceptuales (o de la experiencia) y otro para monitorear y proveer autoconocimiento de nuestras actitudes proposicionales. Estos mecanismos son distintos entre sí, y distintos de los mecanismos para "leer" la mente de las otras personas[48]. De manera que un enfoque de este tipo predice disociaciones dobles entre las capacidades de

[47] Nichols & Stich (2003) asumen que la propuesta de un (o unos) MM para las actitudes proposicionales no puede extenderse simplemente a los estados experienciales, debido al carácter esencialmente representacional de los MM. Como ya se ha mencionado, el MM copia representaciones y las incrusta en un esquema representacional con la forma "yo creo que…" para formar representaciones que colocará en la caja de creencias del sistema cognitivo. Sin embargo, se cuestiona que un enfoque representacional pueda dar cuenta del carácter fenoménico inherente a los estados perceptuales (Tye 1995; Carruthers 2000). En este sentido, los autores son escépticos respecto de que el mecanismo de monitoreo de los mismos sea representacional.

[48] La propuesta de Nichols & Stich (2003) se contrapone, principalmente, al enfoque ampliamente aceptado de que la autoatribución es un proceso paralelo a *mindreading* de tercera persona. Como mencioné en el capítulo 3, este enfoque es sostenido por los psicólogos del desarrollo que defienden la versión de la TT denominada "del niño científico". Apoyados en evidencia que permite sostener un paralelismo en el entendimiento infantil de los estados mentales ajenos y propios, los psicólogos del desarrollo sostienen que el conocimiento de uno mismo y de los otros es resultado de una teoría (Gopnik 1993; Gopnik & Meltzof 1994). La principal crítica a este enfoque, ya mencionada en el capítulo 3, reside en que no parece poder dar cuenta de la intuición de sentido común de que los adultos tienen conocimiento directo de sus estados mentales y, menos aún, de otros hechos obvios tales como que es posible estar sentado, sin exhibir comportamiento manifiesto, y aún así reportar estados mentales ocurrentes (Nichols & Stich 2003: 157).

atribuir estados mentales a otras personas y la autoatribución. Puesto que están realizados en diferentes mecanismos, debería haber casos en los que cada uno está dañado, mientras que la capacidad para *mindreading* de tercera persona está conservada. Así, en principio, se espera que pueda haber casos en los que una persona pueda atribuir estados mentales a otras personas pero no pueda realizar autoatribuciones apropiadamente. Según los autores, este es el caso de los pacientes con esquizofrenia que presentan síntomas de tipo "pasivo". Los síntomas pasivos aluden a la ausencia de control de sus propias acciones. Estos pacientes consideran que sus acciones están siendo controladas por otras personas o externamente, pero no por ellos mismos. Por su parte, el autismo constituye el déficit inverso, en la medida en que la capacidad de autoatribución está conservada y la capacidad de *mindreading* de tercera persona está alterada. De este modo, Nichols & Stich (2003) consideran haber encontrado una doble disociación que brinda apoyo a su enfoque, en detrimento del enfoque del niño científico que predice un déficit paralelo para *mindreading* de primera y tercera persona[49].

A su vez, esta propuesta genera otras predicciones en relación a los déficits. Como he mencionado, según esta propuesta, un mecanismo de monitoreo copia representaciones de la "caja de creencias", les agrega el prefijo actitudinal "yo creo que" y las convierte en representaciones que vuelve a colocar en la caja de creencias. Este enfoque parece plausible para las creencias pero no es claro que pueda generalizarse a todas las actitudes proposicionales. Para que esto sea posible, Nichols & Stich deben postular o bien distintos mecanismos de monitoreo para cada tipo de actitud proposicional, esto es, un mecanismo de monitoreo diferente para copiar una representación de la "caja de deseos" y agregarle el prefijo "yo deseo", para formar la representación que luego colocará en la caja de deseos, y así para el resto de las actitudes proposicionales. O bien, deben postular un mecanismo de monitoreo que

[49] La predicción que se sigue para el desarrollo es que los MM para detectar estados mentales propios deben emerger temprano en el desarrollo, porque son innatos. No obstante, estos mecanismos se encuentran empobrecidos inferencialmente hasta el momento en que el sistema inferencial esté disponible (Nichols & Stich 2003: 163).

recupere instancias de actitud de las múltiples “cajas” de actitudes. En ambos casos, el enfoque predice que cada mecanismo de monitoreo o cada uno de los canales por los que el único mecanismo de monitoreo es conectado con cada una de las “cajas” de actitudes puede estar dañado con independencia de los demás. En base a esto, es esperable encontrar múltiples patrones de disociación. Personas que puedan autoatribuirse deseos pero no creencias, o que puedan autoatribuirse experiencias visuales pero no auditivas, y demás. Sin embargo, no existe evidencia sobre tales disociaciones múltiples (Engelbert & Carruthers 2010).

3.1. Los requisitos para la autoatribución y el criterio de distinción

En relación al requisito (AA1) para la autoatribución, como ya he adelantado, postular la introspección como modo de acceso a los estados mentales propios, en principio, permite dar cuenta de la intuición de sentido común de acceso directo e inmediato a los mismos. Así como también sobre otros hechos obvios como la posibilidad de autoatribuirse estados mentales sin exhibir comportamiento manifiesto, mientras que un enfoque interpretativo tiene dificultades para explicar esto. No obstante, me gustaría mencionar una objeción a la intuición de sentido común relacionada con estudios presentados en el capítulo 3 (sección 3.2). Los datos aportados por los estudios de pacientes con cerebro dividido fuerzan a reconocer que el acceso de las personas a sus juicios e intenciones puede ser interpretativo, tal como es interpretativo el acceso a los juicios e intenciones ajenas (Gazzaniga 1995, 2000). Si bien estos estudios no permiten sostener que los pacientes nunca tienen acceso introspectivo a sus estados mentales, no obstante, permiten sacar otras conclusiones.

A saber, en una ocasión se le presenta a un paciente con cerebro dividido un estímulo en el hemisferio derecho, incomunicado con el izquierdo, que consiste en una tarjeta con la instrucción “¡Camine!”. Inmediatamente, el paciente se para y comienza a caminar por el laboratorio (recuérdese que si bien el hemisferio derecho tiene una capacidad limitada para la comprensión del lenguaje, no interviene en su producción).

Cuando se le pregunta por qué, el paciente contesta “voy a buscar una coca a la casa” (el hemisferio izquierdo aloja el control de la producción del habla). Esta atribución de una intención ocurrente es claramente falsa, sin embargo, se lleva a cabo con una confianza y con un aspecto de inmediatez tal como si fuera una atribución sin vicio. En estos estudios se observa que los pacientes llevan a cabo sus reportes confabulatorios con fluidez y sin dudar. Ahora bien, en la medida en que los pacientes reportan estados confabulados con la misma sensación de obviedad e inmediatez con la que las personas normales (sin cerebro dividido) realizan reportes introspectivos, no hay garantía para afirmar que el acceso a los juicios y las decisiones propias sea introspectivo. Al parecer, los sujetos no disponen de señales que les permitan distinguir entre introspección e interpretación. El punto es que la inhabilidad para discriminar entre interpretación e introspección muestra que no hay una razón subjetiva accesible para creer que tenemos acceso introspectivo a los estados mentales propios (Carruthers 2009).

Además, como señalé en el capítulo 3, actualmente está sólidamente establecido que las personas a menudo interpretan su propio comportamiento sin darse cuenta de que lo están haciendo y como resultado, frecuentemente, llevan a cabo falsas autoatribuciones de estados mentales. Desafortunadamente, en comparación se ha dedicado menos esfuerzo a investigar si, además de auto-interpretación, las personas tienen la capacidad de autoatribuirse estados mentales de manera no interpretativa, por vía de la introspección. Sería preciso que se dedicaran esfuerzos para tratar de establecer si la introspección existe o no (Engelbert & Carruthers 2010). De modo que, a mi entender, el requisito queda satisfecho en la medida en que Nichols & Stich (2003) proponen un mecanismo que puede dar cuenta de la intuición, aunque parcialmente. Esta propuesta depende de la existencia del mecanismo introspectivo, y se precisa algo más que apelar a las intuiciones y a la tradición filosófica. En este sentido, parece necesario el desarrollo de estudios empíricos sobre (la existencia) esta capacidad.

Recapitulando, Nichols & Stich (2003) sostienen que el proceso de monitoreo subyace a la autoatribución de estados mentales ocurrentes, mientras que el proceso de interpretación subyace a las predicciones y explicaciones en términos mentalistas. En este sentido, puesto que la explicación supone causación y que las relaciones causales entre los pensamientos, o entre los pensamientos y las acciones, no se pueden captar mediante la introspección, se captarán mediante la interpretación. De modo que, según Nichols & Stich (2003), la confabulación tendrá lugar cada vez que se le pregunte a una persona por qué hace algo o por qué piensa algo. Contrariamente, cuando una persona tiene que reportar un estado mental ocurrente o un pensamiento reciente, podrá acceder a este directamente y no habrá confabulación. Así, la propuesta no da cuenta del requisito (AA2) de que a menudo las personas interpretan (falsamente) sus pensamientos ocurrentes y del pasado reciente. A continuación mencionaré los casos de evidencia de confabulación, presentados en el capítulo 3, que no pueden ser abarcados por la propuesta del proceso interpretativo de Nichols & Stich (2003). Según la evidencia, no todos los casos de confabulación surgen en el marco de dar cuenta de las causas del comportamiento.

En los estudios de auto-percepción y disonancia, por ejemplo Linder e*t al.* (1967), los participantes son inducidos a escribir un ensayo defendiendo una posición que no comparten (*p.e.* la segregación racial). Según este estudio, al final el experimento las personas tienden a mostrarse más compasivas con la posición a la que no adhieren. Según la interpretación de los investigadores, si una persona es inducida a comportase públicamente de una manera que no coincide con sus actitudes privadas, experimentará disonancia cognitiva. Ante la imposibilidad de apelar a razones de peso (un incentivo o una remuneración) para su comportamiento, cambian sus actitudes respecto de la posición que no compartían para reducir la magnitud de la disonancia cognitiva. Esto sugiere que los participantes interpretan sus pensamientos. Ahora bien, la cuestión es que en este estudio sólo se les pide a las personas que reporten con cuanta convicción creen en algo (la posición que defendieron en el ensayo) pero no se les requiere que brinden explicaciones sobre las causas de su comportamiento. Así, la

evidencia sugiere que las personas interpretan sus pensamientos aún cuando no tienen que explicar su comportamiento (Carruthers 2013). En el estudio de Wegner & Weatley (1999), se les requiere a los participantes que logren un resultado junto con otra persona y que realicen un juicio sobre su contribución (mover el cursor en una pantalla de computadora). El hallazgo reside en que a pesar de que los participantes tienen el control total de la acción que conduce al resultado, sus juicios permanecen anclados en cercanía al punto medio como cuando el control es compartido. Se asume que de haber recurrido a la introspección para realizar un juicio sobre su contribución, deberían haber tenido un sentido contundente de control debido a la contigüidad temporal entre su acción y el resultado. Sin embargo, los participantes siguen juzgando su control en cercanía al punto medio y esto sugiere interpretación en los juicios. Nuevamente, en este caso tampoco se les pregunta a los participantes por las causas de su comportamiento y, sin embargo, confabulan.

En mi opinión, el estudio de Brasil-Neto *et al.* (1992) puede interpretarse de manera similar. En este estudio se requiere a los participantes que elijan qué dedo mover. El hallazgo reside en que las personas se autoatribuyen la decisión de mover el dedo cuando de hecho este movimiento ha sido provocado por la estimulación magnética transcraneana de áreas motoras del hemisferio. Según los investigadores, los participantes no advierten la influencia de la estimulación magnética en la elección de sus repuestas y esto muestra que es posible influenciar los procesos de preparación del movimiento sin alterar la percepción de la volición (Brasil-Neto *et al.* 1992: 966). En términos de interpretación de los estados mentales, la evidencia parece sugerir que en virtud de la información sobre el movimiento del dedo (información perceptual y propioceptiva) las personas interpretan haber decidido el movimiento, cuando de hecho éste es causado por la estimulación de áreas corticales motoras. En este estudio no se les pregunta a los sujetos por qué movieron el dedo. Sólo tienen que elegir qué dedo mover. Sin embargo, la interpretación parece mediar la atribución de estados mentales ocurrentes, contra Nichols & Stich (2003).

Como mencioné en el capítulo 3 (sección 3.3), considero que el requisito de proponer un criterio para determinar el tipo de proceso que subyace a *mindreading* también puede aplicarse a la propuesta para la autoatribución en la medida en que se proponen dos procesos subyacentes, monitoreo e interpretación. En principio la propuesta de Nichols & Stich establece un criterio en base a la distinción ya mencionada entre detección y razonamiento sobre estados mentales. El proceso de monitoreo interviene en la detección y la interpretación media el razonamiento sobre estados mentales propios. No obstante, esto no es del todo satisfactorio en virtud de los casos mencionados anteriormente de "detección" de estados mentales, juicios y decisiones, mediados por inferencias. De modo que el criterio para determinar en qué circunstancias subyace cada proceso parece no funcionar.

4. Conclusión

En este capítulo presenté y analicé la propuesta de Nichols & Stich (2003) de *mindreading* para la tercera y la primera persona. En el caso de la tercera persona, se propone un enfoque híbrido en el que algunos aspectos de *mindreading* se llevan a cabo por teoría y otros por simulación. A mi entender, en esta propuesta el sistema temprano ya es suficiente para predecir la conducta en términos mentalistas, aunque con ciertos límites, y el sistema tardío viene a agregar flexibilidad, en la medida en que permite generar un modelo más acabado del blanco con el auxilio de la CMP. Sin embargo, los errores a los que conducen los mecanismos ricos en información del sistema temprano no se rectifican con este agregado. De modo que este enfoque híbrido no sugiere un complemento entre teoría y simulación. Más bien, la asistencia de la simulación se ocupa de aquello para lo que ésta funciona bien o tiene un caso a favor que es la predicción de inferencias, pero no para subsanar los defectos de la teoría. A su vez, en la medida en que los mecanismos del sistema temprano son en su mayoría ricos en información, a mi entender, éste puede considerarse un sistema híbrido con énfasis en la teoría.

En relación a los requisitos propuestos en el capítulo 3 para *mindreading* de otras personas, considero que algunos se satisfacen y otros no. Los requisitos (HA1) y (HA2) se satisfacen trivialmente, mientras que el requisito (HA3) de que la comprensión de los deseos precede a la comprensión de las creencias en los niños pequeños queda satisfecho parcialmente. Según esta propuesta, los MDD están disponibles tempranamente de modo que posibilitan la comprensión temprana de los deseos ajenos antes de que el sistema tardío esté disponible. Sin embargo, a mi entender, con la postulación de los MDD sólo se da cuenta de la detección de los deseos pero no de la percatación de la variabilidad de los mismos. De modo que este aspecto de la comprensión de los deseos, previo a la comprensión de las creencias, no queda explicado. El requisito (HA4) de que las personas llevan a cabo juicios de *mindreading* proyectando sus propias creencias, sensaciones y conocimientos, en principio, puede considerarse satisfecho con la propuesta de la estrategia de atribución de creencias por *default* y de la predicción de las inferencias. Ambas recurren a la caja de creencias y al sistema inferencial del *mindreader*, respectivamente.

Más allá de la complejidad de la propuesta en relación a los distintos mecanismos, Nichols & Stich (2003) y Stich & Nichols (2003) proponen un criterio de corrección para determinar cuándo subyace un proceso de teoría o de simulación. Si el resultado de *mindreading* resulta correcto es probable que el proceso subyacente sea de tipo simulacional, si el resultado de *mindreading* resulta incorrecto es probable que el proceso subyacente sea de tipo rico en información. En contra del primer aspecto, sostuve que ninguno de los dos argumentos esgrimidos para sostener que *mindreading* exitoso es producto de la simulación logra descartar la posibilidad de que un proceso rico en información subyazca a *mindreading* exitoso. Si no se pueden descartar explicaciones de proceso de ricos en información, la explicación simulacional no es preferible. En contra del segundo aspecto del criterio, sostuve que es concebible que un proceso de tipo simulacional subyazca a *mindreading* erróneo. A saber, si al ejecutar *mindreading* mediante la simulación, se recluta el sistema de toma de decisiones pero no los sesgos que operan sobre el mismo, el *output* de *mindreading* no incluirá el

aporte de los sesgos. En consecuencia, no habrá concordancia entre el *output* de *mindreading* y el *output* de la competencia cognitiva del blanco, que está sometido a los sesgos. Esto generará una predicción errónea sobre el blanco. En este sentido, considero que un proceso simulacional puede generar *mindreading* defectuoso, tal como esto se genera mediante inferencias conducidas por cuerpos de conocimiento psicológico erróneos o incompletos. De este modo, en la medida en que ninguno de los dos aspectos del criterio resulta satisfactorio, concluyo que el criterio de corrección no ofrece un modo de distinguir entre los tipos de procesos subyacentes a *mindreading*.

La propuesta para *mindreading* de primera persona consiste en un método dual de introspección e interpretación. Como he mencionado, este tipo de propuestas permite conservar un enfoque tradicional sobre el autoconocimiento y, a la vez, dar cuenta de la evidencia que sugiere que las personas se autoatribuyen frecuentemente estados mentales mediante interpretación, sin percatarse de esto. En particular, según esta propuesta, el mecanismo de monitoreo se aplica a los casos de detección de estados mentales propios, mientras que la interpretación media el razonamiento sobre los mismos. Es más, los autores sostienen que hay al menos dos (o más) mecanismos de monitoreo. Uno para detectar y proveer autoconocimiento de estados perceptuales (o de la experiencia) y otro para monitorear y proveer autoconocimiento de nuestras actitudes proposicionales. Estos mecanismos son distintos entre sí, y distintos de los mecanismos para "leer" la mente de las otras personas. Los mecanismos para leer la mente de las otras personas son los que se utilizan para razonar sobre los estados mentales propios.

En relación a los requisitos, esta propuesta da cuenta de (AA1) la intuición que tienen los adultos de acceso directo a los estados mentales propios, en la medida en que postula la introspección para la autoatribución. Sin embargo, en la medida en que no parecen haber razones subjetivas accesibles para creer en la introspección, ya que las personas no parecen disponer de señales para distinguir entre introspección e interpretación, y en tanto no hay evidencia contundente que apoye la existencia de un mecanismo introspectivo, la propuesta se ve debilitada. La propuesta de Nichols &

Stich (2003) no da cuenta del requisito (AA2) de que a menudo las personas interpretan (falsamente) sus pensamientos ocurrentes. Si bien se postula un proceso interpretativo subyacente a ciertos casos de confabulación, se trata de unos casos particulares que no contemplan estados ocurrentes sino la autoatribución de las causas del comportamiento. En este sentido, la propuesta para la interpretación como subyacente a la confabulación sólo da cuenta de algunos casos de confabulación mencionados. En contra de Nichols & Stich (2003), la evidencia sugiere que hay casos en los que las personas se autoatribuyen estados mentales ocurrentes mediante interpretación. Si hay casos de "detección" de estados mentales mediados por inferencias en consecuencia el criterio para distinguir entre los procesos subyacentes no funciona.

CAPÍTULO 5. UN ENFOQUE HÍBRIDO DE *MINDREADING* CON ÉNFASIS EN LA SIMULACIÓN

En este capítulo presentaré la propuesta de Goldman (2006) para *mindreading* de primera y tercera persona. En la sección 1, me ocuparé del enfoque híbrido de teoría y simulación con énfasis en el elemento simulacional para *mindreading* de tercera persona. Esta propuesta tiene la particularidad de presentar un modelo simulacional para *mindreading* de nivel inferior (SNI), cuyo prototipo son los procesos espejo (sección 1.1), y un modelo simulacional para *mindreading* de nivel superior (SNS), cuyo prototipo es la imaginación enactiva (sección 1.2). En la sección 2, analizaré el carácter híbrido de teoría y simulación de esta propuesta. Argumentaré que esta propuesta es insuficiente porque no describe el rol de la teoría y apenas señala las posibles relaciones entre teoría y simulación. En este sentido, considero que no se brinda un criterio para determinar el tipo de proceso subyacente a un caso de *mindreading* (el requisito de un criterio para determinar el tipo de proceso subyacente). Además, señalaré que la redefinición de la noción de simulación en que se basa la propuesta es confusa y no queda claro qué tienen en común los prototipos de SNI y SNS como procesos subyacentes a *mindreading*. En relación con esto, mencionaré una crítica a la distinción de niveles (de Vignemot 2009). En la sección 3, analizaré la propuesta de *mindreading* de tercera persona según los requisitos de atribución heterogénea (HA1)-(HA4). El requisito (HA1) se satisface, mientras que los requisitos (HA2) y (HA4) se satisfacen aunque con salvedades. En el caso del requisito (HA3), considero que la propuesta no presenta elementos que permitan satisfacerlo.

En la sección 4, presentaré la propuesta de Goldman (2006) de *mindreading* de primera persona, que consiste en un proceso doble de interpretación e introspección para la autoatribución. Si bien el aspecto interpretativo no está desarrollado, tiene el propósito de dar cuenta de un fenómeno que no se puede dejar de reconocer, presentado en el capítulo 4, la confabulación. Finalmente, en la sección 5, analizaré la propuesta según los requisitos para la autoatribución. Argumentaré que el requisito (AA1) se satisface, aunque con salvedades, y el requisito (AA2) no. En particular, la

propuesta no brinda un criterio que permita establecer satisfactoriamente en qué casos de autoatribución interviene un proceso de tipo interpretativo o introspectivo.

1. *Mindreading* de tercera persona

Goldman (2006) propone un enfoque híbrido de teoría y simulación para *mindreading* de tercera persona. Sin embargo, sólo desarrolla en detalle el aspecto simulacional del mismo mediante la propuesta de un enfoque simulacional para *mindreading* de nivel superior y de nivel inferior. Según la caracterización de Goldman (2006), *mindreading* de nivel superior abarca atribuciones de estados mentales complejos tales como las actitudes proposicionales. Además, algunos de los componentes de este proceso están sujetos al control voluntario y son accesibles a la conciencia. *Mindreading* de nivel inferior, en cambio, abarca estados mentales más simples tales como las emociones y las sensaciones corporales, como el tacto y el dolor. En este caso, el proceso de *mindreading* es primitivo, automático y sin acceso consciente. A su vez, esta propuesta implica cierto refinamiento de la noción de "simulación" y una reconfiguración de la TS con el propósito de hacer frente a la evidencia empírica. Esta evidencia tendrá un rol importante en la caracterización de *mindreading* (mediado por simulación) de nivel inferior y superior.

La noción de "simulación mental" propuesta parte de la predicción de las decisiones ajenas como el caso paradigmático y conserva la idea fundamental de que, en la simulación, los *mindreaders* utilizan su propia capacidad de tomar decisiones para "leer" la mente de los otros. En este sentido, no es necesario poseer una teoría sobre la toma de decisiones en los seres humanos sino, más bien, alimentar el propio sistema cognitivo con el *input* apropiado. Para generar este *input*, se determinan los estados previos del blanco y, a partir de estos, se generan estados ficticios. Este aspecto de la simulación sugiere que el *mindreader* "se pone en el lugar del" blanco. Además, puesto que el sistema de toma de decisiones del *mindreader* es similar al del blanco, el proceso arrojará un *output* que se puede usar para predecir la decisión de este último.

De modo que en la medida en que esta propuesta hace hincapié en un rol para los estados ficticios y en el uso de los mismos mecanismos o procesos cognitivos que se usan *online*, adhiere a la noción de simulación *off-line* (Goldman 2006: 20). No obstante, si bien estos elementos están presentes en la simulación de nivel superior, no tienen ningún rol en la simulación de nivel inferior. Esta última requiere una noción que permita abarcar atribuciones de estados más simples, como las emociones, y que implique procesos automáticos y más primitivos (Fadiga *et al.* 1995; Rizzolatti *et al.* 1996; Grafton *et al.* 1996; Rizzolati, Fogassi & Gallese 2001; Wicker *et al.* 2003; Rizzolatti & Craighero 2004). En particular, una definición según la cual los procesos espejo puedan ser considerados simulaciones. Con este propósito Goldman (2006) refina la noción de simulación aunque, como señalaré en la sección 2, ésta se torna, más bien, confusa.

El refinamiento parte de una noción genérica de simulación según la cual un proceso P es una simulación de P', si P se parece a P' en aspectos significativos (para el propósito de una tarea)[50]. Esta relación es simétrica, dado que nada impide que, a su vez, P' pueda simular a P, si es el caso que P' se parece a P es aspectos significativos. Sin embargo, la simulación mental requiere una relación asimétrica. Con este propósito se agrega la siguiente condición a la definición de simulación:

(S1) *P es una simulación de P'*: si P se parece a P' en aspectos significativos (para el propósito de una tarea) y si, al duplicar P', P satisface su función o propósito.

La simulación mental es aquella en la que un proceso mental se simula mediante otro proceso mental, de modo que:

(S2) *P es una simulación mental de P'*: si P y P' son procesos mentales, y si P y P' ejemplifican la relación de simulación (S1).

[50] Nótese que P y P´ son instancias de procesos y no tipos.

La simulación mental puede ser intrapersonal o interpersonal y se puede utilizar en innumerables tareas, por ejemplo, la visualización (o imaginación visual). Sin embargo, para la tarea de *mindreading* se requiere algo más. Según Goldman (2006), ninguna versión de TS debe sostener que los procesos mentales de los *mindreaders* siempre coinciden, aunque sea aproximadamente, con los del blanco. La simulación tolera casos de *mindreading* incorrecto o, lo que Goldman denomina, intentos de simulación. En este sentido, al *mindreader* puede faltarle información sobre los estados previos del blanco de modo que construya estados ficticios inapropiados que conduzcan a una simulación defectuosa (Goldman 2006: 38). Asimismo, una simulación defectuosa puede ocurrir en el caso de que los estados genuinos del *mindreader* se filtren y alimenten como *input* a la simulación. O bien, puede ser el caso que ciertos estados genuinos del *mindreader* simplemente se proyecten en el blanco (Goldman 2006: 29). De modo que la TS sostiene que *mindreading* consiste en simulaciones mentales exitosas o intentos de simulaciones:

(S3) *Un intento de simulación mental* consiste en que, siendo P y P' procesos mentales, P es ejecutado con el propósito, manifiesto o no, de duplicar o coincidir con P' en aspectos significativos.

El otro aspecto de la redefinición de la simulación consiste en la postulación de la proyección como paso final de una simulación mental (Goldman 2006: 40), con el propósito de dar cuenta de los hallazgos empíricos sobre las tendencias egocéntricas en *mindreading*, observadas en niños y adultos. Si bien este paso ya está implícito en la noción de la simulación *off-line*, aquí se explicita. El rasgo distintivo de la simulación es que el *mindreader* hace un uso especial de la propia mente para atribuir estados mentales a otras personas. Este uso especial consiste en que el *mindreader* toma uno de sus estados mentales (un estado ficticio) y se lo imputa (como un estado genuino) al blanco. El estado ficticio, que alimenta a su sistema de toma de decisiones en funcionamiento *off-line*, y el *output* que resulta de esto son, en rigor, estados mentales

propios del *mindreader* y no creencias acerca de los estados mentales del blanco (como en el caso de la TT). Este *output* (estado mental propio) se utiliza para realizar una atribución al blanco, y así:

(S4) *Proyección* se denomina al acto de asignar un estado propio a otra persona.

Se trata del paso final de la rutina simulacional que no implica ni simulación, ni ficción adicional. Es más, típicamente indica la finalización de la rutina de simulación que, bajo esta especificación, consiste de dos pasos: simulación y proyección. Según Goldman (2006), resaltar el rasgo de la proyección permite apreciar la necesidad de poner en cuarentena los estados mentales genuinos del *mindreader*, que no se corresponden con los estados del blanco. Esto es un requisito para una simulación exitosa. De no ser así, los estados genuinos propios pueden filtrarse en una simulación y conducir a errores o sesgos. Es más, si la filtración de los estados genuinos es desmesurada, ocurren los sesgos egocéntricos (Goldman 2006: 41).

Como he mencionado, esta redefinición le permite a Goldman (2006) afirmar que un proceso de tipo simulacional subyace a *mindreading* de nivel inferior y de nivel superior. En la sección 1.1, me ocuparé de la propuesta simulacional para *mindreading* de nivel inferior. Aquí, se afirma que, según las definiciones (S1)-(S3), un proceso de tipo simulacional subyace a ciertas atribuciones de emociones (*mindreading* de nivel inferior) y que ciertos mecanismos cerebrales, los procesos espejo, pueden considerarse simulacionales. La cuestión de si los procesos espejo subyacen a las atribuciones de estados mentales es controversial, pero Goldman sostiene que hay razones para pensar en esta posibilidad. En la sección 1.2, me ocuparé de la propuesta simulacional para *mindreading* de nivel superior que se identifica con la propuesta simulacional tradicional para las actitudes proposicionales, aunque redefinida como "imaginación enactiva". En este caso, la especificación (S4) del paso de proyección en una rutina de simulación permite dar cuenta de la evidencia sobre los sesgos egocéntricos en *mindreading* de nivel superior.

1.1. La simulación de nivel inferior

La propuesta de *mindreading* simulacional de nivel inferior está relacionada con dos conjuntos de evidencia. Por un lado, el hallazgo de un patrón de déficits conjuntos en la experiencia de cierta emoción y en la atribución de la misma a partir de expresiones faciales. Según los investigadores, este patrón sugiere que la integridad del sustrato para experimentar una emoción es necesaria para realizar atribuciones de la misma. Según Goldman (& Sripada 2005, Goldman 2006) la TS puede explicar la conjunción de los déficits pero la TT no y, en este sentido, se afirma que a este tipo de atribuciones (de emociones a partir de expresiones faciales) les subyace un proceso de tipo simulacional. Por otro lado, cierta evidencia sugiere que los procesos espejo (Iacoboni *et al.* 2005 ver más adelante) pueden subyacer a atribuciones de estados mentales. Según la definición de simulación (S1-S2), los procesos espejo pueden considerarse simulaciones mentales. No obstante, no es claro si los procesos espejo se usan para *mindreading* o no. Goldman (2006) considera que hay razones para sostener que sí.

Se ha acumulado evidencia sobre un déficit en la experiencia de cierta emoción y un déficit selectivo para el reconocimiento facial de esa misma emoción que co-ocurren confiablemente, en particular, para los casos del miedo, el asco y la ira que comentaré a continuación (Goldman & Sripada 2005; Goldman 2006). La amígdala se conoce como la estructura cerebral bilateral cuya función se asocia al condicionamiento del miedo y al almacenamiento de recuerdos emocionales relacionados con el miedo. Se ha descubierto que el daño bilateral de la amígdala causa una experiencia anormal del miedo (por ejemplo, se afrontan las personas y las situaciones con una actitud predominante y excesivamente positiva) y, a la vez, una incapacidad para reconocer la expresión facial del miedo (Adolphs, Tranel, Damasio &

Damasio 1995)[51]. En el caso del asco, se ha observado que la ínsula (especialmente su porción anterior), que tiene un rol en el procesamiento del sentido del gusto en los seres humanos, se activa también durante el reconocimiento de la expresión facial del asco (Phillips *et al.* 1996). En relación con esto, se ha descubierto que el daño de esta región conduce a un déficit conjunto en la experiencia y en el reconocimiento de la expresión facial del asco (Calder *et al*. 2000, Adolphs 2002). Además, usando RMf se ha podido observar que esta región se activa en sujetos normales durante la experiencia del asco provocada por la exposición a olores desagradables y, también, al observar expresiones faciales de asco (Wicker *et al.* 2003). Por su parte, la dopamina se conoce como el neurotransmisor que media la experiencia de la ira y tiene un rol en el procesamiento de la agresión, en diversas especies, durante los encuentros sociales agonistas. En relación con esto, se ha observado que al administrar una droga que bloquea las vías dopaminérgicas (*i.e.* sulpirida), produciendo una disfunción pasajera de las mismas, los sujetos se desempeñan significativamente peor que los controles en el reconocimiento de las expresiones faciales de enojo. Al mismo tiempo, la capacidad para reconocer expresiones faciales asociadas a otras emociones permanece intacta (Lawrence *et al*. 2002). En conjunto, la evidencia mencionada sugiere que es necesario que las estructuras cerebrales que permiten experimentar cierta emoción estén intactas para poder detectar (atribuir) esta misma emoción en otras personas, a partir de su expresión facial.

Según la caracterización de *mindreading* de nivel inferior mencionada en la sección 1, la atribución de emociones basada en expresiones faciales, se puede considerar *mindreading* de nivel inferior, puesto que se trata de una tarea simple en la que sólo hay que reconocer tipos de emociones (*i.e.* miedo, asco, ira) y no contenidos proposicionales. Además, se trata de una tarea automática y sin acceso consciente. Ésta resulta primitiva en la medida en que la "lectura" de emociones tiene un valor

[51] Se han reportado otros estudios de pacientes con daño bilateral de la amígdala (Sprengelmeyer *et al.* 2002) y con actividad reducida de la amígdala (Blair 2002; Blair *et al*. 2004) que muestran la conjunción de los déficits, a saber, la experiencia anormal del miedo y la ausencia del reconocimiento de su expresión facial.

para la supervivencia y es probable que existan procesos cognitivos especializados en esta tarea, que no se utilicen en otras tareas de *mindreading*. Según Goldman & Sripada (2005; Goldman 2006) la simulación puede predecir este patrón de déficit, y su disociación para cada emoción, de la siguiente manera. En analogía con el uso (simulacional) del propio sistema de toma de decisiones, es posible que la gente normal use el mismo equipamiento mental para atribuir una emoción que para experimentarla. Así, la TS predice que si existe algún daño en el equipamiento emocional, éste afectará la habilidad para atribuir adecuadamente las emociones ajenas. Además, en la medida en que el daño no sea tan extenso como para extenderse a otras emociones, la TS predice que el déficit en la tarea de atribuir una emoción a partir de expresiones faciales no se extenderá a otras emociones.

Contrariamente, de acuerdo con todas las versiones de la TT, *mindreading* usa información sobre los estados, en este caso las emociones, pero no usa los estados mismos para realizar las atribuciones. De modo que en la medida en que la TT no afirma ninguna relación entre la experiencia de una emoción y su atribución, la TT no predice la relación entre el sustrato para experimentar una emoción y las atribuciones (basadas en expresiones faciales) de la misma emoción. En este sentido, la TT no ofrece razones para suponer o predecir el patrón de déficits conjunto, menos si se asume que la capacidad de experimentar una emoción y la de razonar sobre las emociones se alojan en lugares diferentes del cerebro (Goldman 2006: 114)[52]. De modo que según

[52] Ahora bien, podría llegar a ser el caso que una región neuronal sea multifuncional de modo que casualmente constituya el sustrato de algún estado mental y, a la vez, del razonamiento sobre ese estado mental. En principio, esto es posible. Por ejemplo, el giro fusiforme derecho es una región multifuncional que da lugar a déficits que no están relacionados funcionalmente entre sí, de manera interesante. El daño de esta región genera acromatopsia (anomalía en la percepción del color) y prosopagnosia (déficit en el reconocimiento de rostros familiares). No obstante, según Goldman (2006), aunque se conceda esto, la TT tampoco parece poder explicar la evidencia relacionada con las emociones. Supóngase, según la versión modular de la TT, que existe un módulo dedicado a *mindreading* de las emociones con la propiedad de un sustrato neuronal coincidente con el sustrato para experimentar emociones. En este caso, si el sustrato neuronal para experimentar las emociones estuviera dañado, se predice un daño en la atribución de emociones en general. La predicción abarca el daño de todos los tipos de emociones que el modulo "lee", pero no predice el daño selectivo para cada emoción. De manera similar, según la versión del niño científico de la TT, si estuviera dañado el sistema de razonamiento de dominio general que subyace a *mindreading*, se predice un daño para "leer" los estados emocionales en general (así como otros estados), pero no hay razones para esperar el daño

Goldman (2006), la simulación subyace a la atribución de emociones a partir de su expresión facial porque los déficits conjuntos antes mencionados sugieren que es preciso que el sustrato cerebral para experimentar una emoción esté intacto para poder realizar las atribuciones correspondientes. Esta es una predicción que la TS puede realizar pero la TT no.

Asimismo, una serie de hallazgos ha conducido a los científicos a establecer, en los seres los seres humanos, la existencia de un amplio rango de mecanismos espejo interpersonales en los cuales ciertos estados cognitivos de un organismo son imitados por estados cognitivos similares en un organismo observador (Fadiga *et al*. 1995; Rizzolatti *et al*. 1996; Grafton *et al*. 1996; Rizzolati, Fogassi & Gallese 2001; Wicker *et al*. 2003; Rizzolatti & Craighero 2004). Los procesos espejo en humanos resuenan en respuesta a un rango mayor de acciones que en los monos. A saber, no sólo respecto de acciones manuales, sino también de acciones relacionadas con los pies, la boca, incluso, con los sonidos asociados a las acciones (Buccino *et al.* 2001). Además, estos responden al tacto, al dolor y a ciertas emociones tales como el asco. Según Goldman (2003), los procesos espejo pueden considerarse simulaciones en virtud de que concuerdan con la idea general de la simulación (definiciones S1 y S2). En los procesos espejo, un proceso, que es relevantemente similar a otro, tiene lugar con el propósito, o la función, de duplicar al primero (Goldman 2006: 132)[53]. Así, un evento de procesos espejo es un caso potencial de un acto de *mindreading* correcto, pero queda por establecer si, y en qué medida, los procesos espejo se usan para *mindreading*. No obstante, siempre que haya un proceso espejo y *mindreading* basado en tal proceso espejo, habrá simulación de nivel inferior (Goldman 2006: 133).

---

selectivo de una emoción (Goldman 2006: 114-115). De modo que suponiendo que la TT pudiera explicar la conjunción de los déficits en la experiencia y la atribución emocional, aún así no podría predecir el patrón de daño selectivo para cada emoción. La ocurrencia de la conjunción de déficits de experiencia y reconocimiento para distintas emociones y numerosas clases de estímulos, hace poco probable que se trate de un caso de co-localización de déficits no relacionados funcionalmente (Goldman 2006).

[53] Es preciso mencionar que no siempre que dos personas comparten un estado mental hay procesos espejo. Si estos se comparten de manera accidental, se trata de un caso de mera sincronía mental.

Ahora bien, no es claro que los procesos espejo se utilicen para atribuir estados mentales a otras personas en tanto la atribución de un estado mental al blanco requiere algo más. Según Goldman (2006), requiere de dos actos mentales por parte del *mindreader*: seleccionar una categoría de estado mental, o clasificación, e imputar la instancia de tal clasificación en un blanco pertinente. Como he mencionado, los procesos espejo sólo requieren que el *mindreader* genere un evento coincidente, y esto no garantiza que éste tenga un repertorio para la clasificación, ni que, en caso de poseerlo, pueda usarlo en esa ocasión. Es más, los procesos espejo no implican *mindreading* porque puede ser el caso que el observador no le adscriba nada al actor (Goldman 2006: 133). No obstante, si bien los procesos espejo no implican *mindreading*, por definición estos pueden considerarse simulaciones mentales y, en tanto simulaciones mentales, es posible que puedan subyacer a *mindreading*.

Goldman (2006) deja abierta la posibilidad de que los procesos espejo intervengan en *mindreading* y, es más, considera que los siguientes experimentos permiten afirmar que los procesos espejo subyacen a *mindreading*. En particular, según Goldman (2006), los estudios del asco donde se observa una activación neuronal solapada en la ínsula anterior, para la experiencia del asco y en la observación de expresiones faciales de asco en otras personas (Wicker *et al.* 2003), sugieren que parece haber una relación contrafáctica entre la integridad del sustrato para experimentar el asco y la facilitación de la atribución interpersonal del asco. “Si uno no tiene el sustrato para el asco intacto, uno no podrá realizar atribuciones de asco normalmente”. Según Goldman (2006), este contrafáctico brinda apoyo a la conexión causal entre experimentar asco mientras se observa una expresión de asco y atribuir asco a la persona que tiene la expresión. En este sentido, se considera que la evidencia sugiere el uso de la experiencia del asco como base causal para la atribución de tercera persona.

Un estudio en particular le sirve a Goldman para sostener su afirmación del rol de los procesos espejo en *mindreading*. El estudio de Iacoboni *et al*. (2005) indaga sobre si las neuronas espejo (NE) codifican la acción o, más bien, la intención. Las

propiedades básicas de las NE pueden ser interpretadas como un mecanismo para reconocer los actos motores tales como agarrar, sostener, llevarse a la boca, sin necesidad de imputarle al blanco un estado mental de intención (de alcanzar una meta). Este estudio intenta discernir si los procesos espejo motores codifican simplemente la acción o más bien las metas, en el sentido de algo que explica por qué el agente hace lo que está haciendo. Si la evidencia sugiere que las NE codifican las metas, esto puede brindar apoyo a la idea de que los procesos espejo motores codifican la intención.

Iacoboni y sus colaboradores advierten que la misma acción llevada a cabo en dos contextos diferentes puede reflejar distintas intenciones. De modo que diseñaron un experimento para investigar si la misma acción de agarre, incrustada en contextos diferentes, produce en el observador la misma actividad neuronal, o una diferente, en las áreas de NE asociadas con la acción de agarre. Los sujetos escaneados observan tres filmaciones como estímulo: (1) agarre con la mano en ausencia de objeto (condición "acción"), (2) escenas con tazas y platos (condición "contexto") y (3) acciones de agarre llevadas a cabo en dos contextos diferentes, durante el té y después del té (condición "intención")[54]. De modo que se espera que si el sistema de NE codifica simplemente el tipo de acción, se observará la misma activación en las condiciones ((1) condición "acción" y (3) condición "intención"). Si, en cambio, las NE codifican la meta asociada con la acción de agarre, se espera una modulación de la actividad en las áreas de NE en la (3) condición "intención". Según los resultados, en comparación con las condiciones (1) y (2), la condición (3) de intención produjo un incremento significativo en la activación de las áreas premotoras (asociadas a las NE). Según los investigadores, este incremento en la activación sugiere que esta área cortical no provee simplemente un mecanismo de reconocimiento de la acción ("eso es un agarre") sino que es crítica para entender las intenciones (metas) de las acciones de los otros.

Además, según Goldman (2006), el rol de los procesos espejo motores en la atribución de intenciones queda sugerido por otro aspecto de este estudio. En el

[54] La tercera condición se denomina "intención" porque cada contexto sugiere una intención asociada con el agarre manual, tomar una taza para beber té o tomar una taza para lavarla.

estudio, los participantes recibieron instrucciones diferentes. La mitad tenía que mirar simplemente las filmaciones (tarea implícita). La otra mitad fue instruida para inferir la intención que conlleva la acción de agarre de acuerdo con el contexto (tarea explícita). Luego del escaneo cerebral, los sujetos fueron entrevistados. Llamativamente, todos los participantes asociaron la intención de tomar el té a la acción de agarre en la condición "durante el té" y la intención de lavar a la acción de agarre durante la condición "después del té", con independencia de la instrucción recibida. Es más, todos los participantes mostraron el mismo patrón de activación de la corteza frontal derecha en la condición (3) en comparación con las condiciones (1) y (2). Y, además, no mostraron diferencias en la activación a pesar de haber recibido instrucciones diferentes (tarea implícita y explícita). Según los investigadores, esto sugiere que las NE hacen la misma contribución a la interpretación de las acciones de agarre, aún cuando el participante no haya recibido la instrucción de inferir las intenciones. En base a la evidencia mencionada, Goldman (2006) afirma que la atribución basada en procesos espejo es contundente para los casos de las emociones, las sensaciones (el dolor y el tacto) y las intenciones (motoras). Además, dado que los procesos espejo son una especie de simulación interpersonal proveen apoyo para sostener que *mindreading* de nivel inferior se lleva a cabo por simulación[55].

Mas allá de la evidencia y el entusiasmo de Goldman (2006), a mi entender, hay una brecha entre los procesos espejo y *mindreading* que puede entenderse como la brecha entre ciertos procesos primitivos y la cognición superior. Este es un problema

[55] Otro estudio que reseña Goldman (2006) con este propósito, es el estudio de Jackson *et al.* (2004), donde los sujetos son expuestos a imágenes de manos y pies en condiciones dolorosas y neutrales. Se les solicita que evalúen (o realicen un juicio sobre) la intensidad del dolor que siente la persona en la foto, pero a la que no se le ve la cara, sólo las manos y los pies. Uno de los resultados de este experimento es que mirar a otros en situaciones dolorosas activa parte de la red neuronal implicada en el procesamiento del dolor propio. Según los investigadores, esto señala la presencia de procesos espejo para el dolor. El otro hallazgo es que se observa una correlación fuerte entre las evaluaciones (atribuciones) de la intensidad del dolor y el nivel de actividad en la región posterior de la corteza cingulada anterior, una parte crucial de la red de procesamiento del dolor. Según Goldman (2006), estos hallazgos confirman la idea de que una sensación inducida por procesos espejo puede servir de base causal para una atribución de *mindreading* de tercera persona, aunque la correlación no establece causalidad (Goldman 2006: 138, nota 18).

que los investigadores empíricos tienen en mente y, por esto, muchos sostienen que los procesos espejo asociados a emociones dan lugar a la empatía, un tipo de estado menos asociado con la clasificación mental y la conceptualización (Gallese 2001; Wicker *et al.* 2003; Decety 2010), y los procesos espejo motores dan lugar a la imitación (Rizzolatti *et al.* 2002; Iacoboni 2008). Por ejemplo, Wicker *et al.* (2003), consideran que probablemente la comprensión de las emociones de los otros depende de sistemas múltiples interrelacionados, entre los cuales se encuentran los de resonancia automática. Es probable que en los seres humanos se haya agregado, a los mecanismos primitivos que compartimos con los animales, otras rutas cognitivas para comprender las emociones (Frith & Frith 1999). Es más, según el enfoque cognitivo, una de estas rutas consiste en el procesamiento de la expresión facial en las regiones corticales visuales que conduce a la representación proposicional del estado de asco inferido (Wicker *et al.* 2003). La cuestión es que parece necesario conectar estos mecanismos primitivos de resonancia con mecanismos de cognición superior, que permitan conceptualizar y verbalizar para que tenga lugar la atribución mentalista. A continuación caracterizaré la brecha en términos de Goldman y mostraré que esta tensión aparece en su propuesta, más allá de los argumentos a favor de los procesos espejo subyacentes a *mindreading*.

Según Goldman (2006) hay sistema espejo:

> …cuando hay un sendero causal, repetible y sistemático que parte del estado mental de un individuo y llega a coincidir con el estado mental en el observador. Cuando este tipo de sendero existe mediado por mecanismos neuronales, debe denominarse sistema espejo. Tal sendero sistemático tiene dos componentes: un sub sendero en el blanco, desde el estado mental hasta la expresión comportamental del mismo, y un sub sendero, en el observador, desde la observación del comportamiento del blanco hacia el estado mental que coincide con el del blanco. Mediante este sendero doble el receptor "imita" un estado

> mental del emisor. El receptor puede o no saber (o creer) que tal evento espejo ocurre. Puede saber o no sobre tal evento mental (tales eventos no precisan ser conscientes), e incluso si la persona es consciente del estado mental puede no saber que coincide con el evento que ocurre en el emisor. Se puede no pensar en el emisor, ni conectar el propio evento mental con el evento del emisor. (Goldman 2006: 133)

Ahora bien, justamente esta caracterización de procesos espejo me permite ilustrar mi punto. Si retomamos la caracterización de la rutina simulacional como simulación-proyección, en base a la definición de procesos espejo es claro que estos pueden dar cuenta del aspecto de simulación, en el sentido en que se genera un estado en el observador que coincide con el del blanco. En la medida en que el proceso puede finalizar en este punto, los procesos espejo no dan cuenta de aspecto de proyección que, según la definición (S4), consiste en el acto de asignar un estado propio a otra persona. Justamente, este último paso permite que se efectúe la atribución mediante la simulación y se complete *mindreading*.

La cuestión es que si los procesos espejo pueden suceder sin *mindreading*, ¿qué hace que suceda *mindreading?* Como mencioné, la brecha entre *mindreading* y los procesos espejo puede caracterizarse como la ausencia del elemento de proyección constituyente de la rutina simulacional. Además, Goldman (2006: 133) sostiene que la diferencia entre los procesos espejo y *mindreading* consiste en clasificar el estado e imputar la instancia de tal clasificación al blanco. A mi entender esto parece sugerir que, en términos de mecanismos cognitivos, para que un proceso espejo subyazca a *mindreading* sería preciso que estos se vincularan con procesos de atribución (clasificación y proyección). En este sentido, parece necesario que los *outputs* de los procesos espejo alimenten algún mecanismo atributivo que permita concretar *mindreading*. Sin embargo, la separación entre procesos simulacionales y atributivos no es parte de la propuesta de Goldman. En mi opinión, aquí hay una tensión. Es más, esta tensión surge en la propuesta misma.

Al discutir la atribución de intenciones en base a procesos de resonancia motora (el experimento de Iacoboni *et al*. 2005) Goldman se adelanta y responde a una objeción. Sostiene que alguien podría cuestionarle que la atribución de intención basada en procesos espejo se trata, más bien, de una atribución de nivel superior y no de nivel inferior. En su defensa, Goldman sostiene que la atribución de intención al nivel de los sistemas motores consiste en codificar una acción como parte de una secuencia de actos motores y que esto implica un nivel cognitivo de funcionamiento inferior. Y que cuando los sujetos son entrevistados verbalmente, los centros cognitivos superiores se ponen en juego y caracterizan el comportamiento en términos proposicionales (Goldman 2006: 140). De modo que parece reconocer que para llevar a cabo atribuciones, los procesos de resonancia motora tienen que conectarse con procesos atributivos de nivel (o cognición) superior. En conclusión, a mi entender, la propuesta de simulación de nivel inferior da cuenta de que ciertos proceso primitivos pueden entenderse como simulacionales, pero esto no alcanza para sostener que estas simulaciones (los procesos espejo) subyacen a atribuciones de estados mentales. Básicamente, los procesos espejo no pueden dar cuenta de los aspectos de clasificación (y conceptualización) y proyección que permiten completar *mindreading*.

1.2. La simulación de nivel superior

Como señalé en la sección 1, *mindreading* de nivel superior se caracteriza como *mindreading* de estados mentales complejos, tales como las actitudes proposicionales, donde algunos aspectos del proceso de *mindreading* pueden estar sujetos a control voluntario (o *top-down*) y tener cierto grado de accesibilidad consciente. Con el propósito de completar esta caracterización, Goldman (2006: 147) sostiene que los procesos simulacionales de tipo espejo pueden considerarse como los prototipos de *mindreading* de nivel inferior conducido por simulación, mientras que el uso de la ficción o la "imaginación enactiva", puede considerarse como el prototipo de

mindreading de nivel superior. De modo que con la caracterización de la imaginación enactiva queda caracterizado *mindreading* de nivel superior conducido por simulación.

La "imaginación enactiva" no es la imaginación tal como la entiende el sentido común, sino que es un término técnico que Goldman (2006) utiliza para introducir un constructo psicológico que identifica con la operación mental de ficción. Se trata de un proceso u operación cuyos *outputs* se denominan estados ficticios. Este tipo de imaginación no genera suposiciones (el producto que se asocia con la imaginación ordinaria), sino muchos tipos de estados mentales tales como sensaciones, emociones, percepciones, creencias, deseos, esperanzas y demás (Goldman 2006: 47). Los estados ficticios no son meras suposiciones puesto que al imaginar enactivamente, por ejemplo "estar eufórico", este estado se reproduce, aunque atenuado. De modo que un estado ficticio es el producto de la imaginación enactiva y los estados ficticios son instancias de creencias, deseos, y demás estados, que son generados endógenamente (o de modo *top-down*) con el propósito de parecerse a las creencias, los deseos, las percepciones y demás estados que, usualmente, se generan exógenamente (Goldman 2006: 48).

La imaginación enactiva se puede utilizar para numerosas tareas, por ejemplo, la visualización. La TS sostiene que la imaginación enactiva se utiliza para *mindreading*. Según Goldman (Goldman 2006: 149), el hecho de que esta capacidad se utilice para *mindreading* y el hecho de que *mindreading* genere a menudo atribuciones correctas, implica que es necesario que los *outputs* de la imaginación enactiva puedan parecerse, en aspectos relevantes, a sus contrapartes, puesto que las atribuciones que se generan suelen ser correctas. No obstante, según la definición (S3), la simulación de nivel superior no requiere que los estados mentales replicados sean confiablemente correctos, sino basta con que estos se hayan generado con el propósito de replicar estados mentales (Goldman 2006: 150). La cuestión es que la imaginación enactiva no depende sólo de la capacidad para imaginar, sino de la información específica para la tarea. Si esta información falta, esto no debe considerarse una falla en la imaginación. Por ejemplo, si nunca vimos a Helena de Troya la imaginación de su rostro no será

correcta por falta de información. De modo que *mindreading* basado en la imaginación enactiva puede ser guiado por conocimiento. Ahora bien, que los estados ficticios se generen a partir de información, por ejemplo almacenada en la memoria, no impide que se trate de un proceso simulacional. Lo que indica que se trata de una simulación mental es que el estado ficticio se genera de manera endógena con el propósito o la función de replicar un estado mental.

Goldman (2006) reseña abundante evidencia experimental con el propósito de fundamentar su propuesta de imaginación enactiva, en particular, respecto de tres aspectos de la simulación como proceso subyacente a *mindreading* de nivel superior. Primero, los estados ficticios son suficientemente similares a los estados mentales que replican. Segundo, es necesario que el *mindreader* monitoree sus estados mentales, por ejemplo, para identificar aquellos que tienen que inhibirse o ponerse en cuarentena con el propósito de excluirlos de la rutina de simulación. Tercero, la necesidad de poner en cuarentena los estados mentales propios en una rutina de simulación. Estos dos últimos aspectos fueron mencionados como características de la simulación en la sección 1.

La estrategia argumentativa de Goldman (2006) consiste en mostrar que la imaginación enactiva es un fenómeno robusto, capaz de producir *outputs* que son similares a sus estados contraparte. Con este propósito reseña evidencia sobre la imaginación visual y motora, ya que en estos dominios se ha indagado sobre la similitud entre los productos de la imaginación y sus contrapartes genuinas[56]. No

[56] En el caso de la imaginación visual, la similitud con la percepción visual ha sido sugerida por evidencia comportamental y, principalmente, neurológica (Kosslyn 1994; Spivey *et al*. 2000). La evidencia comportamental sugiere, por ejemplo, que los movimientos sacádicos del ojo también ocurren durante la imaginación visual y que estos se aproximan a los que ocurren durante la percepción visual. Se ha observado, además, que la imaginación visual interfiere con la percepción. La evidencia neurológica sugiere que las estructuras cerebrales asociadas a la percepción visual intervienen, al menos en parte, en la imaginación visual. Esto se ve apoyado por estudios de neuroimagen que sugieren la activación cerebral solapada durante la visión y la imaginación visual, entre otros hallazgos. Por otro lado, la imaginación motora alude a una representación de la ejecución de un movimiento corporal que, en general, no es consciente (en esto se diferencia de la imaginación visual) pero que está sujeta a control voluntario. Los estados contraparte de la imaginación motora son los eventos de generación de movimientos que ocurren en la corteza motora que produce el comportamiento. Usualmente, en los estudios de imaginación motora se instruye a los sujetos para que se imaginen ciertos movimientos y los sujetos reportan cumplir con tales instrucciones. Entre los hallazgos, se ha observado un incremento de

obstante, se asume que la imaginación enactiva abarca un rango de estados mentales más amplio y, en este sentido, quizás esta evidencia no sea representativa. Sin embargo, Goldman argumenta que si la evidencia para estos casos es robusta puede esperarse algo similar para los otros estados mentales.

Otro aspecto de la simulación que, según Goldman (2006) parece encontrar apoyo evidencial es la necesidad del monitoreo de los estados mentales propios (autoreflexión). Ciertos estudios de neuroimagen sugieren que la actividad cognitiva de autoreflexión está implicada en tareas de *mindreading* de tercera persona. Existe un amplio acuerdo sobre la contribución crítica de la corteza prefrontal medial (CPFm) al entendimiento de la mente de las otras personas. A su vez, una serie de estudios sugiere que la CPFm ventral está selectivamente involucrada en tareas de autoatribución, autoreflexión e introspección. Según Mitchell, Banaji & McRae (2005) esto indica que debe favorecerse una versión de TS para *mindreading*, puesto que este enfoque postula el uso de la autoreflexión como una estrategia para "leer" las otras personas. En particular, el estudio evalúa el vínculo entre la autoreflexión y *mindreading* de tercera persona. Los participantes llevan a cabo una tarea que consiste en evaluar cuán contenta está de haber sido fotografiada la persona que aparece en una fotografía (tarea de *mindreading*) y una segunda tarea que consiste en determinar cuán simétrico es el rostro que aparece en la fotografía (tarea de no-*mindreading*), mientras son escaneados con RMf. Una tercera tarea, que tiene lugar 30 minutos después del escaneo, consiste en evaluar cuán similar a uno mismo es la persona que aparece en la fotografía. En concordancia con lo esperado, el contraste de la activación cerebral en la tarea de *mindreading versus* la tarea de no-*mindreading* muestra una

---

un 22% en la fuerza muscular para los movimientos imaginados en comparación con un incremento del 30% para los movimientos ejecutados. Los investigadores sostienen que este efecto es producto de la activación cortical y no de activación muscular encubierta, porque no se registraron contracciones musculares en los sujetos durante la imaginación motora (Yue & Cole 1992). Además, una serie de hallazgos sugiere que la simulación mental de la acción se basa en las mismas estructuras que se usan para la acción real. La imaginación motora del movimiento parece imitar el tiempo del movimiento real, por ejemplo, el tiempo de una caminata imaginaria es similar al tiempo de una caminata real. Estudios con RMf sugieren que los mismos pixeles se activan durante la contracción muscular y durante la imaginación de movimientos para los mismos músculos.

mayor activación de las regiones cerebrales asociadas a *mindreading* de tercera persona para la condición de *mindreading*, incluida la CPFm dorsal. Además, se pudo apreciar un efecto de similitud en las tareas de *mindreading*. Se observó no sólo la activación de la CPFm ventral, asociada a los procesos de autoreflexión sino, además, una correlación entre el grado de similitud estimado y la activación de esta región. Estos hallazgos son interpretados por los investigadores como evidencia a favor de la predicción de TS según la cual entender los estados mentales de otras personas consideradas similares a uno mismo depende de la autoreflexión. Según Goldman (2006), este estudio brinda apoyo a la TS, porque este enfoque puede predecir el uso de la autoreflexión en *mindreading* de tercera persona, mientras que la TT no.

En mi opinión, este experimento no brinda apoyo directo a la propuesta de Goldman (2006) sino quizás a alguna noción general de la TS. Este estudio supone que si las personas llevan a cabo juicios de similitud respecto de las personas a las que les atribuyen estados mentales, del tipo “el otro es igual a mi”, el proceso que subyace es simulacional. Según los investigadores, este tipo de juicios sugeriría que para realizar atribuciones de tercera persona uno utiliza los propios recursos cognitivos, tal como sostiene la simulación. Si bien considero que de la afirmación del uso de los propios recursos no se sigue necesariamente que las otras personas llevan a cabo juicios de tipo “el otro es igual a mi”, los críticos de la simulación han sugerido algo similar a esto. Para algunos, la simulación está basada en una generalización: “los otros son iguales a mi”. De no ser así, se asume que no podrían utilizarse los recursos propios como modelo del otro. La creencia en esta generalización implica que *mindreading* simulacional depende de una teoría tácita (Jackson 1999). Sin embargo, Goldman (2006: 30) no cree que sea necesario recurrir a la generalización. No obstante, si se llegara a mostrar que la premisa “el otro es igual a mi” es necesaria en la rutina de simulación y que esto implica usar una generalización, esto no le parece preocupante puesto que adhiere a un enfoque híbrido de teoría y simulación. Además, según Goldman (2006), la premisa formaría parte de una rutina simulacional pero la atribución se llevaría a cabo mediante simulación, y esto es lo que importa. Sin

embargo, a mi entender, en tanto Goldman afirma que el juicio "el otro es igual a mi" no es necesario para la simulación, la evidencia que sugiere que las personas llevan a cabo juicios de similitud al realizar atribuciones mentalistas no brinda apoyo evidencial directo a su propuesta simulacional, más allá de que la misma sea compatible con tales juicios. Goldman sólo sostiene que si se llegara a mostrar que los juicios de similitud son parte de la simulación, su propuesta sería compatible con la evidencia.

En segundo lugar, según Mitchel *et al.* (2004), la TS sostiene que, cuando se juzga al otro como similar a uno mismo, se recurre a la autoreflexión como estrategia para atribuir estados mentales a los otros. Es necesario mencionar que "autoreflexión" y "autoatribución" se entienden de muchas maneras. En el estudio empírico de Mitchell y colaboradores, el término "autoreflexión" refiere a un juicio de similitud respecto de otra persona. De modo que por "autoreflexión" se entiende juzgar rasgos de personalidad, y esto no tiene mucho que ver con lo que interesa en el debate teoría-simulación, a saber, la autoatribución de creencias, deseos y demás estados mentales que, según Goldman (2006) y Nichols & Stich (2003), puede llevarse a cabo como un monitoreo de los estados mentales propios. De modo que en este estudio no se considera una noción de "autoreflexión" compatible con la noción que se tiene en cuenta en el debate teoría-simulación (Apperly 2008). A mi entender, no es claro que la noción de "autoreflexión" utilizada por Mitchell, Banaji & McRae (2005) sea útil para evaluar si la predicción que se sigue del estudio puede ser sostenida por un enfoque de TT o de TS. Me ocuparé con más detalle de esta cuestión en el capítulo 6.

El tercer cuerpo de evidencia que brinda apoyo a la propuesta simulacional para *mindreading* de nivel superior es el de los sesgos egocéntricos. En este caso, la evidencia brinda apoyo al aspecto de proyección de la rutina simulacional, más que a algún rasgo de la imaginación enactiva. Según Goldman (2006), la evidencia sugiere cierto patrón de error específico en *mindreading* que la TS puede explicar pero la TT no. El patrón de sesgos egocéntrico hallado sugiere que la raíz de, al menos, algunos sesgos egocéntricos reside en una falla para inhibir la propia perspectiva. Además, la

propuesta de cuarentena para estados mentales genuinos durante la simulación puede entenderse a nivel cerebral como inhibición (Goldman 2006: 170).

Goldman reseña evidencia de sesgos egocéntricos en relación al conocimiento, las valoraciones y las sensaciones[57]. Según Goldman todos estos casos son interpretables en términos de rutinas simulacionales afectadas incorrectamente por estados mentales no sometidos a cuarentena. Por ejemplo, el conocido estudio de Camerer, Loewenstein & Weber (1989) indaga sobre situaciones en las que se le requiere a personas bien informadas que predigan el pronóstico de las ganancias que llevarán a cabo personas menos informadas. Este estudio, y otros estudios subsiguientes, establecen la tendencia en los adultos a evaluar sesgadamente el conocimiento de una persona menos informada (o ingenua), según su propio conocimiento ("la maldición del conocimiento"). Este caso permite apreciar que si bien los participantes están notificados del hecho de que cierta información que poseen les es propia y que no la comparten con el blanco, la proyectan de todos modos. En este sentido, según Goldman (2006: 168), es evidente que estos estudios muestran que las personas son incapaces de poner en cuarentena su conocimiento (valoraciones o sensaciones) al realizar predicciones sobre las otras personas. De esta evidencia se sigue que el conocimiento y las creencias propias se usan en *mindreading*.

Los hallazgos concuerdan con la TS porque la cuarentena, o la inhibición de los estados mentales propios, tiene un rol importante en la rutina de TS, pero no lo tiene

[57] En relación a las sensaciones, Van Boven & Loewestein (2003) estudiaron la predicción de estados hipotéticos de hambre y sed. En el estudio se les requiere a los participantes que predigan las sensaciones de un grupo de excursionistas perdidos en el bosque sin agua y sin comida. Los participantes realizan las predicciones en dos condiciones, antes y después de realizar ejercicio físico intenso, que presumiblemente les genera sed y calor. Los que se ejercitaron se mostraron más propensos a predecir que los excursionistas estarían más afectados por la sed que por el hambre. A los participantes también se les pregunta cómo se sentirían si ellos fueran los excursionistas. De la misma manera, los participantes que se ejercitaron, en comparación con los que no, tienden a predecir que estarían más afectados por la sed que por el hambre en el caso de los excursionistas. Esto sugiere que también hay una relación en las auto-predicciones, esto es, los participantes tienden a proyectar los estados mentales ocurrentes, en este caso las sensaciones, en situaciones hipotéticas. De modo que las sensaciones de sed y calor de los participantes se asocian positivamente a sus predicciones sobre las sensaciones de los excursionistas. Van Boven & Loewestein (2003) concluyen que las personas predicen como se sienten los otros imaginándose como se sentirían ellos mismos en la misma situación.

en la TT. Como he mencionado en el capítulo 4, la TT propone una explicación de los sesgos egocéntricos a partir de la posesión de teorías defectuosas que generan predicciones erróneas pero, según Goldman (2006), esta hipótesis no concuerda con la tesis de la inhibición. Es raro que un "teórico de la teoría" sostenga que las inferencias de *mindreading* implican creencias genuinas del agente que pueden o tienen que inhibirse. A lo sumo, éstas implican creencias acerca del blanco y acerca de generalizaciones de la psicología de sentido común (Goldman 2006: 172). Además, según Goldman, su propuesta da cuenta de evidencia adicional que permite sostener que este tipo de errores se explican como problemas de cuarentena o inhibición, y no con teorías erróneas.

Samson *et al.* (2005) estudiaron un paciente (WBA) con daño en un área cerebral asociada con la habilidad para inhibir la perspectiva propia (Vogeley *et al.* 2001). Según los investigadores, esta incapacidad condujo a WBA a cometer errores egocéntricos en numerosas tareas de *mindreading* de tercera persona (falsa creencia, perspectiva visual, perspectiva emocional y atribución de deseos). La evidencia en relación a WBA parece mostrar que la raíz de los errores egocéntricos en *mindreading*, consiste en una anomalía en la inhibición de la perspectiva propia. Si la simulación implica inhibición de la auto-perspectiva, se sigue que alguien con un daño en la capacidad de inhibir la perspectiva propia tendrá problemas en generar *mindreading* correcto. Esto es lo que la TS predice y esta predicción se ve confirmada en el caso de ese paciente con daño neurológico. De modo que esto favorece la propuesta de TS de Goldman (2006). En la sección 3, al evaluar la propuesta de Goldman (2006) según el requisito (HA4) relacionado con los sesgos egocéntricos, me ocuparé de las limitaciones de la explicación de TS de los sesgos egocéntricos y, en este sentido, mostraré que el apoyo empírico que esta evidencia otorga a TS se debilita.

2. Acerca del carácter híbrido de teoría y simulación

Goldman (2006) propone un enfoque híbrido de teoría y simulación donde los roles para los procesos de teoría y simulación apenas se establecen estipulando modos en los que pueden combinarse para generar *mindreading*. En el capítulo 3 (sección 3.3), sostuve que es un requisito para un enfoque híbrido de teoría y simulación proponer un criterio para determinar el tipo de proceso que subyace a un caso de *mindreading*. Sin embargo, esta propuesta apenas estipula las posibles relaciones entre teoría y simulación. Goldman considera que las relaciones más plausibles entre teoría y simulación son la independencia y la cooperación (Goldman 2006: 41-46). A continuación analizaré estas relaciones y argumentaré que no son suficientes para establecer un criterio que permita distinguir entre teoría y simulación.

La relación de independencia alude a la posibilidad de que ciertas tareas de *mindreading* se lleven a cabo completamente por simulación o completamente por teoría. Esta relación está poco descripta y no es claro qué implica. En principio, es posible pensar que ambos procesos, en tanto independientes, son suficientes para producir *mindreading* por separado. Sin embargo, no se especifican las condiciones según las cuales actúan los procesos de teoría y simulación por separado. No se sabe si cada proceso se pone en funcionamiento en casos particulares, o en un rango de casos o en todos los casos. Dada la falta de especificación, quedan abiertas todas las posibilidades. Además, la relación de independencia propuesta deja abierta la posibilidad del funcionamiento simultáneo de los procesos de teoría y simulación. Si esto es posible, surgen otras cuestiones tales como qué mecanismo determina el *output* de *mindreading* cuando los procesos funcionan en simultáneo.

En cuanto a la cooperación, se asume, más bien, que la teoría ayuda a establecer las condiciones para que un proceso simulacional se inicie. Esto se aprecia en los ejemplos de cooperación propuestos por Goldman (2006). Primero, es posible recurrir a teorización para inferir los estados previos del blanco para los que, luego, se generan estados mentales ficticios, que alimentan (como *input*) al sistema de toma de decisiones propio. Éste, que funciona en modo *off-line*, arroja un *output* que se usa para realizar una atribución al blanco. Nótese que si bien el primer paso se lleva a cabo

por teorización, el proceso mediante el cual se genera la atribución es de tipo simulacional. El segundo ejemplo de cooperación se relaciona con la estrategia de "generación y testeo". Como mencioné en el capítulo 2, según Goldman no es posible iniciar una rutina de simulación utilizando como *input* un comportamiento y obtener como *output* los estados que le dieron origen. Sin embargo, muchas veces las atribuciones se realizan en este sentido en particular, las explicaciones en términos mentalistas. Para dar cuenta de este tipo de atribuciones, el enfoque simulacional propone la estrategia de "generación y testeo". En esta estrategia se generan hipótesis sobre estados mentales que pueden dar lugar a la conducta de interés. Luego, estos se ponen a prueba por simulación. En este contexto, es posible usar generalizaciones sobre las creencias y los deseos que las personas pueden tener en determinadas circunstancias, para generar hipótesis que luego serán testeadas por simulación. Si la decisión que se produce por simulación coincide con la acción de la persona observada, la hipótesis es confirmada y la adscripción aceptada. Si no, se repite el proceso hasta encontrar la coincidencia (Goldman 2006: 45). En ambos casos, la teoría no es suficiente para producir el *output* de *mindreading* sino, más bien, viene a establecer las "condiciones iniciales" para una simulación.

Claramente, la teoría y la simulación se conciben como sub-competencias que colaboran para llevar a cabo *mindreading*. No obstante, esta propuesta no brinda un criterio satisfactorio que permita establecer qué tipo de proceso subyace a un caso de *mindreading*, en general, por la falta de descripción de la interacción entre teoría y simulación y, en particular, por la falta de descripción del rol de la teoría. En este sentido, considero que la propuesta de Goldman (2006) no cumple con el requisito de un criterio para determinar el tipo de proceso subyacente para los enfoques híbridos de teoría y simulación. La propuesta de Goldman resulta más bien una propuesta de simulación refinada. No obstante, una propuesta con limitaciones.

Como señalé anteriormente, la redefinición de la simulación conduce a una noción confusa. Por un lado, se agrupan una serie de características con las que tiene que cumplir algo para ser una simulación (las definiciones S1-S4) y, por otro lado, se

determinan unos prototipos de simulación de nivel inferior (SIN) o superior (SNS) para *mindreading* (los procesos espejo y la imaginación enactiva, respectivamente). No obstante, en mi opinión, ya no queda claro de qué se trata ser una simulación, si es un proceso cognitivo, si es un proceso cerebral o si meramente es una relación entre eventos, que a veces pueden ser procesos cognitivos y otras veces algún otro evento. Es más, no parece quedar claro si la SNI y la SNS comparten algún rasgo común, algo así como un sentido mínimo de simulación bajo el cual ambas puedan considerarse una simulación.

De modo que las nociones "*mindreading* de nivel inferior" y "de nivel superior" no están bien definidas. Goldman advierte esto y para intentar aclarar las nociones propone prototipos de *mindreading* de SIN y SNS: los procesos espejo y la imaginación enactiva, respectivamente. Pero qué hace que estos procesos, tan distintos, sean ambos simulaciones. Se podría pensar que la condición mínima queda establecida en la definición (S3) de ser un proceso mental ejecutado con el propósito de duplicar o coincidir en aspectos significativos (intento de simulación). Esto es así en el caso de la imaginación enactiva donde no hace falta coincidencia relevante entre los procesos mentales. Basta que haya propósito de coincidir, aunque esto no se logre. Sin embargo, en el caso de los procesos espejo esta definición no parece suficiente. Los procesos espejo se llevan a cabo con el propósito encubierto de duplicar en aspectos relevantes un evento mental y precisan que haya coincidencia entre los procesos o eventos mentales. Si el blanco experimenta asco, se duplica la emoción de asco en el observador. De modo que las condiciones para una SNI parecen ser más fuertes que para SNS. En este sentido, la definición (S3) de intento de simulación no puede constituir la condición mínima de simulación mental, porque no cubre el caso de los procesos espejo (SNI).

En el caso de la definición (S2) de simulación mental pasa lo contrario. Si bien alcanza para cubrir el caso de los procesos espejo (SIN), no cubre el caso de la imaginación enactiva (SNS). Ésta precisa dar cuenta de casos en los que no hay coincidencia entre los procesos que se duplican y, aún así estos deben ser considerados

(intentos de) simulaciones, en tanto el propósito fue el de duplicar. En el caso de la definición (S1) de simulación, puesto que se trata de la noción genérica de simulación abarca los casos de procesos espejo y de imaginación enactiva pero también los junta con demasiados otros eventos, a saber, simulación mental para tareas que no son de *mindreading* y simulaciones en general. A mi entender, esto no es relevante para saber qué tienen en común los procesos espejo y la imaginación enactiva como procesos subyacentes a *mindreading*. Estos procesos son tan distintos y no queda claro en qué consiste ser una simulación mental subyacente a *mindreading*.

En relación con esto, la propuesta de distintos niveles también ha sido cuestionada. Según de Vignemont 2009, se han señalado los niveles pero sus diferencias no han sido caracterizadas propiamente y, por esto, lo que aparenta ser una distinción conceptual puede no serlo. La distinción de niveles apenas se ha especificado con detalle a partir de la naturaleza de sus prototipos, los procesos espejo y la imaginación. No obstante, estos constituyen realizaciones de los niveles pero no niveles de simulación propiamente dichos (de Vignemont 2009). Al analizar los procesos espejo y la capacidad de ficción en relación a los rasgos asociados a cada nivel, tales como la automaticidad, parece que la distinción de niveles implica, más bien, una cuestión de grado, es decir, se trata de procesos más automáticos o menos conscientes, y no de una distinción conceptual[58].

Según la definición de niveles, la SNI es automática. Por definición, los procesos automáticos son inmunes a la interferencia y no pueden ser guiados por conocimiento. Sin embargo, una serie de estudios han mostrado que los procesos espejo, los prototipos de la SNI, pueden ser modulados e interferidos y, por lo tanto, no pueden considerarse procesos automáticos. Específicamente, una serie de estudios de empatía emocional sugiere que ésta no se activa automáticamente frente a un estímulo doloroso. Contrariamente, el tipo de relación que se establece con la persona que

[58] El análisis de Vignemont (2009) es más extenso y abarca otros rasgos tales como los estados mentales que implican la accesibilidad a la conciencia, la base neuronal, la confiabilidad (adecuación de la información) y la fecundidad (riqueza de la información que se adquiere), queda en evidencia que tales variables no constituyen buenos criterios para establecer una distinción conceptual.

padece dolor modula la respuesta empática hacia la misma (Singer *et al*. 2006), además la respuesta empática puede ser regulada intencionalmente (Cheng *et al*. 2007) y la empatía puede ser guiada por conocimiento semántico (Lamm *et al.* 2007)[59]. Ahora bien, si la simulación de nivel inferior está sujeta, en parte, a un control *top-down*, entonces ya no puede definirse por la automaticidad. En este sentido, no parece haber una distinción conceptual entre niveles de simulación. De modo que esta propuesta tampoco ofrece un criterio satisfactorio para diferenciar los niveles de SIN y SNS. No está claro tampoco si existe alguna diferencia entre *mindreading* de nivel inferior y de nivel superior y la SNI-SNS.

3. Los requisitos de hetero-atribución

En esta sección me ocuparé de evaluar si la propuesta de Goldman (2006) satisface los requisitos (HA1)-(HA4) de *mindreading* de estados mentales de otras personas, establecidos en el capítulo 3, sección 3.1. A mi entender, el requisito (HA1) de que los adultos normales atribuyen percepciones, conocimientos, creencias, deseos, intenciones, decisiones y razonamientos a otras personas queda trivialmente satisfecho con una propuesta para *mindreading* de tercera persona. En principio, la propuesta de *mindreading* simulacional de nivel superior da cuenta de este rango de atribuciones, mientras que la propuesta de *mindreading* simulacional de nivel inferior

[59] Un grupo de personas, varones y mujeres, participó en un juego económico donde algunos contrincantes les hacían trampa mientras que otros no. Posteriormente, los participantes fueron sometidos a estudios de neuroimagen de RMf que mostraron, en los participantes mujeres, una activación en la red asociada al dolor cuando observaban escenas donde sus contrincantes tramposos resultaban lastimados. Contrariamente, en los participantes varones, no se observó la activación en la red asociada al dolor al ser expuestos al mismo escenario (Singer *et al*. 2006). Los investigadores concluyeron que la respuesta emocional empática hacia una persona es modulada según la relación que se ha establecido con la misma. El estudio de neuroimagen con RMf (Cheng *et al*. 2007) muestra que, ante eventos dolorosos, no se observa activación en la red del dolor en los sujetos que son médicos practicantes (Cheng *et al*. 2007). En otro estudio con RMf, los participantes, que son expuestos a escenas donde hay personas que sienten dolor físico, mostraron un patrón de respuesta empática modulada según el conocimiento. Más específicamente, la activación de las regiones asociadas al dolor resultó menor en la condición donde los participantes saben que el dolor ocasionado es útil para quien lo padece porque, por ejemplo, es parte de la curación, en comparación con la condición bajo la cual se sabe que el dolor no ayuda a la persona (Lamm *et al.* 2007).

da cuenta de la atribución de otros estados mentales que no aparecen en esta lista, como las emociones y las sensaciones, tales como el tacto y el dolor.

Los requisitos (HA2) y (HA3) se relacionan con ciertas características de *mindreading* infantil ampliamente establecidas en la literatura, tal como señalé en el capítulo 3. En la propuesta de Goldman (2006), la relación entre la simulación y el desarrollo de *mindreading* no es central. No obstante, se establece cierta concordancia entre teorías que hacen hincapié en el déficit de *mindreading* en el autismo (Baron-Cohen 1999, 2003) y la simulación (Goldman 2006: 200-211). A su vez, se propone una explicación del desempeño infantil en tareas de falsa creencia basada en la simulación. Este último aspecto es relevante para el análisis de los requisitos (HA2) y (HA3) sobre el desarrollo. A continuación, reconstruiré la explicación del desempeño en tareas de falsa creencia propuesta por el enfoque simulacional. Luego, señalaré algunas dificultades de esta explicación y concluiré que la explicación simulacional del desempeño en tareas de falsa creencia da cuenta del requisito (HA2), pero no del requisito (HA3).

Según Goldman (2006), la simulación tiene una explicación simple del desempeño de los niños en las tareas de falsa creencia y del hallazgo en el desarrollo respecto del cambio que se produce entre los 3 y los 5 años. Dado un escenario de falsa creencia, el *mindreader* debe simular al blanco a partir de una creencia ficticia que contradice lo que él mismo sabe. El *mindreader* debe usar su estado mental ficticio, más que su estado mental genuino, para predecir la creencia del blanco. En una tarea de falsa creencia de cambio de lugar, por ejemplo, debe usar la creencia ficticia de que el objeto está en cierta ubicación, cuando él tiene la creencia de que la ubicación real es otra. De modo que el simulador debe poner en cuarentena o inhibir su creencia genuina para que no contamine la simulación. En este sentido, el cambio en el patrón de desempeño entre los 3 y los 5 años se relaciona con un cambio en la capacidad para poner en cuarentena o inhibir los estados mentales genuinos (Goldman 2006: 197).

Para sostener su afirmación, Goldman (2006) reseña evidencia relacionada con explicaciones del cambio en el desempeño en tareas de falsa creencia basadas en funciones ejecutivas, en particular, en el control inhibitorio[60]. El control inhibitorio es una habilidad ejecutiva que permite ignorar las tendencias dominantes o habituales. En el caso de *mindreading*, el *mindreader* debe focalizar en lo que el blanco cree, más que en lo que él mismo sabe o cree. Las tareas de falsa creencia requieren que el *mindreader* ignore su tendencia a lo que (él cree que) es real. Según el enfoque de las funciones ejecutivas, las fallas de los niños en varias tareas de falsa creencia (cambio de lugar, realidad-apariencia, contenedor engañoso y demás) se explican como una dificultad para ignorar la tendencia "a la realidad". De modo que un factor crucial en el desarrollo de *mindreading* lo constituye el desarrollo del control inhibitorio.

Carlson & Moses (2001) señalan que las capacidades de *mindreading* y el control inhibitorio tienen varios puntos en común. En primer lugar, se observan cambios en el desarrollo de estas capacidades en el mismo período etario, entre los 3 y los 5 años. Además, comparten su ubicación cerebral (los lóbulos frontales) y ambas capacidades parecen estar ausentes en el autismo. En base a esto, el estudio de Carlson & Moses (2001) indaga sobre la relación entre el control inhibitorio y *mindreading*. Con este propósito, se evalúa el desempeño de los niños en una batería de tareas de *mindreading*, que incluye dos tareas de falsa creencia, y una batería de tareas de control inhibitorio. Según los resultados, cada medida de control inhibitorio está significativamente relacionada con la batería de *mindreading* y cada medida de *mindreading* se correlaciona con la batería de control inhibitorio. Si bien este estudio sugiere una relación entre las funciones ejecutivas, o más precisamente el control inhibitorio, y *mindreading*, la naturaleza correlacional del mismo no permite conocer la dirección de la relación. En línea con esto, el estudio de Hughes (1998) sugiere que el

[60] Esta tesis está en concordancia con la propuesta de que las fallas en *mindreading* de los adultos están asociadas a déficits en la capacidad de poner en cuarentena los estados genuinos. Como mencioné en la sección 1.2, la evidencia empírica que brinda apoyo a esta tesis se relaciona con los sesgos egocéntricos en *mindreading* y con el caso del paciente WBA que, debido a su incapacidad para inhibir la perspectiva propia, muestra un desempeño deficiente en una amplia gama de tareas de *mindreading* (Samson *et al.* 2003)

desempeño en tareas de funciones ejecutivas a los 3 años predice el desempeño en *mindreading* un año después, pero no a la inversa. Además, Birch & Bloom (2003) muestran que la magnitud de la dificultad en el control inhibitorio (o el efecto de la maldición del conocimiento) disminuye con la edad de los sujetos entre los 3 y los 5 años. Esto sugiere un rol de las funciones ejecutivas en el desarrollo de *mindreading*.

Estos estudios señalan una relación entre el control inhibitorio y el desarrollo de *mindreading*. A diferencia de lo que ocurre entre *mindreading* y el lenguaje, donde el lenguaje parece estar implicado sólo en el desarrollo de *mindreading*, la evidencia sugiere un rol para el control inhibitorio en la capacidad madura de *mindreading*. En particular, el estudio ya mencionado de la paciente WBA que, debido a su incapacidad para inhibir la perspectiva propia, muestra un desempeño deficiente en una amplia gama de tareas de *mindreading*. Esta evidencia sugiere una relación entre el control inhibitorio y *mindreading* en la adultez.

Según Goldman (2006), este conjunto de evidencia señala que el déficit en la capacidad para inhibir la perspectiva propia conduce a errores masivos en *mindreading* con la forma de sesgos egocéntricos. Si bien Carlson & Moses (2001) no sugieren que sus hallazgos puedan ser explicados por la TS, Goldman (2006) sostiene que estos son consistentes con la propuesta simulacional. Además, Samson *et al.* (2005) señalan la correspondencia entre las áreas de lesión de WBA y las señaladas por Vogeley *et al*. (2001) como las áreas que vinculan la auto-perspectiva con *mindreading* de tercera persona, y esto brinda apoyo a la TS. Según Goldman (2006), la TS no sólo es consistente con todos estos hallazgos sobre la relación entre *mindreading* y el control inhibitorio, sino que se ve favorecida (apoyada empíricamente) por los mismos. Es más, la TS provee la mejor explicación para esta clase de errores egocéntricos en *mindreading* identificados en la psicología del desarrollo, la psicología social y la neuropsicología.

En mi opinión, el problema de esta explicación es que la cuestión del déficit en el control inhibitorio no da cuenta de todos los aspectos del cambio en el desempeño infantil en tareas de falsa creencia. Al reflexionar sobre las tareas de falsa creencia y

otras tareas de *mindreading* se advierte que las funciones ejecutivas pueden estar involucradas de muchas maneras. A saber, para mantener en mente la secuencia de eventos, para inferir y mantener en mente la creencia falsa de *Sally* así como nuestro conocimiento sobre la ubicación del objeto, para resistir la interferencia entre estos registros, para formular una respuesta en base a la creencia de *Sally* y no en base al propio conocimiento (Apperly 2011). Muchos estudios han encontrado correlaciones entre funciones ejecutivas y *mindreading*. Ahora bien, si bien algunos estudios han señalado que ciertas funciones ejecutivas, por ejemplo las tareas de inhibición de conflicto, se correlacionan más estrechamente con las tareas de falsa creencia, esto no indica qué aspecto de la tarea de falsa creencia demanda la inhibición del conflicto. En la literatura, hay muchas interpretaciones del vínculo entre *mindreading* y funciones ejecutivas. Según algunos, en concordancia con Goldman (2006), razonar con falsas creencias requiere que los niños resistan la tendencia a responder en base a su propio conocimiento (Leslie 1994; Carlson & Moses 2001; Birch & Bloom 2007). Según otros, las funciones ejecutivas son requeridas por las demandas incidentales de las tareas de falsa creencia como atender y recordar la secuencia de eventos, interpretar las preguntas (Bloom & German 2000). En suma, no es claro que el control inhibitorio sea la historia completa sobre el desempeño infantil en tareas de falsa creencia. Esto ha sido mostrado críticamente por Call & Tomasello (1999). Los investigadores diseñaron una tarea de falsa creencia no verbal que reduce al máximo la posibilidad de interferencia de la perspectiva propia del niño y, sin embargo, los niños aún se desempeñan con dificultad[61]. Además, hay evidencia que sugiere cierta capacidad de

[61] En la tarea falsa creencia no verbal (Call & Tomasello 1999), la tarea del participante es identificar que caja contiene el objeto oculto. En la fase de entrenamiento, los niños aprenden que hay un premio en una de dos cajas, y que Sally les ayudará a encontrarlo poniendo una marca como señal. En la tarea, *Sally* mira en las dos cajas y sabe dónde está el objeto. En el ensayo crítico para la falsa creencia, Sally sale de la escena, y en su ausencia *Andrew* cambia las cajas de lugar, de modo que ahora *Sally* tiene una creencia falsa acerca de la ubicación de premio. *Sally* vuelve a escena y pone una marca en una de las cajas para indicar la ubicación del premio. Si el niño tiene en cuenta la creencia falsa de *Sally* puede encontrar el premio. El aspecto crítico de la tarea es que cuando los niños tienen que tener en cuenta la creencia falsa de *Sally*, ellos mismos no saben donde está ubicado el objeto, de modo que no necesitan resistir a la interferencia de esta información (o al menos la posibilidad queda reducida). Sin embargo, no mejora el desempeño de los niños que sigue siendo como en las tareas de falsa creencia estándar.

mindreading en los bebés (Onishi & Baillargeon 2005) y en los primates no humanos (Call & Tomasello 2008), que no pueden estar relacionadas con las funciones ejecutivas porque esta población carece de las mismas.

A su vez, la relación entre *mindreading* y el control inhibitorio no parece ser toda la historia sobre los sesgos egocéntricos en los adultos. Como mencioné, se ha mostrado que el desempeño en los adultos se relaciona con el desempeño en pruebas independientes de funciones ejecutivas, y que la capacidad de *mindreading*, en los adultos, puede verse disminuida si hay demanda de funciones ejecutivas para una segunda tarea o si están deterioradas por daño neurológico. No obstante, al parecer otras funciones ejecutivas, distintas al control inhibitorio, están involucradas en *mindreading*. Algunos estudios indagan sobre el rol de las funciones ejecutivas en *mindreading* en adultos neurológicamente intactos. Un estudio basado en un juego de comunicación donde los participantes tienen que usar información sobre la perspectiva visual del hablante muestra que estos comenten errores egocéntricos y que los mismos se correlacionan con tareas de funciones ejecutivas que requieren que los participantes retengan mentalmente alternativas y elijan entre las mismas, pero no se correlacionan con una variedad de pruebas de inhibición y de memoria de trabajo (Apperly, Carroll *et al.* 2010).

De modo que la explicación basada en el control inhibitorio no es toda la historia sobre los errores egocéntricos, ni del desempeño infantil en tareas de falsa creencia entendido como el producto de errores egocéntricos. En este sentido, en la medida en que la propuesta brinda una explicación respecto del salto significativo en la habilidades de *mindreading* alrededor de los tres años como un el desarrollo de las funciones ejecutivas, el desarrollo del control inhibitorio permite a los niños desempeñarse correctamente en la tarea de falsa creencias, el requisito (HA2) queda satisfecho. No obstante, como he mencionado, la explicación del control inhibitorio no es del todo satisfactoria. En relación al requisito (HA3) de que la comprensión de los deseos precede a la comprensión de las creencias, ningún elemento en esta explicación

permite dar cuenta de este aspecto del desarrollo. En este sentido considero que el requisito (HA3) no queda satisfecho y en este sentido la propuesta es insuficiente.

Finalmente, considero que esta propuesta satisface el requisito (HA4) de que las personas llevan a cabo juicios de *mindreading* proyectando sus propias sensaciones, creencias y conocimiento. Al hacer énfasis en el aspecto de la proyección y el requisito concomitante de cuarentena de los estados mentales genuinos propios en la rutina de simulación, la propuesta simulacional da cuenta de cómo las rutinas de simulación pueden ser propensas a la proyección de las propias creencias, sensaciones y conocimiento en las atribuciones de tercera persona. No obstante quisiera señalar que, en mi opinión, esta explicación no permite discriminar entre un error ocasional producto de la filtración de un estado genuino y las atribuciones sistemáticamente incorrectas o los sesgos egocéntricos. Según la TS, la cuarentena de los estados mentales propios en una rutina de simulación es necesaria para que *mindreading* sea correcto. De modo que si hay violación de cuarentena esto conducirá a una atribución errónea porque se usan los estados genuinos propios que no coinciden con el blanco. En este sentido, se trata de un error egocéntrico. Esto sólo es una explicación de la posibilidad de que un error sea de tipo egocéntrico, pero no explica la sistematicidad de los sesgos egocéntricos. Es más, Goldman (2006: 41) sugiere que hay sesgos egocéntricos si la filtración es desmesurada. La cuestión consiste en definir qué hace que la filtración sea desmesurada. Como he mencionado, por ejemplo, Birch & Bloom (2003) sugieren que los adultos cometen errores sistemáticos en tareas de *mindreading* cuando tienen que contemplar probabilidades. Nickerson (1999; Royzman, Cassidy & Baron 2003) sostiene que cuando los adultos llevan a cabo *mindreading* en condiciones de incertidumbre, a menudo, utilizan el atajo de asumir que el "otro es como yo" y, a partir de allí, realizan ajustes. Con esta salvedad, considero que el requisito (HA4) queda satisfecho.

4. *Mindreading* de primera persona

Todo enfoque de *mindreading* debe poder dar cuenta no sólo de la atribución de estados mentales a otras personas sino también de la autoatribución[62]. Es preciso notar que todo enfoque puro de simulación siempre va a constituir una explicación parcial de *mindreading* en tanto éste no puede dar cuenta de la autoadscripción de estados mentales ocurrentes. No obstante, la simulación intrapersonal puede usarse para la autoadscripción de estados mentales pasados, futuros e hipotéticos (Goldman 2006: 223). En este sentido, el enfoque simulacional deberá ser complementado con algún enfoque sobre la autoadscripción de estados mentales ocurrentes para dar cuenta de *mindreading*. Además, en el caso particular de la propuesta de Goldman (2006), se asume que la autoatribución tiene un rol central en *mindreading* de tercera persona. A saber, para atribuir estados mentales ajenos es necesario clasificar previamente los estados mentales propios.

Según Goldman (2006), la autoatribución no sólo consiste en la detección de los estados mentales sino también en la clasificación de los mismos. En este sentido, es necesario que el modelo de autoatribución pueda dar cuenta de la clasificación de estados mentales según su tipo. El enfoque introspectivo propuesto por Nichols & Stich (2003), según Goldman (2006), no da cuenta de esto. En esta propuesta introspeccionista, la autoatribución se lleva a cabo mediante un mecanismo de monitoreo (MM) que toma una representación de una "caja", le agrega un prefijo, por ejemplo "yo deseo", y vuelve a colocar la representación, esta vez, en la "caja de creencias". El problema reside, según Goldman (2006), en que las cajas sobre las que opera el mecanismo de monitoreo no son "cajas cerebrales" genuinas sino denominaciones para roles funcionales. De modo que si hay un único MM, este enfoque no da cuenta de cómo éste determina el rol funcional del estado mental. Una

[62] En filosofía, la cuestión de la autoatribución remite a la cuestión del auto-conocimiento. Sin embargo, como ya he mencionado en el capítulo 2, en el ámbito de *mindreading*, "conocimiento" no se entiende como un concepto epistemológico en el sentido de una "creencia verdadera justificada" o a una "creencia verdadera formada confiablemente". En el caso del auto-conocimiento, se trata de la cuestión más simple de cómo las personas captan y clasifican los estados mentales propios, con independencia de si son confiables, justificables y demás.

posible solución consiste en postular múltiples MM, uno para cada actitud. Sin embargo, esto parece poco parsimonioso (Goldman 2006: 239)[63].

La propuesta de Goldman (2006) para el reconocimiento del tipo de estado mental se basa en un modelo cuasi perceptual de la introspección. El proceso de identificación de los estados mentales propios está modelado sobre los procesos perceptuales de ver y escuchar (Goldman 2006: 225) y se denomina "introspección", "sentido interno" o "monitoreo" [64]. Según este modelo, la introspección implica dos operaciones. Primero, la elección de un estado mental para analizar, que puede

[63] Esta crítica a Nichols & Stich (2003) está basada en la adhesión de Goldman (2006) a un enfoque unitario sobre la introspección. Sin embargo, en principio, no hay ninguna razón para sostener que la introspección necesariamente tenga que consistir en un único proceso o mecanismo. En este sentido, que no se trate de un proceso unitario no implica la falta de parsimonia, si es que se explica mejor el fenómeno. Si bien Nichols & Stich (2003) no terminan de describir su propuesta para otro tipo de estados que no sean las creencias y los deseos, en principio, consideran la posibilidad de múltiples mecanismos de monitoreo. A su vez, si bien Goldman (2006) adhiriere a un enfoque unitario respecto de la introspección, sin embargo, la propuesta de las propiedades neuronales como realizadoras de los tipos de estados mentales no parece conducir a un enfoque unitario: en la medida en que los realizadores neuronales de las diferentes actitudes y de los tipos perceptivos estén localizados en diferentes regiones cerebrales, cada actitud o tipo perceptivo dependerá de un canal distinto para conducir información a la introspección. De modo que Goldman también tendría que predecir disociaciones múltiples en la capacidad de introspección (Engelbert & Carruthers 2010).

[64] Una de las críticas que recibe la propuesta cuasi perceptual de la introspección es que la analogía con la percepción no puede establecerse, entre otras razones, porque la introspección carece de un órgano receptor, que pueda poner al sujeto en una relación adecuada con el blanco cognitivo (Shoemaker 1996). En contra de esto, Goldman (2006: 244) sostiene que la atención puede considerarse el órgano de orientación de la introspección. Para sostener su propuesta y reforzar el modelo perceptual de la introspección, Goldman (2006) recurre a cierta evidencia. Por un lado, evidencia que sugiere que la introspección no ocurre sin la guía de la atención. La atención se requiere o, al menos, facilita la introspección. Esta puede ocurrir de manera automática, voluntaria o mediante una intervención externa (un sonido). En los estudios de Hulburt y sus colegas, se desarrolla un método para estudiar la introspección. Éste consiste en que los sujetos presten atención inmediata lo que están experimentando cuando suena una alarma. Ésta se activa de manera aleatoria y los sujetos tienen que anotar lo que les ocurre. Posteriormente, esto será descripto con profundidad en una entrevista que tiene lugar 24 horas después (Hulburt & Heavey 2001). Al parecer, mediante este método los sujetos descubren pensamientos o imágenes mentales sobre las que no tenían conocimiento, aunque pueden acceder a las mismas de manera consciente (con ocasión someterse a este método de "muestreo descriptivo de experiencia"). Por otro lado, Goldman reseña evidencia que sugiere que la atención tiene un rol en la percepción. Se trata del fenómeno de "ceguera al cambio" o "ceguera por desatención". En estos estudios, los participantes observan escenas naturales y se les pide que detecten cambios, por ejemplo, el cambio de color de un objeto o su aparición o desaparición. El hallazgo consiste en que si estos cambios se hacen coincidir con los movimientos sacádicos del ojo, con los pestañeos o con un corte en la secuencia de una película, los participantes no advierten los cambios aunque estos sean notorios y estén a la vista. Incluso aunque las medidas de la visión señalen que ésta está dirigida directamente al cambio (O'Reagan 1992; O'Reagan, Rensik & Clark 1999). De modo que si la percepción y la introspección parecen requerir de la atención, según Goldman, este modelo cuasi perceptual encuentra cierta plausibilidad.

entenderse como la atención que se dirige hacia adentro. Segundo, el análisis y clasificación del estado mental que, según este modelo cuasi perceptual, debe incluir un proceso de transducción (Goldman 2006: 246). Se considera que todo proceso de transducción tiene un *input*, ciertos eventos o propiedades a la cuales el proceso es sensible, y un *output*, o unas representaciones generadas en respuesta a tales *inputs*. Según Goldman (2006), los *outputs* de la introspección son representaciones de instancias de estados mentales que se van a clasificar según una o más dimensiones. A saber, según el tipo de estado mental, el contenido y la intensidad del mismo. Ahora bien, la cuestión de las propiedades que pueden servir de *inputs* es más delicada. En principio, las candidatas son varias pero las propiedades neuronales, según Goldman (2006), son las más plausibles por las siguientes razones[65]. Éstas presentan la ventaja de ser propiedades físicas y, en este sentido, no se puede cuestionar su eficacia causal. Además, pueden ser detectables con independencia de ser conscientes o no. Para sostener esta propuesta, se recurre al ejemplo de la discriminación introspectiva de las sensaciones corporales.

[65] Las propiedades funcionales son descartadas con argumentos similares a los mencionados en el capítulo 2. En principio, las propiedades funcionales no son candidatas a ser las propiedades a las que puede ser sensible un sistema introspectivo porque se trata de propiedades disposicionales y relacionales. Nótese que no se niega que los estados mentales tengan roles funcionales sino que este rasgo no puede ser captado por un modelo cuasi perceptual de la introspección. En pocas palabras, ningún sistema perceptual puede atender a este tipo de propiedades. Así como la visión no puede captar la solubilidad del azúcar, aunque este rasgo se puede inferir, el modelo cuasi perceptual de la introspección sólo puede dar cuenta de estados mentales ocurrentes. El carácter relacional tampoco es apropiado. Por ejemplo, el rol funcional asociado al miedo incluye la tendencia a aliviar el dolor. El problema es que el sistema introspectivo no puede dar cuenta de si el estado mental ocurrente (dolor) implica también esta tendencia. Para resto será preciso determinar si el estado mental viene acompañado de esa tendencia o no, puesto que se capta directamente es el estado mental ocurrente y no todas sus relaciones. Asimismo, las propiedades fenoménicas tampoco son buenas candidatas. La introspección requiere que las propiedades a identificar puedan ser discriminadas para las instancias ocurrentes de distintos tipos de actitudes proposicionales (*i.e.* deseos, creencias, dudas y demás). En general se asume que las actitudes proposicionales no poseen cualidades fenoménicas, pero aún si se concediera que sus instancias tienen algún rasgo cualitativo, esto no garantiza la posibilidad de clasificarlas según su fenomenología. Para esto, será necesario que cada tipo de actitud proposicional tenga una propiedad fenoménica característica que pueda servir para su calcificación. En este argumento, Goldman (2006) toma distancia se su postura anterior, desarrollada en el capítulo 2, en la que afirmaba que las instancias de actitudes proposicionales ocurrentes poseen rasgos fenoménicos (Goldman 1993a).

Basado en un sistema neurofisiológico que se conoce con detalle (la lamina I del sistema espino-tálamo-cortical), Craig (2002) propone un enfoque de la interocepción, esto es, la sensación de la condición fisiológica del cuerpo[66]. En este sistema convergen señales de aferentes que representan el estado fisiológico de todos los tejidos del cuerpo de modo que en el núcleo ventral posterior de tálamo, están alojadas las representaciones corticales de las distintas sensaciones como el dolor, la temperatura, el picor, las sensaciones viscerales y musculares, y demás sensaciones corporales. En este sentido, según Craig (2002), esta región cortical puede considerarse "la corteza interoceptiva". En este sistema aferente, dos clases diferentes de neuronas conducen las señales de dolor agudo (por presión) y de dolor por quemaduras, que reciben *inputs* selectivos de los receptores del dolor. Además, hay dos tipos de células termo receptoras que responden selectivamente al frío y al calor, así como otras clases de neuronas que responden selectivamente a los músculos, las articulaciones y las cosquillas.

En base a esta descripción, el enfoque de las propiedades neuronales sostiene que, la clasificación de los tipos de sensaciones se basa en la percepción de la clase de neuronas que está activada en el sistema del núcleo ventral posterior del tálamo. De modo que si están activadas las neuronas termo receptoras de frío, la sensación se clasifica como el estado de frío. Ahora bien, la interocepción no debe confundirse con la introspección. La interocepción consiste en la percepción del estado fisiológico de los tejidos del cuerpo que culmina en sensaciones (a menudo descriptas por los neurólogos en términos de representaciones del estado del tejido en las diferentes regiones corporales). Goldman (2006) considera que estas representaciones sensoriales son el contenido de las sensaciones y que la introspección es una respuesta a estas sensaciones. La introspección consiste en una meta-representación y clasificación de las sensaciones corporales en términos de estados mentales (Goldman

[66] Se trata del sistema aferente en el que los axones somatosensoriales procedentes de la piel, los músculos y los órganos internos entran al sistema nervioso central vía los nervios espinales, pasan por el núcleo ventral posterior de tálamo (un núcleo de relevo de la información sensorial) y, desde el tálamo, los axones proyectan a la corteza somato sensorial, la corteza insular anterior.

2006: 252). Según la propuesta de las propiedades neuronales, la clasificación del estado mental según su tipo se ejecuta en base al grupo de células neuronales que esté activado. Además, se puede determinar su intensidad, en relación a la intensidad de la activación[67].

A su vez, la introspección debe involucrar algún aspecto que permita determinar el contenido del estado mental. Según Goldman (2006), esto no puede llevarse a cabo por el mismo proceso que clasifica los tipos, porque mientras las categorías de estados mentales son unas pocas, los contenidos son innumerables (Goldman 2006: 253). El modo más plausible de "introspectar" un contenido es "reutilizarlo". De modo que el contenido del deseo es replicado en la meta-representación introspectiva del deseo. A su vez, la "reutilización" no sirve para representar (clasificar) el tipo de estado mental, porque la meta-representación no es en si misma un deseo sino una creencia o un juicio (Goldman 2006: 254). Ahora bien, si este contenido está codificado en distinto formato que el de la autoatribución, la reutilización implicará, además, una traducción. Por ejemplo, el sistema de introspección o monitoreo no puede simplemente tomar una representación visual y agregarle el prefijo "yo creo que" de modo que resulte una representación para colocar en la caja de creencias. Para esto, es necesario que el formato visual sea "traducido" a un formato apropiado aunque, cuando el sistema de introspección opera sobre estados sensoriales o perceptuales, a menudo la traducción será parcial. De modo que la clasificación introspectiva implica algún tipo de traducción de representaciones sensoriales, y ésta es limitada y aproximada (Goldman 2006: 254).

Además de sostener que existe un método especial para la autoatribución, Goldman (2006) concede que, a veces, ésta resulta en confabulación. Sin embargo, se considera que la evidencia respecto de la confabulación (la misma que mencioné en el

[67] Goldman (2006: 253) advierte que si bien la propuesta de las propiedades neurológicas es prometedora para clasificar los tipos de sensaciones en tanto hay circuitos neuronales diferentes dedicados a cada tipo de sensación (calor, dolor, cosquillas y demás), es poco probable que este tipo de circuitos esté disponible para las actitudes proposicionales. Sin embargo, considera que afirmar que las propiedades neuronales son utilizadas por la introspección no implica necesariamente la existencia de circuitos dedicados. De modo que el enfoque de las propiedades neuronales sigue siendo prometedor.

capítulo 3) no permite afirmar que se trate del único método, ni siquiera del método principal. Es más, se considera que los casos de confabulación son raros, mientras que los casos estándar de autoatribución involucran la introspección o el auto-monitoreo (Goldman 2006: 232). No obstante, según esto, la propuesta de Goldman consiste en un método doble de introspección o monitoreo, que sólo puede usarse para la autoatribución y, un segundo método, la interpretación, que puede usarse en las atribuciones de tercera y de primera persona. Sin embargo, el aspecto interpretativo no se desarrolla demasiado, sólo se indica que subyace a la confabulación (un fenómeno que no se puede dejar de reconocer).

Recapitulando, este modelo propone un proceso doble de introspección y de interpretación para la autoatribución o *mindreading* de primera persona. Este último aspecto no se desarrolla, mientras que el aspecto introspectivo consta de dos procesos: reconocimiento y reutilización. El reconocimiento permite determinar el tipo de estado mental, y si tiene contenido o no. Además, permite clasificar el estado según su intensidad. La reutilización, en cambio, se utiliza para asignar contenido a la meta-representación introspectiva.

5. Los requisitos para la autoatribución

En relación al requisito (AA1) para la autoatribución, como ya he mencionado en los capítulos 3 y 4, la propuesta de la introspección como un modo de acceso directo a los estados mentales propios satisface trivialmente este requisito, en tanto puede dar cuenta de la intuición de sentido común del acceso directo e inmediato a los mismos. A su vez, puede dar cuenta de otros hechos obvios como la posibilidad de auto-adscribirse estados mentales sin exhibir comportamiento manifiesto, mientras que un enfoque interpretativo tiene dificultades para explicar ambas cuestiones. Sin embargo, quisiera señalar que más allá de la propuesta de un modelo cuasi perceptual de la introspección, y de las ventajas o desventajas de las propiedades neuronales como *input* de la misma, la existencia de la introspección se da por sentada y se

argumenta a favor de un modelo particular de la misma, incluso reseñando evidencia empírica que parece sostenerlo (ver nota 64). Sin embargo, a mi entender, hay una cuestión que es anterior a la discusión sobre uno u otro modelo para la introspección, respecto de la cual esta propuesta no brinda argumentos. La cuestión es si existe o no un proceso o mecanismo cognitivo de acceso directo a los estados mentales propios, y qué tipo de evidencia permitiría sostener esto. En mi opinión, esta cuestión es relevante porque la existencia de procesos de tipo interpretativo para la autoatribución está suficientemente documentada por la evidencia, en especial, en relación con la confabulación. Esta cuestión es reconocida por Goldman (2006: 224) sólo que la respuesta que brinda consiste en ofrecer argumentos, y reseñar evidencia, para defender un modelo particular de la introspección y dar por sentada la misma.

La propuesta de Goldman (2006) propone un método doble de introspección e interpretación para la autoatribución. Sin embargo, el aspecto interpretativo de la propuesta no está desarrollado. Apenas se afirma que la evidencia sobre la confabulación sugiere que la interpretación se usa en ocasiones inusuales, pero que no se trata del proceso estándar de autoatribución. No obstante, como todo teórico de la introspección, Goldman (2006: 233) nunca afirmaría que las causas del comportamiento pueden conocerse de manera directa (Goldman 2006: 233). En este sentido, y en concordancia con Nichols & Stich (2003), puede especularse que Goldman estaría dispuesto a sostener que las personas confabulan cuando tienen que dar cuenta de su comportamiento, pero que en los casos de autoatribuciones de estados concurrentes interviene la introspección. Si esto es así, la propuesta de Goldman (2006) no da cuenta del requisito (AA2) de que a menudo las personas interpretan falsamente sus pensamientos ocurrentes y del pasado reciente, tal como sugiere la evidencia del fenómeno de la confabulación.

Dado que Goldman (2006) no desarrolla el aspecto interpretativo de su propuesta, no proporciona un criterio satisfactorio para establecer cuando subyace la introspección o la interpretación a la autoatribución. No obstante, en una respuesta a Carruthers (2009), Goldman (2009) propone el siguiente criterio. La introspección se

emplea para pensamientos conscientes mientras que la interpretación se aplica a pensamientos inconscientes. De modo que todas las instancias de confabulación se explican como ocurriendo en circunstancias donde los pensamientos relevantes no son conscientes. Carruthers (2013) señala que la respuesta de Goldman resulta problemática a la luz de los dos grandes tipos de enfoques sobre el pensamiento consciente. Un enfoque afirma que los pensamientos conscientes son pensamientos que sabemos que poseemos, en general, o mediante un acceso directo no interpretativo. El otro enfoque, sostiene que los pensamientos conscientes son los que se transmiten globalmente a un rango amplio de sistemas ejecutivos, afectivos e inferenciales.

Según el primer enfoque, los pensamientos conscientes son los que sabemos que poseemos ya sea por introspección o por interpretación. De modo que no tiene sentido apelar a la distinción consciente/inconsciente para explicar las instancias de confabulación. Los pensamientos conscientes también parecen estar involucrados en la interpretación, entonces se puede esperar que la confabulación también ocurra en estos casos. La otra opción consiste en sostener que hay acceso directo a los pensamientos que se transmiten globalmente, mientras que para el resto se precisa interpretación. El problema con esta opción es que hay poca evidencia de que los pensamientos (juicios y decisiones) se transmitan de este modo, tal como se transmiten las sensaciones o los estados que involucran sensaciones (Carruthers 2013:472). De modo que el criterio consciente/inconsciente tampoco parece un buen criterio para determinar los casos de autoatribución a los que se aplica introspección o interpretación. En este sentido, la propuesta de doble proceso de Goldman (2006) no satisface el requisito de proponer un criterio que permita determinar el proceso subyacente a un caso de autoatribución. Esta propuesta apenas permite afirmar que en algunos casos de atribución se apela a la introspección y en otros a la interpretación.

6. Conclusión

En este capítulo presenté y analicé la propuesta de Goldman (2006) para *mindreading* de tercera y primera persona. Esta propuesta es satisfactoria en tanto da cuenta de ambos tipos de atribuciones, sin embargo, no propone un criterio que permita determinar el tipo de proceso subyacente, ya sea en *mindreading* de tercera persona o de primera persona. Esto es relevante ya que ambas propuestas postulan más de un proceso subyacente a *mindreading*.

En el caso de *mindreading* de tercera persona, se propone un enfoque híbrido de teoría y simulación. No obstante, como señalé en la sección 2, la propuesta no brinda un criterio para determinar el tipo de proceso que subyace a un caso de *mindreading*. Es más, el elemento de teoría no se desarrolla, y apenas se describen las posibles relaciones entre teoría y simulación. El énfasis en la simulación se aprecia en la propuesta de SNI y SNS para *mindreading*. Goldman se ha encargado de reseñar evidencia para fundamentar ambas propuestas. En la sección 1.1, me ocupé de la SNI, cuyo prototipo son los procesos espejo. Goldman afirma, basado en la interpretación de cierta evidencia, que los procesos espejo pueden considerarse simulaciones y que subyacen a *mindreading* de nivel inferior. A mi entender, los procesos espejo pueden dar cuenta del aspecto simulacional de las rutinas de simulación pero no del aspecto de proyección. De modo que para que un proceso espejo pueda subyacer a *mindreading* tendría que conectarse con algún proceso atributivo. Esto no es algo que concuerde con la propuesta de Goldman. A mi entender, la propuesta de SNI es suficiente para sostener que los procesos espejo pueden entenderse como simulaciones pero no alcanza para sostener que estos subyacen a *mindreading*. De manera similar considero que la evidencia mencionada para sostener la propuesta de SNS tiene limitaciones. En particular, el estudio de Mitchell *et al.* (2004) no brinda apoyo directo a la propuesta de Goldman (la necesidad del monitoreo de los estados mentales propios en la atribución de tercera persona), sino a alguna versión general de simulación (sección 1.2). Además, la distinción de niveles ha sido cuestionada (de Vignemont 2009), de modo que no es tan claro en qué consiste la SNI y la SNS (sección 2). En particular,

considero que la redefinición de la noción de simulación es confusa y no es claro qué tienen en común los procesos espejo y la imaginación enactiva como procesos simulacionales subyacentes a *mindreading* (sección 2). En conjunto, estos argumentos sugieren que la propuesta de *mindreading* simulacional de tercera persona no está tan fundamentada por la evidencia como puede parecer en un principio.

En la sección 3 analicé la propuesta según los requisitos de hetero-atribución definidos en el capítulo 3. El requisito (HA1) queda trivialmente satisfecho con la propuesta del enfoque simulacional de *mindreading* de tercera persona. En principio, la SNS da cuenta del rango de atribuciones del requisito, mientras que la propuesta de SNI da cuenta de la atribución de otros estados mentales que no aparecen en esta lista, tales como las emociones y las sensaciones. El requisito (HA2) queda satisfecho en la medida en que la propuesta brinda una explicación sobre el salto significativo en las habilidades de *mindreading* alrededor de los tres años a partir del desarrollo de las funciones ejecutivas (el desarrollo del control inhibitorio permite a los niños desempeñarse correctamente en la tarea de falsa creencias). No obstante, con cierta salvedad. Como he mencionado, la historia del control inhibitorio no es toda la historia sobre los errores egocéntricos, ni lo es el desempeño infantil en tareas de falsa creencia entendido como el producto de errores egocéntricos. En relación al requisito (HA3) de que la comprensión de los deseos precede a la comprensión de las creencias, ningún elemento en esta propuesta permite dar cuenta de este aspecto del desarrollo. En este sentido considero que el requisito (HA3) no queda satisfecho. El requisito (HA4) queda satisfecho con la postulación de la proyección como el último paso en la rutina simulación y la necesidad de poner en cuarentena los estados mentales propios para que *mindreading* sea correcto. Si esto no ocurre, los estados genuinos del *mindreader* se filtran en la rutina de simulación dando lugar a errores egocéntricos. No obstante, considero que esta explicación da cuenta de la posibilidad de que un error en *mindreading* sea egocéntrico pero no del carácter sistemático de los sesgos egocéntricos.

En la sección 4 presenté la propuesta de proceso doble de introspección e interpretación para la autoatribución. Esta propuesta hace hincapié en el aspecto introspectivo y de esta manera, como toda propuesta introspectiva, satisface el requisito (AA1) para la autoatribución (sección 5). En particular, propone un modelo cuasi perceptual de la introspección. Ahora bien, más allá de las ventajas y desventajas respecto de esta propuesta y respecto de las propiedades neuronales como *input* de la introspección, la existencia de la introspección se da por sentada y considero que esto no es una virtud de este enfoque. A mi entender, es una cuestión relevante si existe o no un proceso o mecanismo cognitivo de acceso directo a los estados mentales propios, y qué tipo de evidencia permitiría sostener esto, porque la existencia de procesos de tipo interpretativo para la autoatribución está suficientemente documentada por la evidencia, en especial, en relación con la confabulación (sección 5). El aspecto interpretativo de la propuesta no está desarrollado. Se reconoce el fenómeno de la confabulación pero se sugiere que la interpretación se usa en ocasiones inusuales y que no se trata del proceso estándar de autoatribución, sin mucho más argumento. En este sentido el requisito (AA2) no se satisface. Además, no se ofrece un criterio para establecer cuándo se recurre a la interpretación y cuando a la introspección. En este sentido, la propuesta de doble proceso de interpretación e introspección para la autoatribución de Goldman (2006) no satisface el requisito de proponer un criterio que permita determinar el proceso subyacente a un caso de autoatribución.

CAPÍTULO 6. PROPUESTA PARA UN ENFOQUE HÍBRIDO DE *MINDREADING*

1. Introducción

Como ya he mencionado, en mi opinión, es esencial que los enfoques híbridos de teoría y simulación de *mindreading* puedan brindar un criterio que permita distinguir entre los procesos subyacentes a un caso de *mindreading*, de modo que se pueda determinar si interviene un proceso de teoría o de simulación (el requisito definido en el capítulo 3, sección 3.3). Si bien es cierto que las propuestas evaluadas en la tesis (Nichols & Stich 2003; Goldman 2006) dan cuenta de *mindreading* de primera persona (proponen enfoques de introspección e interpretación), en este capítulo me centraré en *mindreading* de tercera persona. Es preciso mencionar que la cuestión de la introspección es un tema importante aunque no está tan desarrollado en la literatura de neurociencia cognitiva social. En general, en los estudios empíricos la capacidad de *mindreading* se asocia a *mindreading* de tercera persona, y esto conlleva el estudio de este fenómeno por separado de su contraparte, la autoatribución. Si bien hay estudios sobre las bases cerebrales de la autoreflexión, como señalaré en la sección 3, no es claro que en estos estudios se esté abordando el fenómeno de autoatribución que interesa a los enfoques de *mindreading*. *Mindreading* se asocia a la autoadscripción de estados mentales tales como creencias y deseos, más que a la autoatribución de, por ejemplo, rasgos de personalidad como implican las tareas que se utilizan usualmente en los estudios de neuroimagen.

En la sección 2, me ocuparé del estudio de *mindreading* con técnicas de neuroimagen, el método más usual para su estudio en la neurociencia cognitiva social, y de algunas discusiones en torno al hallazgo de la "red neuronal de *mindreading*". En la sección 3, presentaré algunos estudios de neuroimagen que han puesto a prueba las hipótesis de TT y TS. Señalaré algunas desventajas metodológicas de los mismos y, en relación a estos, mostraré la importancia de poseer un criterio para distinguir entre teoría y simulación. Particularmente, considero que esto es crítico para el estudio de *mindreading* con métodos de neuroimagen.

Los enfoques híbridos de teoría y simulación evaluados en los capítulos 4 y 5 (Nichols & Stich 2003; Goldman 2006), no proponen un criterio satisfactorio para determinar el tipo de proceso que subyace a un caso de *mindreading*. En la sección 4, propondré que, un criterio proveniente de la neurociencia cognitiva social (Keysers & Gazzola 2006, 2007), permite defender los enfoques híbridos de teoría y simulación para *mindreading* de tercera persona. No obstante, éste implica algunas concesiones en relación al alcance explicativo de la TS. Específicamente, en relación al tipo estados mentales a los que subyace la simulación. Asimismo, la evidencia proveniente de las neurociencias sugiere un criterio que implica una división de tareas entre teoría y simulación, y esta división configura un sistema de doble proceso subyacente a *mindreading* que, en mi opinión, no puede equipararse con las propuestas de doble proceso usuales en ciencias cognitivas como, por ejemplo, las propuestas de doble proceso para el razonamiento (Evans & Stanocvich 2013) o los sistemas multiprocesos (Machery 2009). Más bien, se trata de algún otro tipo de proceso doble (sección 5).

2. Los estudios de las bases neuronales de *mindreading*

El desarrollo de las técnicas de neuroimagen, que permiten el acceso al cerebro intacto en funcionamiento, condujo a una explosión de la investigación en neurociencia cognitiva. En particular, el campo de la neurociencia cognitiva social asume que ciertas partes del cerebro podrían estar dedicadas a la percepción y la cognición sociales. En este campo se inserta la investigación sobre *mindreading*. Desde la psicología cognitiva, se asume que entender la base neuronal, y sus funciones, puede proveer evidencia para entender la capacidad cognitiva de *mindreading*. En neurociencia cognitiva social, la mayoría de los estudios se dedica a buscar la base neuronal de *mindreading* de tercera persona. En esta sección, brindaré un panorama breve de cómo se estudia *mindreading* y presentaré el hallazgo más notorio al respecto, la "red neuronal de *mindreading*", así como también algunas discusiones en torno al mismo. Básicamente, se discute cuán especializada está esta red.

El panorama que surge de los estudios de la base cerebral de *mindreading* es consistente. Esta capacidad recluta una red compleja de regiones y procesos neuronales, entre estos se incluyen regiones neuronales que parecen estar dedicadas selectivamente a *mindreading*. No obstante, como señalé en la introducción de la tesis, algunos consideran que existe una región cerebral especializada en la representación del pensamiento de otras personas, principalmente, de las creencias y los deseos (Saxe & Kanwisher 2003). En esta sección, el panorama de las bases neuronales de *mindreading* estará orientado a presentar los aciertos y las limitaciones de esta última propuesta. Si bien hay evidencia a favor de la misma, la evidencia sobre la base neuronal de *mindreading* en su conjunto parece sugerir que esta capacidad es propiamente una función de la red y no de alguna región cerebral especializada.

El método más usual para medir la actividad neuronal son las imágenes por resonancia magnética funcional (IRMf). Este método divide al cerebro en numerosas unidades cúbicas de volumen y monitorea, en el tiempo, el nivel de oxígeno en cada uno de estos *voxels*. Esto otorga una alta resolución espacial (de 3-6 milímetros), pero una baja resolución temporal. Si bien el nivel de oxígeno en la sangre está fisiológicamente ligado a la actividad neuronal local, en este método de registro hay una demora significativa. Una única neurona puede disparar 200 veces por segundo, pero la resolución temporal de la IRMf es de unos segundos. Esto significa que los estudios de IRMf tienden a hacer hincapié en las diferencias de ubicación de la activación neuronal entre las diferentes condiciones, más que en la medición temporal de los procesos funcionales.

Un problema esencial que presenta este método es que, al registrar la actividad neuronal del participante mientras lleva a cabo una tarea de *mindreading*, el registro resultante también incluye la actividad neuronal correspondiente al procesamiento perceptual y cognitivo general relacionado con el estímulo que se utiliza en la tarea (*i. e.* caricaturas, videos, estímulos verbales). El método usual para resolver este problema consiste en que los participantes resuelvan una segunda tarea (la tarea de sustracción o control) que implica las mismas demandas perceptuales y cognitivas

generales, pero sin la demanda de *mindreading*. Si estas tareas están bien combinadas, al sustraer la actividad neuronal registrada durante la tarea de control, se podrá aislar la activación que está asociada sólo a *mindreading*. De modo que, el punto crítico de este método es que las tareas de interés y de control se tienen que corresponder en las demandas perceptuales y cognitivas generales. Asimismo, es preciso que se posea un buen enfoque de qué implica *mindreading* para poder diseñar tareas experimentales adecuadas que permitan evaluar la base cognitiva (los procesos cognitivos involucrados) de la función que se estudia.

Una serie de estudios ha proporcionado evidencia contundente sobre la existencia de una red neuronal para *mindreading* utilizando tareas de sustracción. Ésta abarca las siguientes áreas cerebrales: la juntura temporo-parietal derecha e izquierda (JTP-D y JTP-I), los polos temporales (PT), el precúneo medial, la corteza cingulada posterior (CCp) y la corteza prefrontal medial (CPFm). Como mencioné, en estos estudios, se han utilizado numerosos estímulos (caricaturas, historias, animaciones de figuras geométricas "intencionales", juegos estratégicos que implican adelantar la estrategia del competidor y demás) y se han requerido razonamientos sobre una gran variedad de estados mentales tales como creencias, deseos, conocimiento e intenciones. Esto ha generado confianza en la propuesta de que esta red está implicada en *mindreading* con independencia de si la tarea utilizada es visual o verbal, o de si involucra pensar sobre estados mentales en general o en particular (*p.e.* en creencias falsas). La consistencia de los estudios ha llevado a algunos a denominar a este conjunto de regiones cerebrales la "red de la 'Teoría de la Mente'" (Gallagher & Frith 2003; Carrington & Bailey 2009) o la "red de *mentalizing*" (Frith & Frith (2003). Es más, algunos han sugerido que esta especialización neuronal constituye evidencia a favor de un módulo cognitivo especializado en *mindreading* (Gallagher & Frith 2003; Leslie 2005).

No obstante, la red de *mindreading* no parece estar especializada. El rango de funciones que se han asociado a estas regiones (la CPFm, la JTP, los PT y el PC) es muy amplio. Algunos enfoques sugieren que estas áreas tienen un rol en la integración y el

procesamiento de información abstracta, porque están estrechamente conectadas entre sí y, a su vez, con muchas otras regiones del cerebro. Según este enfoque, *mindreading* es una función más entre las distintas funciones de dominio general asociadas a estas regiones tales como recuperar información de la memoria, la autoatribución, el razonamiento en general, el razonamiento inductivo bajo incertidumbre (Legrand & Ruby 2009). Otros le asignan funciones más especializadas a cada una de las regiones (Frith 2007). Por ejemplo, se ha sugerido que la JTP está implicada en el procesamiento de la toma de perspectiva, ya sea visual o mental. Si bien ésta, y otras funciones sugeridas para el resto de las regiones, son más especializadas, no lo son tanto como para sostener que la red constituye el módulo específico de *mindreading*.

Una de las tesis más fuertes respecto de la base neuronal de *mindreading* es que existe una región cerebral especializada en los pensamientos ajenos (Saxe & Kanwisher 2003). Para sostener esto, los investigadores se basan en un método que han diseñado para identificar las regiones de interés para *mindreading*. Se trata de un "localizador de la función", que también se puede utilizar para cualquier otra capacidad cognitiva. Este método está basado en el método de sustracción mencionado más arriba. Saxe y sus colaboradores usan una tarea de falsa fotografía (TFF) como la tarea de sustracción en relación a una tarea de falsa creencia (TFC) para estudiar *mindreading*[68]. En base a esta combinación de tareas, se analiza el contraste entre las activaciones asociadas a las mismas para localizar la función. Primero, se determinan las áreas más activas durante la TFF y la TFC. En este caso, se han identificado tres regiones de interés: la JTP bilateral, el PC y la CPFm. Luego, se resta la actividad

[68] En la literatura del desarrollo, se asume que si *mindreading* implica razonar sobre actitudes de un agente hacia una representación mental (Juan cree que llueve), parece apropiado comparar esto con el razonamiento acera de representaciones no mentales, por ejemplo, fotografías. En la tarea de falsa fotografía (Zaitchik 1990; Perner & Leekman 1991), se crea una secuencia similar a la de la tarea de falsa creencia. La situación inicial del objeto en una caja roja es captada por la creencia de *Sally* y por la fotografía. Luego, la ubicación del objeto se cambia a la caja azul. A continuación, se le pregunta a los niños dónde cree *Sally* que está el objeto o (con la fotografía cara abajo) dónde está el objeto en la fotografía. El problema con esta tarea es que las fotografías falsas no son realmente "falsas". La fotografía es la representación de una situación anterior y no de la situación actual, en este sentido, no es falsa. Contrariamente, la creencia de *Sally* es sobre la ubicación actual del objeto y en este sentido es falsa.

asociada a la TFF y así, se aísla la activación que se asocia a pensar sobre falsas creencias pero que no se asocia a pensar sobre fotografías falsas.

En las regiones mencionadas se observa una respuesta neuronal significativa ante historias sobre deseos y creencias, pero no ante historias que involucran situaciones físicas o que describen características físicas del los individuos (altura, peso). De modo que, las regiones que más se activan durante las TFC también se activan selectivamente cuando los participantes piensan en estados mentales, pero no cuando piensan meramente sobre otras personas. Estas conclusiones fueron refinadas en estudios posteriores y, en su conjunto, estos sugieren que la JTP-D y, en menor medida, la JTP-I y el PC responden selectivamente a estímulos que involucran estados mentales, y responden mucho menos a estímulos que describen situaciones físicas o de apariencia física de las otras personas (Saxe & Kanwisher 2003). El uso del método sustracción con la combinación de TFC-TFF ha conducido a una hipótesis fuerte. La JTP-D está especializada en la atribución de creencias a otras personas. Estos resultados han sido replicados por Perner *et al.* (2006), que agregan una condición de signos falsos[69]. Los resultados de Perner y colaboradores sugieren, en concordancia con la afirmación de Saxe y colegas, que la JTP-D se especializa en "pensar" sobre estados mentales, más que en problemas de toma de perspectiva.

No obstante, los estudios de sustracción tienen limitaciones. En particular, existe la posibilidad de que el método de sustracción TFF-TFC excluya procesos y

[69] En virtud de la crítica realizada a la TFF (mencionada en la nota 68), se han desarrollado otras tareas con representaciones que sean genuinamente falsas, por ejemplo, con señales falsas. "La señal que indica que el monasterio está en dirección al bosque fue dada vuelta por los niños mientras jugaban. De acuerdo a la señal, ahora ¿el monasterio está en dirección al bosque o al club de golf?". Perner *et al*. (2006) agregan esta condición al estudio de la base neuronal de *mindreading* con los estímulos de Saxe y colaboradores. Este experimento replica los resultados de Saxe en relación a las activaciones comunes en las TFC y TFF, y las activaciones de la JTP-D para pensar sobre estados mentales. El hecho de que la JTP-D se active selectivamente para razonar sobre estados mentales y no simplemente ante la toma de perspectiva se interpreta en el sentido de que es más sensible al contenido de los pensamientos (pensar sobre creencias) que al tipo de procesamiento implicado (toma de perspectiva). Sin embargo, este estudio encuentra otro patrón de activación para la JTP-I. Ésta está significativamente más activada durante las tareas de signos falsos que en las TFF. Esto sugiere que las regiones no se dedican selectivamente a los estados mentales ajenos, sino a la toma de perspectiva. De modo que, según la interpretación de los investigadores, *mindreading* parece ser uno entre los muchos tipos de problemas de toma de perspectiva.

activaciones neuronales que son críticos para el entendimiento apropiado de la habilidad de razonar con creencias (Apperly 2011). Desde una perspectiva funcional, la literatura sugiere que razonar con creencias implica mucho más que poseer conceptos de creencia, implica realizar inferencias que pueden requerir demandas de funciones ejecutivas. Desde una perspectiva neuroanatómica, los estudios de Saxe y sus colaboradores sugieren un rol menor de la CPF y del PC, y ningún rol para los PT, a pesar de que estas áreas están implicadas en numerosos estudios de *mindreading*. Esto puede tener dos explicaciones. O bien, los estudios anteriores a Saxe y colaboradores usaron tareas de sustracción menos específicas que no permitieron excluir actividad neuronal que es irrelevante para *mindreading*. O bien, esta combinación de tareas de sustracción identifica sólo un subconjunto de los sistemas neuronales involucrados en *mindreading*. En relación con esto, otros estudios sugieren un rol para la CPFm en *mindreading*. En el estudio de Gilbert *et al.* (2007) se manipula la presencia o ausencia de la demanda de *mindreading* junto a otra tarea de interés[70]. Según este estudio, la condición de *mindreading* se muestra asociada a activaciones significativas de la CPFm y el PT derecho pero no a activaciones de la JTP, tal como sostienen Saxe y sus

[70] En el estudio de Gilbert *et al*. (2007), los participantes realizan tareas que requieren dirigir su atención hacia un objeto disponible perceptualmente (*p.e.* presionar botones para girar a la izquierda o a la derecha mientras navegan mentalmente en un laberinto presentado visualmente) o dirigir su atención a un objeto generado internamente (la misma tarea pero con una versión del laberinto imaginada mentalmente). Estudio anteriores han mostrado una mayor activación de la CPFm cuando los participantes dirigen su atención hacia el estímulo presentado visualmente en comparación con el estimulo imaginado mentalmente (Gilbert, Frith & Burgess 2007). Esto parece contradictorio con los estudios que sugiere que la misma región está implicada en al autoreflexión (Vogeley *et al.* 2001). Retomando el estudio, los intervalos entre las diferentes fases de las tareas mencionada son largos o cortos, y en la mitad de los bloques experimentales Gilbert *et al*. (2007) les pidieron a los participantes que juzguen el tiempo de los mismos. Esta es la condición de no *mindreading*. Para la otra mitad de los bloques, se les dice a los participantes que un experimentador controla el tiempo de los intervalos y, al final del bloque, tienen que juzgar si el investigador fue colaborador o no (condición de *mindreading*). Es preciso mencionar que los intervalos son iguales en ambas condiciones (*mindreading* y no *mindreading*). El segundo hallazgo asociado a este estudio, es que las regiones de activación de la CPFm asociada a *mindreading* no mostraron solapamiento con las áreas activadas durante la orientación de la atención a un estímulo externo. De modo que no hay contradicción en entre los hallazgos mencionados anteriormente (Vogeley *et al.* 2001; Gilbert, Frith & Burgess 2007). Una interpretación de esto es que la CPFm tiene un rol en *mindreading* que es independiente de su rol en la orientación de la atención (Apperly 2011). Una interpretación alternativa, es que el rol de la CPFm se asocia a la orientación atencional sólo que las regiones más caudales se especializan en esta operación con un propósito social (Gilbert *et al*. 2007).

colaboradores. En relación a esto, es preciso señalar que el tipo de juicio que realizan los participantes en el estudio de Gilbert *et al*. (2007) es general. Esto es, si el investigador se comporta de manera colaborativa con el participante o no. Este tipo de juicio no requiere juicios asociados a creencias y deseos, los que usualmente se asocian a las activaciones de la JTP-D, o juicios de toma de perspectiva, que se asocian a la activación de la JPT-I.

El propósito de este panorama, por cierto bastante complejo, es mostrar que hay evidencia que apoya la hipótesis de Saxe respecto de que la JTP-D es una región especializada en la atribución de creencias (mientras que la JTP-I está más especializada en la toma de perspectiva). Ahora bien, el hecho de que otros estudios sugieran que ciertas regiones que los estudios de Saxe excluyen parecen tener un rol en *mindreading*, particularmente en lo que respecta a la CPFm, no cuestiona el hallazgo de Saxe. Se cuestiona la afirmación de que la base neuronal de *mindreading* se limite a esta región (la JTP-D). El uso de los localizadores funcionales puede excluir regiones importantes de la red de *mindreading*. En este sentido, la activación de la CPFm parece estar asociada a atribuciones más generales (*p.e.* si el experimentador es colaborador o no) y no a la atribución de una creencia específica al experimentador. La cuestión es que con el método de localización de función no se pueden abordar todas las preguntas sobre *mindreading*.

3. Los estudios que ponen a prueba las predicciones de la "Teoría de la Teoría" y la "Teoría de la Simulación"

Un grupo de estudios se dedican a evaluar las predicciones que se desprenden de los enfoques de teoría y simulación. Se trata de unos pocos estudios con técnicas de neuroimagen que han puesto a prueba estas predicciones para tratar de determinar el tipo de proceso subyacente a *mindreading*, aunque con resultados dispares (Vogeley *et al.* 2001; Ramnani & Miall 2004; Grezes, Frith & Passingham 2004). Del análisis particular de estos estudios, y de los supuestos que se asumen en los mismos,

mostraré por qué resulta crítico que los enfoques híbridos de teoría y simulación puedan proveer un criterio para distinguir entre los procesos subyacentes.

Como he mencionado anteriormente, desde el inicio del debate teoría-simulación, se ha asumido que de los enfoques de TT y TS se desprenderían predicciones que, al ser puestas a prueba empíricamente, permitirían decidir entre estas alternativas. Ante el fracaso de los estudios en psicología para ofrecer evidencia a favor o en contra, la expectativa se posó en los estudios de neurociencia, particularmente, en los estudios de neuroimagen (Stich & Nichols 1997). Sin embargo, hasta el momento la neurociencia cognitiva social no ha tenido más éxito que los estudios comportamentales para ofrecer evidencia contundente que permita discriminar entre los enfoques. El desafío para los investigadores experimentales continúa siendo el mismo. Dado un juicio acerca de otra persona ¿es posible diseñar un experimento que provea evidencia clara sobre el rol de la simulación o de procesos ricos en información? (Apperly 2008: 270). En los últimos años, los estudios de neuroimagen de *mindreading* se han multiplicado. Si bien la TT y la TS forman parte del marco teórico de los mismos, los estudios se han dedicado a estudiar el fenómeno pero no a evaluar las predicciones específicas que surgen de estos enfoques.

No obstante, algunas de estas predicciones se asumen como afirmaciones en las investigaciones empíricas. En particular, se asume que si se observa activación en las mismas regiones cerebrales cuando se lleva a cabo una acción (o se realiza una tarea cognitiva desde la perspectiva propia) y cuando se predice la acción de otra persona (o se realiza la tarea desde la perspectiva ajena), entonces el proceso subyacente a *mindreading* es de tipo simulacional. Esto, en virtud de que la TS sostiene que los procesos cognitivos que utilizamos para desempeñarnos en una tarea pueden servir como un modelo útil de los procesos cognitivos que otra persona usaría para desempeñar la misma tarea (Apperly 2008: 272). De modo que, si la capacidad para anticipar un juicio ajeno emplea los mismos procesos cognitivos que se utilizan para realizar el juicio propio, la TS afirma que ambas capacidades (atribución y predicción)

están funcionalmente relacionadas de manera estrecha y predice que emplearán los mismos mecanismos neuronales (Vogeley *et al*. 2001: 171).

Algunos estudios de neuroimagen han puesto a prueba esta predicción para tratar de determinar el tipo de proceso subyacente a *mindreading*, aunque con resultados dispares. Grezes, Frith & Passingham (2004) encuentran activaciones solapadas y sugieren un proceso de simulación subyacente, mientras que Ramnani & Miall (2004) observan activación en una misma región cerebral, aunque en áreas de la misma que no se solapan. Esto sugiere que se utilizan distintos sistemas cerebrales de modo que, según los investigadores, esta evidencia favorece a un enfoque de TT. De la cuestión de que la activación en diferentes regiones neuronales brinda apoyo empírico a la TT me ocupo más adelante, en esta misma sección. En el tercer estudio, se ha observado activación solapada en ciertas regiones cerebrales y, al mismo tiempo, activación en regiones diferentes para la condición "yo" (Vogeley *et al*. 2001). Según los investigadores, esto sugiere que las habilidades de realizar juicios y anticipar juicios ajenos dependen, a la vez, de mecanismos neuronales comunes y diferentes. Este último hallazgo se ha interpretado a favor de la existencia de procesos híbridos de teoría y simulación subyacentes a *mindreading*.

Además de la disparidad de los resultados, se ha señalado que los estudios mencionados son controversiales. La principal fuente de controversia proviene de cuestiones conceptuales que necesitan ser atendidas si se pretenden estudiar las predicciones que se siguen de los enfoques de TT y TS con técnicas de neuroimagen (Apperly 2008). En primer lugar, es preciso determinar qué cuenta como una tarea de *mindreading*. En este sentido, las tareas deben involucrar la atribución de estados mentales tales como creencias, intenciones y otras actitudes proposicionales. De modo que, si la condición "otro" consiste en predecir la acción de otra persona, la condición "yo" tiene que consistir en realizar una acción similar. Además, el diseño experimental tiene que asegurarse de que la predicción del comportamiento de la otra persona involucre necesariamente la consideración de estados mentales. Como es sabido, el comportamiento puede anticiparse simplemente como una respuesta a un estímulo,

sin necesidad de recurrir a la atribución de estados internos o mentales. Este es el problema del estudio de Ramnani & Miall (2004). Aquí, *mindreading* no se evalúa adecuadamente porque sólo se requiere que los participantes anticipen respuestas en relación con simples conjuntos de emparejamiento estímulo-respuesta[71]. De modo que los estados mentales no están necesariamente involucrados en las predicciones, ni en las activaciones asociadas (Apperly 2008).

La segunda cuestión conceptual se relaciona con el diseño de la condición "yo" para la tarea de *mindreading*. Si bien existen varios abordajes conceptuales sobre el "yo", no todos estos enfoques aportan una noción de proceso relacionado al "yo" que sea adecuado para diseñar una tarea de *mindreading* (Apperly 2008). Para poder comparar las activaciones que se corresponden a la condición de atribuciones de primera y de tercera persona, según los postulados de la TT y la TS, la condición "yo" (autoatribución) debería involucrar al participante con la experiencia de un estado mental ocurrente, que éste pueda autoadscribirse. En particular, porque según la TS, la simulación implica el recurso a los mismos sistemas (mentales o cerebrales) subyacentes a la experiencia de un estado mental para realizar atribuciones de ese mismo estado. En otras palabras, la TS asume que el participante usa los procesos causales de su cognición para modelar estos mismos procesos en otra mente. En este sentido, considero que Apperly (2008) está en lo cierto cuando sostiene que no es lo mismo autoatribuirse estados mentales ocurrentes que autoatribuirse rasgos de personalidad, como suelen involucrar algunos estudios (*p.e.* un juicio de similitud). Por más que la autoatribución de un rasgo de carácter pueda considerarse una creencia acerca de tal rasgo de carácter, no parece ser el tipo de creencia que alimenta al

[71] El propósito del estudio de Ramnani & Miall (2004) es evaluar si los procesos neuronales involucrados en la preparación de las acciones propias están también involucrados en la predicción de las acciones futuras de los otros. Dentro del resonador, el participante mira una pantalla en la que se indica con una señal (*i.e.* un color) si él o el otro participante tiene que realizar una acción. Además, se indica con señales distintas (*i.e.* formas geométricas) qué movimiento realizar con los dedos. El problema de esta tarea es que consiste en la mera asociación entre un estímulo (*i.e.* forma geométrica) y una respuesta (*i. e.* movimiento con los dedos). De modo que, al momento de predecir la respuesta de la otra persona al estímulo, basta con conocer el movimiento que va a asociado a la forma geométrica, sin necesidad de atribuir al otro, estados mentales o internos.

sistema de toma de decisiones para decidir qué acción llevar a cabo (el caso paradigmático de la simulación).

De modo que, el enfoque de la TS implica que la condición "yo" tiene que asociarse con la autoatribución de un estado mental ocurrente, más que con la autoatribución en general. Una condición de autoatribución adecuada requiere que los participantes tengan creencias, deseos o intenciones ocurrentes y que éstas, a su vez, sean relevantemente similares a las que se les adscriben a los otros en la condición "otro". Así, si se observa una activación solapada en las condiciones de primera y de tercera persona, se puede interpretar que se usan los mismos recursos funcionales y neuronales (la predicción simulacional). El problema del estudio de Grezes, Frith & Passingham (2004) es que los juicios de los participantes en la condición "yo" se realizan respecto de filmaciones en las que se observan a ellos mismos como agentes. Y adscribir un estado mental a un "yo" observado no es lo mismo que estar en un estado mental, en tanto los participantes están tomando una actitud de observadores. De modo que, es posible que las activaciones estén solapadas porque, en ambas condiciones, los participantes realizan juicios a partir de la observación de agentes, ya sea uno mismo como agente (condición "yo") u otra persona como agente (condición "otro"). En este sentido, el diseño no es adecuado para testear TT *versus* TS, porque la condición "yo" no recoge el requerimiento de que el participante esté en un estado mental actual (Apperly 2008).

Ahora bien, si la condición "yo" no implica la autoatribución de un estado mental ocurrente sino de un juicio autoatributivo de algún rasgo de personalidad, puede ser el caso que el solapamiento de las activaciones asociadas con las dos tareas esté indicando alguna relación entre *mindreading* de tercera persona y la autoatribución, que no sea la relación que supone la TS. Por ejemplo, en el caso del estudio de Mitchell, Banaji & McRae (2005), la evidencia sugiere que la tarea de *mindreading* implica un juicio de similitud (autoatribución) en relación con la persona que se observa en la foto (capítulo 5, sección 1.2). Como ya mencioné, esta relación no es algo que se siga de la propuesta simulacional de Goldman (2006). En esta propuesta,

si bien se postula un rol para las autoatribuciones en *mindreading* de tercera persona, éstas no se relacionan con los juicios de similitud. Específicamente, se requieren autoatribuciones relacionadas con el monitoreo de los estados mentales propios en la rutina de simulacional, por ejemplo, para excluir los estados mentales genuinos con el propósito de que *mindreading* sea correcto. En mi opinión, no se puede equiparar la autoatribución implicada por el juicio de similitud y la autoatribución requerida por el monitoreo de estados mentales propios. De modo que, la relación señalada en este estudio debe ser algún otro tipo relación entre *mindreading* de tercera persona y la autoatribución. Se ha sugerido, por ejemplo, que este tipo de juicios de similitud en relación con uno mismo, quizás no sean fundamentales para *mindreading* sino que estén implicados en aquellas tareas de *mindreading* que se llevan a cabo en condiciones de incertidumbre. De modo que, ante la falta de información, estos juicios permitirían acotar el conjunto de los posibles pensamientos que pueden tener las otras personas a un subconjunto abarcable (Jenkins & Mitchell 2009). En suma, en virtud de la activación observada en regiones cerebrales asociadas a la autoreflexión, en el contexto de una tarea de *mindreading* de tercera persona, parece existir una relación entre estas capacidades pero no se conoce su naturaleza, ni es claro que esta relación brinde apoyo empírico a un enfoque de TS.

Si bien los defectos en los diseños experimentales mencionados son graves, considero que hay un problema más de fondo que se relaciona con las predicciones asociadas con la TT y la TS. Estas predicciones, que han guiado los estudios con técnicas de neuroimagen, se asumen como obvias, pero considero que hay razones para cuestionarlas. En primer lugar, en todos estos estudios se asume que la activación neuronal solapada sugiere un proceso de tipo simulacional subyacente. Esto está relacionado con un razonamiento ya mencionado. Según la TS, tal como la entienden los neurocientíficos, si la capacidad para anticipar un juicio ajeno emplea los mismos procesos cognitivos que se utilizan para realizar el juicio, ambas capacidades están funcionalmente relacionadas de manera estrecha y emplearán los mismos mecanismos neuronales. Además, esta afirmación ha encontrado apoyo empírico en la evidencia

proveniente de los procesos espejo, como he mencionado en el capítulo 5. Estas neuronas, ubicadas principalmente en áreas cerebrales premotoras, se activan selectivamente ante la ejecución de una acción motora y ante la observación de la misma acción motora en otros agentes. No obstante, como señalé en el capítulo 5, el rol de las neuronas espejo en *mindreading* es controversial.

Más allá de esta controversia, a mi entender, el problema de fondo reside en que se asume que un enfoque de TT no puede generar la predicción de una activación cerebral solapada. Considero que hay razones para sostener que esto no es así. En paralelo a la afirmación de que sólo la TS predice la activación solapada, se afirma que si el patrón de activación asociado a *mindreading* muestra áreas cerebrales diferentes esto brinda apoyo a un enfoque de TT (Vogeley *et al.* 2001; Frith & Frith 2003; Grezes, Frith & Passingham 2004; Ramnani & Miall 2004; Apperly 2008). Esta predicción se relaciona con la afirmación de que, según el enfoque de TT, las predicciones sobre los otros dependen de diferentes conjuntos de procesos (que implican conceptos y principios) diferentes a los implicados cuando nuestras propias creencias, deseos e intenciones forman la base (causal) de nuestro comportamiento (Apperly 2008: 270). En base a esto, considero que los investigadores identifican al enfoque de TT con una de sus versiones. La versión que postula la posesión de un cuerpo de información para realizar los juicios propios y de otro cuerpo de información diferente, sobre cómo los otros llevan a cabo juicios, para realizar atribuciones a otras personas (Harris 1995; Stich & Nichols 2003; Nichols & Stich 2003). Sin embargo, como he señalado en el capítulo 4, un enfoque rico en información no implica necesariamente postular la reduplicación de teorías. En particular, cuando nos referimos a juicios que implican creencias y deseos, una de las versiones más difundidas de la TT, el “enfoque del niño científico”, afirma que utilizamos un mismo cuerpo de información o teoría para atribuir estados mentales, y explicar y predecir el comportamiento propio y ajeno (Sellars 1956; Gopnik & Wellman 1994; Gopnik 1993; Gopnik & Meltzoff 1994).

De modo que, la activación solapada puede estar relacionada, a su vez, con un proceso de tipo rico en información si se adopta esta versión de la TT. En este sentido,

considero que del enfoque de TT también surge la predicción de una activación en regiones cerebrales solapadas para las atribuciones de tercera y de primera persona. Se trata de la activación de las regiones neuronales asociadas a los procesos cognitivos que emplean tal cuerpo de información. Ahora bien, la consecuencia crítica que se sigue de esto es que ya no es claro que sólo el enfoque de TS predice la activación solapada de regiones neuronales. Si esta predicción no sólo es generada por la TS, ya no hay un criterio que permita "leer" las activaciones neuronales como brindando apoyo a un enfoque de teoría o a un enfoque de simulación.

En este sentido, considero que es crítico que se pueda proveer un criterio para distinguir entre teoría y simulación. Si la versión de la TT que postula el recurso a una única teoría puede predecir la activación neuronal solapada esto implica que, por más que los procesos espejo sean considerados simulaciones y muestren activación solapada, no toda activación solapada sugiere procesos espejo. Los procesos espejo implican activaciones solapadas pero no toda activación solapada implica procesos espejo. De modo que, si se hallara activación solapada en regiones asociadas a procesos de autoatribución y a *mindreading* de tercera persona, sobre todo en regiones que no están asociadas con los procesos espejo (como parece ocurrir en el caso de la CPFm), quizás no pueda afirmase simplemente que esto favorece a un enfoque de simulación. Como he mencionado, si un mismo cuerpo de información subyace a las atribuciones de primera y de tercera persona, es probable que se observe activación en regiones solapadas. De modo que, el solapamiento o no solapamiento de las activaciones en regiones cerebrales no proporciona un criterio para distinguir entre procesos subyacentes. Así, parece crítico proveer un criterio que permita determinar el tipo de proceso subyacente a un caso de *mindreading* para llevar adelante estudios de neuroimagen que intenten poner a prueba sus predicciones o bien utilicen estos dos enfoques como marco teórico.

4. Los lineamientos para un enfoque híbrido de *mindreading* de tercera persona

Una serie de estudios en neurociencia cognitiva social han brindado propuestas híbridas de teoría y simulación (Uddin *et al*. 2007; Pineda & Hetch 2008; Keysers & Gazzola 2006, 2007). En principio, estas propuestas de enfoques híbridos no resultan satisfactorias para dar cuenta de *mindreading* en tanto no dan cuenta de la autoatribución. No obstante, a mi entender, sugieren un criterio para determinar el tipo proceso subyacente a *mindreading* de tercera persona. En particular, me interesa la propuesta especulativa de Keysers & Gazzola (2006, 2007), que intenta reconciliar los hallazgos de correlatos neuronales y da cuenta de una dificultad, que se mencionó en el capítulo 5 (sección 1.2), respecto de los procesos espejo como subyacentes a *mindreading*.

Keysers & Gazzola (2006, 2007) sostienen que los estudios de las bases neuronales de *mindreading* han señalado distintas áreas subyacentes, según se adhiera a un enfoque de TS o TT. Esto se debe a que los estudios utilizan tareas muy diferentes. En este sentido, los estudios están evaluando distintos fenómenos de *mindreading*, en particular, los extremos de un espectro (atribuciones muy simples y muy complejas). Si se comparan el tipo de tareas que se utilizan en los estudios de neuroimagen de procesos espejo y en los estudios de neuroimagen que emplean tareas de falsa creencia, es claro que las tareas difieren dramáticamente en la cantidad de pensamientos explícitos que involucran (2006: 400). En los estudios de neuroimagen de procesos espejo, los participantes usualmente observan videos cortos relacionados con acciones, emociones o sensaciones sin que se les requiera reflexionar sobre el significado de los estímulos o los estados mentales de los agentes observados. En las tareas de falsa creencia, en cambio, se les pide a los participantes que piensen acerca de los estados mentales ajenos. Los estudios de neuroimagen que emplean tareas de falsa creencia señalan a la CPFm como la región subyacente, en comparación con tareas que no requieren *mindreading*.

Keysers & Gazzola (2006, 2007) señalan que los procesos espejo pueden dar cuenta de nuestra comprensión de la vida interna de otras personas pero con limitaciones. La evidencia sugiere que hay sistemas espejo para las acciones, las

sensaciones y las emociones. La corteza premotora para las acciones, la ínsula para el asco, la corteza cingulada anterior y la ínsula anterior para el dolor, las cortezas somato sensoriales para el tacto, la amígdala para el miedo. En todos estos casos, observar lo que otras personas hacen o sienten es transformado en una representación interna de una acción o una sensación propia en una situación similar[72]. No obstante, hay cierto entendimiento al que estos procesos no pueden subyacer. En particular, aquel que implica la imaginación abstracta o desacoplada de la experiencia corporal. Por ejemplo, no debería adscribirse a los procesos espejo la capacidad para imaginar qué se siente volar como un pájaro (Keysers & Gazzola 2006: 399). Así, una vez que los procesos espejo han transformado las acciones, las sensaciones y las emociones de los otros en nuestras propias representaciones de acciones, sensaciones y emociones, entender a las otras personas se reduce a entendernos a nosotros mismos, en el sentido de, entender nuestras propias acciones, sensaciones y emociones (Keysers & Gazzola 2006: 395).

Ahora bien, curiosamente, se ha señalado que regiones similares de la CPFm están involucras en *mindreading* de tercera persona y en la autoreflexión sobre las emociones (Gusnard *et al.* 2001). Esto ha llevado a algunos investigadores a especular que los procesos que subyacen a la capacidad de pensar sobre otras personas pueden estar relacionados con los procesos que subyacen a la autoreflexión. Keysers & Gazzolla (2006) consideran esta propuesta como un marco para asociar los procesos espejo con *mindreading*. La CPFm recibe *inputs* indirectos del cuerpo incluyendo regiones donde se alojan representaciones motoras, somatosensoriales y viscerales, que pueden permitir una representación secundaria del estado del propio cuerpo (Frith & Frith 2003). Keysers & Gazzola especulan que en este punto los procesos espejo se conectan con *mindreading*. Mediante los procesos espejo, las sensaciones, las emociones y las acciones de los otros son "traducidas" al lenguaje neuronal de las propias sensaciones, emociones y acciones. De modo que, éstas se han transformado

[72] En rigor, Keysers & Gazzola (2006, 2007) denominan a los procesos espejo "circuitos compartidos" pero en la reconstrucción de su propuesta utilizaré el término "procesos espejo" que es el que he utilizado a lo largo de la tesis para nombrar a estos sistemas cerebrales.

en representaciones primarias de tales estados. Este proceso puede generar un compartir implícito (una resonancia) y el consiguiente entendimiento de los estados de los otros. No obstante, en los humanos existen procesos explícitos que complementan estos procesos automáticos (206: 399). Según Keysers & Gazzola, si se pregunta explícitamente por el estado de otra persona, será preciso generar una representación secundaria, más consciente y cognitiva de este estado (Keysers & Gazzola 2006: 401). No obstante, puesto que este estado ya ha sido traducido al lenguaje de los estados propios, los investigadores especulan que esta tarea no será diferente a la tarea de reflexionar acera de los estados mentales propios y, en este sentido, se activarán las regiones de la CPFm asociadas a la autoreflexión. No obstante, como señalan los autores, esta es una hipótesis que tiene que ponerse a prueba con neuroimagen.

En particular, me interesa resaltar los siguientes aspectos de la propuesta. Los procesos espejo se consideran "traductores" de los estados mentales ajenos en estados mentales propios, en el sentido de que convierten los estados de las otras personas en representaciones primarias de estados propios. Ahora bien, según los autores estos estados se pueden compartir simplemente, y la cognición social puede terminar en esta instancia (nótese que no se trata de *mindreading*). Sin embargo, para que haya una atribución explícita que esté basada en las representaciones primarias que generan los procesos espejo tiene que operar una segunda transformación. Las representaciones primarias tienen que ser transformadas en representaciones secundarias más cognitivas, más conscientes. Con este propósito, intervienen los proceso cognitivos superiores asociados a las regiones de la CPFm. En otras ocasiones, parece preciso sólo reflexionar sobre los estados mentales ajenos, para obtener un entendimiento más elaborado y cognitivo de los mismos. Según Keysers & Gazzola, estos son los casos de atribución que se estudian, por ejemplo, con tareas de falsa creencia. De modo que, estos procesos de *mindreading* se colocan por encima de las simulaciones y permiten alcanzar niveles de sofisticación que van más allá de los que se pueden alcanzar usando sólo la simulación. Al usar la simulación asumimos que todos somos iguales, al recurrir a la teorización podemos concebir las diferencias (Keysers &

Gazzola 2006: 401). De modo que, según esta propuesta existen unos mecanismos pre reflexivos (los procesos espejo) que permiten compartir estados mentales. Estos, a su vez, pueden dar lugar a atribuciones explícitas (*mindreading*) pero, para que éstas se lleven a cabo, es preciso que intervengan procesos cognitivos superiores. A su vez, se pueden llevar a cabo atribuciones completamente por reflexión.

Como señalé en los capítulos 4 y 5, los enfoques híbridos de teoría y simulación analizados (Nichols & Stich 2003; Goldman 2006) no son satisfactorios porque no proveen un criterio para determinar el tipo de proceso que subyace a un caso de *mindreading*. Desde la perspectiva de la neurociencia cognitiva social, se ha propuesto un criterio que permite determinar qué tipo de proceso subyace a un caso de *mindreading*. La adopción de este criterio puede articular de un modo novedoso los enfoques híbridos de teoría y simulación para *mindreading* de tercera persona. Considero que la propuesta de Keysers & Gazzola (2006, 2007) tiene, por un lado, la ventaja de explicar cómo los procesos espejo podrían subyacer a *mindreading* pero, por el otro, la desventaja de un costo para el enfoque de TS. El costo consiste en que, primero, la simulación sólo es implementada por los procesos espejo. En este sentido, procesos tales como la imaginación enactiva postulada por Goldman (2006) quedan excluidos. Segundo, la simulación implementada neuronalmente por los procesos espejo, en principio, sólo puede subyacer a atribuciones automáticas, pre reflexivas, no explícitas y, en principio, esto no parece constituir *mindreading* propiamente dicho, sino algún otro tipo de fenómeno de cognición social como la resonancia (quizás el contagio emocional o la empatía, incluso cierto tipo de comportamientos imitativos). Tercero, para que el entendimiento de las otras personas provisto por los procesos espejo se vuelva una atribución explícita, las representaciones primarias generadas por los procesos espejo tienen que transformarse en representaciones secundarias con la mediación de los procesos cognitivos superiores. Estos se relacionan con la conceptualización y la verbalización. A mi entender, este aspecto de la propuesta puede considerarse una solución al problema mencionado en el capítulo 5 (sección 1.2), respecto de la posibilidad de que los procesos espejo subyazcan a *mindreading*.

Cuarto, y no menos importante, las atribuciones a las que puede subyacer la simulación implementada por los procesos espejo se acota a los procesos espejo que existen. Según la evidencia, los procesos espejo intervienen en el procesamiento de estados mentales tales como la intención motora, ciertas sensaciones (tacto y dolor) y ciertas emociones (miedo, asco, ira).

De modo que, de este enfoque especulativo se sigue que los procesos espejo proveen un entendimiento simple, primitivo, automático y pre reflexivo de las otras personas. En este sentido, pueden dar lugar a varios fenómenos de cognición social (como la resonancia) pero, para que tenga lugar *mindreading*, las atribuciones tienen que ser explícitas en el sentido de ser verbalizables, conceptualizables, reflexivas y conscientes. En este sentido, *mindreading* basado en los procesos espejo implica que las representaciones generadas por tales procesos se volvieron explícitas mediante la intervención de algún proceso de cognición superior. Por supuesto, *mindreading* también puede llevarse a cabo completamente por un proceso completamente reflexivo. En este caso se espera que sólo las regiones frontales y temporo-parietales sean reclutadas. Estas están asociadas al procesamiento de estados mentales más complejos como las creencias. En este sentido considero que la propuesta brinda un criterio que implica una división de tareas entre los procesos de teoría y simulación respecto de los tipos atribuciones a los que los procesos subyacen. En particular, en relación al tipo de estado mental y a la condición de reflexivo o pre reflexivo. La simulación sólo subyace atribuciones de estados emocionales, sensaciones e intenciones (estados para los que hay procesos espejo) pero no a atribuciones, por ejemplo, de falsa creencia. Además, subyace a las atribuciones pre reflexivas, es decir cuando las atribuciones de emociones, sensaciones e intenciones (motoras simpes) se tienen que verbalizar (y conceptualizar) es precisa la intervención de un proceso de cognición superior. Por todas estas razones, el enfoque simulacional implicado por esta propuesta es muy diferente a la propuesta de Goldman (2006), aunque propone un modo de dar cuenta de la afirmación de que los procesos espejo subyacen a las atribuciones de *mindreading* (Goldman 2006).

5. El enfoque híbrido de teoría y simulación como una teoría de doble proceso

En base a la propuesta antes mencionada, un enfoque híbrido de teoría y simulación tendría que regirse por el siguiente criterio. Un proceso de tipo simulacional, implementado por procesos espejo, subyace a atribuciones de estados mentales simples tales como emociones, sensaciones e intenciones (motoras simples). Además, para que estas atribuciones devengan explícitas es preciso que los procesos primitivos, automáticos y pre reflexivos sean complementados por procesos cognitivos superiores que aporten conceptualización y verbalización. Ahora bien, no es claro que los procesos cognitivos superiores se correspondan con procesos ricos en información subyacentes a *mindreading*, ya que, en principio, no implican llevar a cabo inferencias a partir de generalizaciones psicológicas. El resto de las atribuciones se llevaría a cabo por procesos de tipo ricos en información. En particular, las atribuciones de actitudes proposicionales.

De modo que, desde el ámbito de la neurociencia cognitiva social, en base a la evidencia neuroanatómica y funcional, se propone un criterio claro para establecer el tipo de proceso que subyace a un caso de *mindreading*. Ahora bien, como ya he mencionado, el criterio propuesto implica una división de tareas a nivel funcional entre los procesos ricos en información y la simulación. Esto tiene dos consecuencias para los enfoques híbridos de teoría y simulación de *mindreading* de tercera persona. Primero, implica que el enfoque de simulación reduce su alcance explicativo, en tanto la simulación sólo subyace a atribuciones de emoción, sensaciones e intenciones motoras. Segundo, en la medida en que más de un proceso cognitivo subyace a una capacidad cognitiva, esta propuesta híbrida puede considerarse una propuesta de doble proceso. No obstante, a mi entender, ésta no se puede equiparar a otras propuestas de doble proceso en ciencias cognitivas (Machery 2009; Evans & Stanovich 2013) precisamente, por la supuesta división funcional entre los procesos. El propósito de esta sección es describir la particularidad de la propuesta de teoría y simulación de

Keysers & Gazzola (2006, 2007) como una teoría de doble proceso subyacente a *mindreading* de tercera persona.

Según Machery (2009), los psicólogos dividen la cognición en varias competencias cognitivas. Estas se definen funcionalmente, por aquello que producen[73]. Un proceso cognitivo es un modo específico de producir aquello para lo que es una competencia cognitiva. Los procesos cognitivos *subyacen* a las competencias cognitivas. Describir un proceso cognitivo implica caracterizar los pasos en virtud de los cuales, en base a sus *inputs*, se produce aquello para lo que es la competencia cognitiva[74]. En línea con esto, las Teorías Multi-Proceso (TMP) asumen que distintos procesos cognitivos le subyacen a una única competencia cognitiva. Sin embargo, esta afirmación no debe entenderse en el sentido trivial de una competencia conformada por sub-competencias, sino en el sentido de que cada uno de los procesos postulados es suficiente para producir la competencia cognitiva. De modo que, si todos los procesos menos uno fueran eliminados, el organismo aún poseería la competencia, aunque quizás sus productos difieran de aquellos que produce la misma cuando todos los procesos están en funcionamiento (Machery 2009: 127). En este sentido, *mindreading* se caracteriza como la "atribución de estados mentales" y, según las propuestas híbridas de teoría y simulación, los procesos cognitivos subyacentes a la competencia de *mindreading* son la simulación y los procesos ricos en información. Ambos constituyen modos de producir atribuciones de estados mentales. Ahora bien, la propuesta de Keysers & Gazzola (2006, 2007), a mi entender, no se puede concebir como una teoría de doble proceso en el sentido de Machery (2009)[75]. Un

[73] Pueden considerarse ejemplos de competencias cognitivas: reconocer caras visualmente, ser capaz de estimar visualmente la cardinalidad de una clase, o de estimar la cardinalidad de una secuencia de sonidos, ser capaz de clasificar objetos físicos en clases, ser capaz de determinar la estructura sintáctica de oraciones, distinguir fonemas, identificar sombras en el campo visual. Las competencias cognitivas están anidadas. Esto es, típicamente, tener una competencia cognitiva implica tener sub-competencias. Por ejemplo, la capacidad de distinguir objetos en nuestro campo visual implica ser capaz de identificar las sombras proyectadas por estos objetos (Machery 2009).

[74] Un proceso cognitivo, asumido por los defensores del enfoque de prototipos para la categorización, es comparar la representación de un objeto con un prototipo con el propósito de decidir si este objeto pertenece a la categoría representada por el prototipo (Machery 2009).

[75] La teoría de doble proceso es un caso particular de las teorías multi-procesos (Machery 2009).

ejemplo de TMP para Machery (2009) es la teoría de doble proceso que da lugar a los juicios morales (Greene & Haidt 2002), según la cual dos procesos subyacen a los juicios morales, uno de tipo emocional y otro de tipo racional. De modo que dañado un proceso aún se pueden llevar a cabo juicios morales aunque con un perfil de desempeño diferente del que surgiría si ambos procesos estuvieran intactos. Como ya he mencionado, según la división de funciones en la propuesta de Keysers & Gazzola (2006, 2007), la simulación subyace a las atribuciones de emociones, sensaciones e intenciones (motoras) y los procesos ricos en información subyacen al resto de las atribuciones, en especial, a las atribuciones de actitudes proposicionales (creencias y demás). En este sentido, si se dañara alguno de los procesos ¿qué perfil de desempeño surgiría de este daño? Por ejemplo, si se dañan los procesos espejo que subyacen a la emoción del asco, ya no sería posible atribuir asco, sino sólo el resto de los estados mentales. Si bien es cierto que esto se puede interpretar como un cambio en el perfil de desempeño de *mindreading*, tal como predice la caracterización de las TMP, en rigor ya no se puede atribuir asco en ningún sentido. En mi opinión, este no es el tipo de perfil de desempeño que tienen en mente los "teóricos de doble proceso". En todo caso, una teoría de doble proceso postularía dos procesos subyacentes a la atribución del asco, por ejemplo uno de cognición fría y otro de cognición caliente[76]. De modo que, si una ruta (un proceso) se avería, aún se podrían realizar atribuciones de asco a partir de la segunda ruta. A mi entender, este es el sentido en que los "teóricos de doble proceso" entienden el "perfil de desempeño diferente". La división funcional que implica la propuesta de Keysers & Gazzola (2006, 2007) sugiere que si se avería de uno de los procesos subyacentes, hay una serie de atribuciones que ya no se pueden realizar, y este no es el perfil de desempeño acorde con las teorías de doble proceso subyacentes a una misma competencia cognitiva. En este sentido, de la propuesta de teoría y simulación para *mindreading* de tercera persona que surge del criterio

[76] Algo de esto parece estar sugerido por la idea de que la atribución emocional se lleva a cabo por varios caminos, algunos primitivos como los procesos espejo o de resonancia y otros más cognitivos (o de cognición superior) como el procesamiento de la expresión facial en las regiones corticales visuales que conduce a la representación proposicional del estado emocional inferido (Wicker *et al*. 2003).

propuesto por Keysers & Gazzola (2006, 2007), dos procesos subyacen a *mindreading* pero no implican una teoría de doble proceso tal como se entiende usualmente en ciencias cognitivas.

6. Conclusión

En este capítulo me ocupé del estudio de *mindreading* con técnicas de neuroimagen y del problema esencial que se presenta en torno al diseño de la tarea de sustracción (sección 1). El panorama sobre las bases neuronales de *mindreading* es contundente. Una red está implicada en *mindreading* con independencia del tipo de tarea con que se estudie y del tipo de estado mental que se atribuya. Esta red parece estar especializada en cierta medida, pero el grado de especialización no es suficiente para sostener la existencia de un módulo de *mindreading*. En particular, se ha propuesto la existencia de un área especializada en el pensamiento sobre los estados mentales, en particular, de las creencias y los deseos (Saxe & Kanwisher 2003). Según el método de localización de la función de *mindreading*, basado en la TFF como tarea de sustracción de la TFC, se ha asociado la activación de la JTP-D a los juicios sobre los pensamientos de las otras personas. No obstante, esta propuesta debe tomarse con precaución porque deja afuera numerosas regiones señaladas por otros estudios como subyacentes a *mindreading*. En particular, la CPFm. En este sentido, se ha sugerido que el método de localización de función quizás no permita abordar todas las preguntas de *mindreading*.

En la sección 2, presenté ciertos estudios de neuroimagen que se han dedicado específicamente a evaluar las predicciones de TT y TS (Vogeley *et al.* 2001; Ramnani & Miall 2004; Grezes, Frith & Passingham 2004). Mencioné algunos problemas metodológicos que se han señalado en relación a tales estudios (Apperly 2008). No obstante, considero que el problema principal de los mismos se relaciona con las predicciones de TT y TS que asumen. Según la TS, las capacidades para realizar juicios y para anticipar juicios ajenos emplean los mismos recursos cognitivos. En este sentido,

si ambas capacidades (de juicio y predicción) están relacionadas funcionalmente, se espera que empleen los mismos mecanismos cerebrales. Así, los estudios mencionados asumen que la TS predice la activación solapada en regiones cerebrales para las condiciones de *mindreading* de primera y de tercera persona, pero la TT no. A mi entender, hay razones para sostener que esto no es así. Considero que esta misma predicción puede surgir de un enfoque de TT. Si se adhiere a la versión de TT que sostiene que un mismo cuerpo de información se emplea en atribuciones de primera y de tercera persona (enfoque del niño científico), es esperable que la activación neuronal resulte solapada en ambas condiciones. Si esto es así, ya no sólo el enfoque de TS predice la activación solapada en regiones neuronales. De modo que, se diluye el criterio que ha permitido "leer" las activaciones neuronales solapadas como indicando un proceso simulacional subyacente. Los procesos espejo (simulaciones) suponen activaciones solapadas, pero no toda activación solapada supone procesos espejo. En este sentido, considero que resulta crítico proponer un criterio satisfactorio para distinguir entre teoría y simulación con el propósito de estudiar *mindreading* con neuroimagen.

En la sección 4, he presentado los lineamientos para un enfoque híbrido de teoría y simulación de tercera persona basado en un criterio proveniente de la neurociencia cognitiva social (Keysers & Gazzola 2006, 2007). La adopción de este criterio puede articular de manera novedosa los enfoques híbridos de teoría y simulación. Recuérdese que los enfoques analizados en los capítulos 4 y 5 de la tesis (Nichols & Stich 2003; Goldman 2006) no han logrado proporcionar un criterio adecuado para determinar el tipo de proceso subyacente a un caso de *mindreading*. En mi opinión, la propuesta de Keysers & Gazzola propone un criterio claro. Además, ésta tiene la ventaja de sugerir un modo en que los procesos espejo pueden subyacer a *mindreading*, proporcionando, así; una respuesta a una de las dificultades que presenta la propuesta de Goldman (2006). Sin embargo, tiene un costo para el enfoque de la TS. En esta propuesta, la simulación se asocia solamente a los procesos espejo y a los estados mentales a los que estos pueden subyacer (emociones, sensaciones e

intenciones motoras simples). Para que *mindreading* basado en procesos espejo tenga lugar es necesario, además, que intervengan proceso cognitivos superiores que traduzcan las representaciones primarias de los procesos espejo, y posibiliten la conceptualización y la verbalización.

De este modo, el criterio para distinguir entre teoría y simulación sugerido por Keysers & Gazzola (2006, 2007) implica una división de tareas entre los procesos. La simulación, implementada por los proceso espejo, va a subyacer a atribuciones de emociones, sensaciones e intenciones (motoras simples). El resto de las atribuciones se lleva a cabo por procesos reflexivos, ricos en información. Si bien la propuesta híbrida de teoría y simulación postula dos procesos subyacentes a una competencia cognitiva (*mindreading*), ésta no puede equiparase con las teorías de doble proceso que usualmente se postulan en ciencias cognitivas. La división de tareas implica, a mi entender, que ante el daño de alguno de los procesos, la capacidad para realizar las atribuciones que se asocian específicamente al proceso dañado se pierde totalmente. En este sentido, no cumple con uno de los requisitos de las teorías multiproceso (Machery 2009) y no concuerda con lo que típicamente sostienen los "teóricos de doble proceso" del razonamiento (Evans & Stanovich 2013), esto es, ante el daño de uno de los sistemas, la capacidad se conserva aunque con un perfil de desempeño diferente.

CONCLUSIONES

El fenómeno de atribuir estados mentales, y de explicar y predecir la conducta y el pensamiento propio y ajeno, en base a éstos, desde siempre ha llamado la atención de los filósofos y, más recientemente, de los psicólogos. La cuestión central es cómo podemos entender y predecir, sin esfuerzo, el comportamiento de las otras personas en la mayoría de las situaciones cotidianas. En este sentido, se ha indagado sobre los procesos que subyacen a *mindreading*. Desde la filosofía, el enfrentamiento entre los enfoques de TT y TS ha constituido el modo tradicional de abordar este fenómeno. No obstante, no ha sido posible decidir entre los mismos desde una perspectiva teórica ni desde una perspectiva empírica. Esta dificultad, junto con la creciente opinión respecto de que *mindreading* es un fenómeno complejo al que no le subyace un único proceso o mecanismo, condujo a un cambio. En la literatura sobre *mindreading* se puede observar la tendencia a abandonar posturas puras de teoría y simulación, incluso entre quienes las sostuvieron alguna vez. En este sentido, se asume que las cuestiones que no pueden ser abordadas por las teorías puras, pueden serlo por enfoques mixtos de teoría y simulación. En esta tesis me he ocupado de propuestas teóricas de los procesos subyacentes a *mindreading* que postulan enfoques híbridos de teoría y simulación (Nichols & Stich 2003; Goldman 2006).

Todo enfoque de *mindreading* debe poder dar cuenta tanto de las atribuciones de estados mentales a otras personas así como de la autoatribución, por tratarse de dos aspectos de una misma capacidad: la atribución mentalista. En este sentido, propuse unos requisitos fundamentales o de mínima para la hetero-atribución (HA1-HA4) y la autoatribución (AA1, AA2) que toda teoría de *mindreading* debe satisfacer. En base a estos criterios evalúe enfoques híbridos de teoría y simulación de *mindreading* mencionados, y analicé los alcances y límites explicativos de las mismos. A continuación retomaré los aspectos cruciales del análisis en base a estos requisitos que me llevaron a proponer que resulta esencial que un enfoque híbrido proporcione un

criterio para distinguir entre los procesos de teoría y simulación subyacentes a *mindreading*.

En relación a *mindreading* de tercera persona, las propuestas analizadas (Nichols & Sitch 2003; Goldman 2006) satisfacen el criterio (HA1) de manera trivial. De los requisitos, relacionados con el desarrollo, (HA2) y (HA3), surgen algunas limitaciones explicativas. Ambas propuestas dan cuenta del (HA2) salto significativo en el desarrollo asociado al desempeño exitoso en la TFC, observado en los niños entre los 3 y 4 años. Según Nichols & Stich (2003), este salto significativo en el desempeño en la TFC se relaciona con la disponibilidad de la caja de mundos posibles (CMP), y del actualizador, para *mindreading*. Así, tiene lugar el paso del sistema temprano al sistema tardío de *mindreading*. A partir de este momento, los niños pueden construir un modelo de las creencias del blanco, y esto les permite contemplar la posibilidad de creencias discrepantes. Al construir un modelo más acabado del blanco se generan predicciones óptimas sobre el comportamiento ajeno. Según Goldman (2006), la simulación tiene una explicación simple de este salto. Dado un escenario de TFC, el *mindreader* debe simular al blanco a partir de una creencia ficticia que contradice lo que él mismo sabe. Para predecir la creencia del blanco, el *mindreader* debe usar su estado mental ficticio (la creencia que tiene *Sally* acerca de la ubicación del objeto) y no su estado mental genuino (la ubicación actual de objeto). De modo que, el simulador debe poner en cuarentena o inhibir su creencia genuina para que no contamine la simulación. El cambio en el desempeño infantil en la TFC se relaciona con un cambio en la capacidad para poner en cuarentena los estados mentales genuinos. En otras palabras, el cambio en el desarrollo se explica como la adquisición de las funciones ejecutivas. En particular, la adquisición del control inhibitorio permite inhibir los estados mentales genuinos o "la tendencia a la realidad", como la llaman los psicólogos.

Sin embargo, nótese que la asociación entre *mindreading* y funciones ejecutivas no es privativa de la simulación. La propuesta modular/nativista del MTdM/PS, presentada en el capítulo 1 (sección 3.3), postula un mecanismo de "Teoría de la

Mente" que se complementa con un procesador de selección (Leslie & Scholl 1999; Scholl & Leslie 2001). Éste último resulta un proceso ejecutivo general requerido para inhibir la respuesta saliente, pero indeseada. En este sentido, consiste en la función ejecutiva de control inhibitorio. En la propuesta modular, el procesador de selección se considera necesario porque el MTdM atribuye automáticamente creencias verdaderas y, en la TFC estándar, es preciso inhibir la respuesta privilegiada por el MTdM. La misma está asociada a la ubicación actual y saliente del objeto. El procesador de selección permite captar el contenido de la creencia de *Sally* respecto de la ubicación del objeto[77]. En este breve panorama se puede apreciar que los hallazgos en el desarrollo, *i.e.* el rol de los procesos ejecutivos en *mindreading*, pueden ser compatibles con distintos enfoques, ya sea puros de teoría (enfoque nativista) o de simulación (Goldman 2006), ya sea por enfoques híbridos de teoría y simulación (Nichols & Stich 2003)[78]. Es más, las propuestas de Leslie & Scholl (1999; Scholl & Leslie 2001) y Goldman (2006) recurren al mismo cuerpo de evidencia respecto de la relación entre *mindreading* y funciones ejecutivas para fundamentar sus propuestas[79].

Según el requisito (HA4) las personas llevan a cabo juicios de *mindreading* proyectando sus propias sensaciones, creencias y conocimientos. En ambas propuestas parece haber acuerdo respecto de que el recurso a los estados mentales propios en

[77] Como mencioné, si bien hay evidencia que documenta la relación entre *mindreading* y el control inhibitorio (Carlson & Moses 2001), no es clara la naturaleza de la misma. Al parecer los niños de 3 años se desempeñar sin éxito en una TFC modificada de modo tal que las demandas de control inhibitorio resultan casi nulas (Call & Tomasello 1999). Esto sugiere, primero, que no es claro qué aspecto de la TFC requiere de control inhibitorio y, segundo, que la relación entre *mindreading* y control inhibitorio no puede ser toda la historia sobre el salto en el desarrollo. Se ha sugerido que la relación entre *mindreading* y funciones ejecutivas tiene que ver con las demandas incidentales de la TFC. En conclusión, la evidencia sugiere una relación entre las competencias pero no es clara la naturaleza de la misma.

[78] Recuérdese que la CMP y el actualizador son considerados procesos simulacionales pero la predicción propiamente dicha de lleva a cabo por inferencia.

[79] La cuestión de que la comprensión de los deseos precede a la comprensión de las creencias (el requisito HA3) no se relaciona directamente con la propuesta de la necesidad de un criterio para distinguir entre procesos subyacentes. No obstante, muestra las limitaciones de ambos enfoques. Según Nichols & Stich (2003) los mecanismos detectores de deseos, que forman parte del sistema de atribución en base a metas, están presentes tempranamente. En este sentido dan cuenta de que la comprensión de los deseos es previa a la de las creencias. No obstante, considero que esta propuesta no puede dar cuenta de que los niños advierten tempranamente la variabilidad de las creencias. Así, el requisito (HA3) se satisface parcialmente. En el enfoque de Goldman no hay elementos que den cuenta de esta característica de la comprensión temprana, y el requisito (HA3) no se satisface.

mindreading de tercera persona sugiere que el proceso subyacente es de tipo simulacional. Esto está en concordancia con la propuesta de que la TS tiene un caso a favor en la predicción de inferencias (Harris 1995, Nichols & Stich 2003; Stich & Nichols 2003; Goldman 2006; Apperly 2008). En este sentido, se podría considera el caso a favor de la simulación como un posible criterio para determinar el proceso que subyace a un caso de *mindreading*. De modo que la simulación subyace a la predicción de inferencias, mientras que la teoría subyace al resto de *mindreading*. Sin embargo, esto no puede considerarse un criterio satisfactorio porque su alcance es acotado. Sólo indica el proceso subyacente para un solo caso de *mindreading*, la predicción de inferencias. Otro caso que se considera a favor de la simulación es el recurso de atribución por *default* de las creencias propias a la otra persona. Si bien Nichols & Stich (2003) afirman que este recurso es central, ya que está involucrado en la construcción de un modelo de las creencias del blanco en la CMP del *mindreader*, no es claro cuán a menudo se utiliza el mismo. La evidencia sugiere el recurso de la atribución de los estados mentales propios a otras personas en la medida en que solemos cometer sesgos egocéntricos en *mindreading*. Sin embargo, el recurso de apelar a las propias creencias, conocimientos, sensaciones parece tener lugar en casos particulares de *mindreading*. Por ejemplo, cuando hay que razonar con creencias en condiciones de incertidumbre (Nickerson 1999; Royzman, Cassidy & Baron 2003) o juzgar la probabilidad con la que alguien tiene una creencia (Birch & Bloom 2007). De modo que un criterio basado en el caso a favor de la simulación no parece tener un alcance suficiente, y aún hace falta un criterio más abarcativo para determinar el proceso que subyace a un caso de *mindreading*.

En virtud de estas limitaciones explicativas considero que es necesario un criterio para distinguir entre los procesos subyacentes. A mi entender, los enfoques híbridos de teoría y simulación para *mindreading* de tercera persona no pueden sostener simplemente que en algunas ocasiones subyace un proceso y en otras ocasiones otro. Esto no permite realizar las predicciones adecuadas. En este sentido, considero que la adecuación de las propuestas híbridas depende de que puedan

brindar un criterio claro para poder distinguir el tipo de proceso subyacente a cierta instancia de *mindreading*. Los enfoques híbridos que analicé (Nichols & Stich 2003; Goldman 2006) tienen dificultades para brindar un criterio que permita distinguir entre procesos subyacentes de teoría o simulación.

Nichols & Stich (2003) y Stich & Nichols (2003) proponen un criterio de corrección para determinar cuándo subyace un proceso de teoría o de simulación a *mindreading*. Si el resultado de *mindreading* resulta correcto es probable que el proceso subyacente sea de tipo simulacional, si el resultado de *mindreading* resulta incorrecto es probable que el proceso subyacente sea de tipo rico en información. Este criterio falla. En contra del primer aspecto, sostuve que ninguno de los argumentos esgrimidos para sostener que el resultado de *mindreading* exitoso es producto de la simulación logra descartar la posibilidad de que un proceso rico en información subyazca a *mindreading* con un resultado correcto. Al no descartarse las explicaciones de procesos ricos en información, no se ve por qué la explicación simulacional es preferible. En contra del segundo aspecto del criterio, sostuve que es concebible que un proceso de tipo simulacional subyazca a *mindreading* con un resultado incorrecto. Si al llevar a cabo *mindreading* mediante la simulación, se recluta el sistema de toma de decisiones pero no los sesgos que operan sobre el mismo, el *output* de *mindreading* no incluirá el aporte de los mismos. En consecuencia, no habrá concordancia entre el *output* de *mindreading* y el *output* de la competencia cognitiva del blanco, el sistema de toma de decisiones que está sometido a los sesgos. Esto generará una predicción errónea sobre el comportamiento del blanco. En este sentido, es posible que un proceso simulacional pueda generar *mindreading* con un resultado incorrecto, tal como esto se tener lugar mediante inferencias conducidas por cuerpos de conocimiento psicológico erróneos o parcialmente incompletos. En la medida en que ninguno de los dos aspectos del criterio de corrección resulta satisfactorio, concluyo que este criterio no ofrece un modo de distinguir entre los tipos de procesos subyacentes a *mindreading*.

En el caso de Goldman, la dificultad reside en que no se propone criterio alguno. En este enfoque los roles para los procesos de teoría y simulación apenas se establecen estipulando los distintos modos en que pueden combinarse para generar *mindreading*. Goldman considera que las relaciones más plausibles entre teoría y simulación son la independencia y la cooperación (Goldman 2006: 41-46). La independencia consiste en que algunos casos de *mindreading* se llevan a cabo completamente por teoría y otros completamente por simulación. Esta relación está poco descripta. En principio, es posible pensar que ambos procesos, en tanto independientes, son suficientes para producir *mindreading* por separado. Sin embargo, no se especifican las condiciones según las cuales actúan los procesos de teoría y simulación separadamente. No se sabe si cada proceso se pone en funcionamiento en casos particulares, o en un rango de casos o en todos los casos. Dada la falta de especificación, quedan abiertas todas las posibilidades. Además, la relación de independencia propuesta deja abierta la posibilidad del funcionamiento simultáneo de los procesos de teoría y simulación. Si esto es posible, surgen otras cuestiones tales como qué mecanismo determina el *output* de *mindreading* cuando los procesos funcionan en simultáneo.

En cuanto a la cooperación, se asume, más bien, que la teoría ayuda a establecer las condiciones para que un proceso simulacional se inicie. Esto se aprecia en los ejemplos de cooperación propuestos por Goldman (2006). Primero, es posible recurrir a teorización para inferir los estados previos del blanco para los que, luego, se generan estados mentales ficticios, que alimentan como *input* al sistema de toma de decisiones propio. Éste, que funciona en modo *off-line*, arroja un *output* que se usa para realizar una atribución al blanco. Nótese que si bien el primer paso se lleva a cabo por teorización, el proceso mediante el cual se genera la atribución es de tipo simulacional. En la relación de cooperación, la teoría no es suficiente para producir el *output* de *mindreading* sino, más bien, viene a establecer las condiciones iniciales para una simulación.

Además, quisiera hacer hincapié en que tanto la TT, como la TS y sus híbridos, asumen que los estados previos de blanco se pueden determinar fácilmente. La posibilidad de determinar estos estados es esencial en los procesos ricos en información y en la simulación. Respecto de un proceso rico en información constituye uno de los elementos que intervienen en las inferencias que generan las predicciones del comportamiento. En el caso de la rutina simulacional, estos son esenciales para generar los estados mentales ficticios en el *mindreader* que alimentarán el sistema cognitivo reclutado en la simulación. Sin embargo, esto se da por sentado y quizás no sea una cuestión tan simple determinar tales estados previos, sino algo más bien complicado. Incluso, se puede tratar de una subcapacidad de la capacidad de *mindreading*.

Claramente, en la propuesta de Goldman (2006) la teoría y la simulación se conciben como sub-competencias que colaboran para llevar a cabo *mindreading*. No obstante, esta propuesta no brinda un criterio satisfactorio que permita establecer qué tipo de proceso subyace a un caso de *mindreading*, en general, por la falta de descripción de la interacción entre teoría y simulación y, en particular, por la falta de descripción del rol de la teoría. En este sentido, considero que esta propuesta no cumple con el requisito de un criterio para determinar el tipo de proceso subyacente para los enfoques híbridos de teoría y simulación. Más bien, resulta una propuesta de simulación refinada.

Dada esta situación, intenté ofrecer un criterio para distinguir entre teoría y simulación atendiendo a los hallazgos empíricos, particularmente, en neurociencias. A partir de un criterio que se desprende de la propuesta de integración de teoría y simulación de Keysers & Gazzola (2006) se pueden articular los enfoques híbridos de teoría y simulación. Esto implica, sin embargo, dividir las funciones de *mindreading* entre los procesos subyacentes. La simulación subyace a las atribuciones de emociones, sensaciones e intenciones motoras simples, para las que hay procesos espejo, mientras que la teoría subyace al resto de *mindreading*. La ventaja de este criterio es que es más abarcativo que, por ejemplo, el criterio que surge de afirmar que

la simulación subyace a la predicción de inferencias. Otra ventaja reside en que propone un modo en que los procesos espejo podrían subyacer a *mindreading*, dando cuenta de una de las dificultades de la propuesta simulacional de Goldman (2006). Para que el entendimiento de las otras personas provisto por los procesos espejo se vuelva una atribución explícita, las representaciones primarias generadas por los procesos espejo tienen que transformase en representaciones secundarias con la mediación de los procesos cognitivos superiores asociados a la conceptualización y la verbalización. La desventaja es que esta división de funciones es costosa para la simulación. En primer lugar, la simulación se identifica con los procesos espejo, y se excluyen otros procesos como imaginación enactiva. En segundo lugar, el rango de atribuciones se reduce a las atribuciones mencionadas, cediendo el resto de *mindreading* a un proceso subyacente de tipo rico en información. Particularmente, la simulación no subyace a la las atribuciones de actitudes proposicionales.

Además, la división de funciones que implica el criterio de Keysers & Gazzola (2006, 2007) tiene otra consecuencia. Si se dañara alguno de los procesos subyacentes a *mindreading* ¿qué perfil de desempeño surgiría de esto? Por ejemplo, si se dañan los procesos espejo que subyacen a la emoción del asco, dada la división de funciones ¿es posible atribuir asco? Al parecer ya no se podría atribuir asco en ningún sentido, sino sólo el resto de los estados mentales. Este perfil de desempeño conlleva que la propuesta de dos procesos subyacentes a *mindreading* no puede considerarse un sistema de doble proceso, tal como usualmente se postulan en ciencias cognitivas. Según los "teóricos del doble proceso", si alguna ruta se avería, aún queda la otra vía para realizar atribuciones del mismo tipo. Por ejemplo, si se daña el sistema racional para realizar juicios morales, aún queda el sistema de tipo emocional (Greene & Haidt 2002). Cambia el perfil de los juicios morales, en tanto que todos estarán originados en el proceso de tipo emocional subyacente, pero aún se llevarán a cabo juicios morales. En el caso del asco, ya no se puede atribuir asco en ningún sentido. Salvo que se adhiera a la idea de que la atribución emocional se lleva a cabo por varios caminos y hay otro camino además de los procesos espejo.

En este caso, ya no se trataría de un sistema de doble proceso, porque varios procesos subyacen a la atribución emocional[80]. De modo que el perfil de desempeño que surge ante el daño potencial de teoría o simulación, según la propuesta basada en el criterio de Keysers & Gazzola (2003), no es el tipo de perfil de desempeño que tienen en mente los "teóricos de doble proceso". En este sentido, si bien dos procesos subyacen a *mindreading*, no implican una teoría de doble proceso tal como se entiende usualmente en ciencias cognitivas.

Es preciso mencionar que la propuesta anterior sólo abarca a *mindreading* de tercera persona y, en este sentido, no resulta una propuesta adecuada de *mindreading*. En relación a la autoatribución, las propuestas teóricas analizadas postulan un proceso dual de introspección e interpretación. En el caso de Nichols & Stich (2003), el criterio para determinar qué tipo de proceso subyace a una autoatribución no es satisfactorio. La introspección es el proceso principal y la interpretación es excepcional. Según los autores, el proceso interpretativo sólo subyace a los casos en los que se les pregunta a las personas por las causas de su comportamiento. En concordancia con la afirmación de los "teóricos de la introspección" de que no se puede acceder directamente a las causas del comportamiento. De modo que cuando se nos pregunta por la causas de las acciones no nos queda más opción que recurrir a la interpretación y confabular. No obstante, la evidencia sugiere que la confabulación surge también en otros caso, contra Stich & Nichols (2003). De modo que el criterio según el cual la interpretación sólo subyace a la pregunta por las causas del comportamiento y la introspección al resto de los casos autoatribución, no es satisfactorio según la evidencia. En el caso de Goldman (2006), no se propone ningún criterio aunque, en principio, en tanto teórico de la

[80] Se ha sugerido que la atribución emocional se lleva a cabo por varios caminos, algunos primitivos como los procesos espejo o de resonancia y otros más cognitivos (o de cognición superior). Se podría pensar que estos caminos de cognición superior son los procesos que según Keysers & Gazzola (2006) complementan a los procesos espejo. No obstante, justamente porque los procesos cognitivos superiores son traductores de las representaciones de los procesos espejo no concuerdan con la idea mencionada. Se trata de vías paralelas de procesamiento o independientes que son completamente suficientes para atribuir emociones, como el procesamiento de la expresión facial en las regiones corticales visuales que conduce a la representación proposicional del estado emocional inferido (Wicker *et al*. 2003). De modo que no hay ninguna representación de procesos espejo que haya que traducir.

introspección estaría de acuerdo con el criterio propuesto por Nichols & Stich (2003)para la interpretación y la introspección.

En relación a la introspección me gustaría señalar que resulta una cuestión relevante establecer la plausibilidad de un mecanismo o proceso cognitivo de acceso directo a los estados mentales así como establecer qué tipo de evidencia permitiría sostener esto. La presencia de procesos de tipo interpretativos subyacentes a la autoatribución está suficientemente documentada en la evidencia, en especial, en relación a la confabulación. Se ha sugerido que los estudios que indagan sobre las causas del comportamiento así como también los que requieren que los sujetos reporten sus actitudes con una posterioridad significativa no resultan adecuados para estudiar la introspección. Si los datos implican reportes introspectivos de eventos que ocurrieron con mucha anterioridad será preciso que los sujetos rememoren. En estos casos, los sujetos no tendrán más remedio que interpretar el comportamiento recordado en la ocasión. En este sentido, tal como han sugerido Engelbert & Carruthers (2010), parecen adecuados los estudios en los que los sujetos tienen que brindar reportes introspectivos luego de en un lapso de tiempo breve posterior al evento.

REFERENCIAS

Adophs, R., Tranel, D., Damasio, H. & Damasio, A. (1994) "Impaired Recognition of Emotion in Facial Expressions Following Bilateral Damage to the Amygdala", *Nature*, 372, 669-672.

Apperly, I. (2008) "Beyond Simulation-Theory and Theory-Theory: Why social cognitive neuroscience should use its own concepts to study 'theory of mind'", *Cognition, 107*, 266-283.

Apperly, I. (2011) *Mindreaders: The Cognitive Basis of "Theory of Mind",* New York, Psychology Press.

Apperly, I. & Butterfill, S. (2009) "Do Humans Have Two Systems to Track Beliefs and Belief-Like States?", *Psychological Review*, 116 (4), 953-970.

Apperly, I. A., Carroll, D. J., Samson, D., Qureshi, A., Humphreys, G. W., & Moffatt, G. (2010) "Why Are There Limits on Theory of Mind Use? Evidence from Adults' Ability to Follow Instructions from an Ignorant Speaker", *Quarterly Journal of Experimental Psychology*.

Aydede, M. (2010) "The Language of Thought Hypothesis", *The Stanford Encyclopedia of Philosophy* (Fall 2010 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2010/entries/language-thought/>.

Baars, B. J. (1988) *A Cognitive Theory of Consciousness*, Cambridge, Cambridge University Press.

Baars, B. J. (1997) *In the Theatre of Consciousness*, Oxford, Oxford University Press.

Baillargeon, R., He, Z., Setoh, P., Scott, R. M., Sloane, S., & Yang, D. Y.-J. (2013) "False-Belief Understanding and Why It Matters: The Social-Acting Hypothesis", en M. R. Banaji & S. A. Gelman (Eds.), *Navigating the Social World: What Infants, Children, and Other Species Can Teach Us*, NY, Oxford University Press.

Baron-Cohen, S. (1995) *Mindblidness: An Essay on Autism and Theory of Mind*, Cambridge, MIT Press.

Baron-Cohen, S. (1999) "The Extreme Male Brain Theory of Autism", en Tager-Flusberg H. (ed) *Neurodevelopmental Disorders,* Cambridge, MIT Press.

Baron-Cohen, S. (2003) *The essencial Diference: the Truth about the Male and Female Brain*, New York, Basci Books.

Baron-Cohen, S., Leslie, A. & Frith, U. (1985) "Does the Autistic Child Have a 'Theory Of Mind'?", *Cognition*, 21, 37-46.

Baron-Cohen, S., Leslie, A. & Frith, U. (1986) "Mechanical, Behavioral, and Intentional Understanding of Picture Stories in Autistic Children", *British Journal of Developmental Psychology*, 4, 283-286.

Bartsch & Wellman (1995) *Children Talk about the Mind*, New York, Oxford University Press.

Bennett, J. (1978) "Some Remarks about Concepts", *Behavioral and Brain Sciences*, 1, 557-560.

Bermúdez, J. (2005a) *Philosophy of Psychology: A contemporary Introduction*, New York, Routeldge.

Bermúdez, J. (2005b) “Arguing for eliminativism”, en Keeley, B. (ed.) *Paul Churchland*, Cambridge, Cambridge University Press.

Birch, S. A. J., & Bloom, P. (2007) “The Curse of Knowledge in Reasoning about False Beliefs”, *Psychological Science*, 18, (5), 382-386.

Blair, R. J. R. (2002) “A Neuro-Cognitive Model of the Psychopathic Individual”, en M. Ron (ed.) *Disorders of Brain and Mind II*, Cambirdge, Cambridge University Press.

Blair, R.J.R.; Mitchell, D. G. V. Peschardt, K. S., Colledge, E., Leonard, R. A., Shine, J H., Murray, L. K., and Perrett, D. I. (2004) “Reduced Sensitivity to Others’ Fearful Expressions in Psychopatic Individuals”, *Personality and Individual Differences*, 37, 1111-1122.

Blakemore, S.J., Winston, J. & Frith, U. (2004) “Social Cognitive Neuroscience: Where Are We Heading?”, *Trends in Cognitive Sciences*, 8, 216-222.

Block, N. (1994) “Functionalism (2)”, en Guttenplan, S. (ed.) *A Companion to the Philosophy of Mind*, Oxford, Blackwell.

Bloom, P., & German, T. P. (2000) “Two Reasons to Abandon the False Belief Task as A Test of Theory of Mind”, *Cognition*, 77, B25-B31.

Botterill, G. & Carruthers, P. (1999/2003) *The philosophy of Psychology*. Cambridge, Cambridge University Press.

Brasil-Neto, J., Pascual-Leone, A., Valls-Sole, J., Cohen, L. & Hallett, M. (1992) “Focal transcranial magnetic stimulation and response bias in a forced choice task”, *Journal of Neurology, Neurosurgery, and Psychiatry*, 55, 964–66.

Briñol, P. & Petty, R. (2003) “Overt head movements and persuasion: A self-validation analysis”, *Journal of Personality and Social Psychology*, 84, 1123–39.

Buccino, G., Binkofski, F., Fink, G.R., Fadiga, L., Fogassi, L., Gallese, V., Seitz, R.J., Zilles, K., Rizzolatti, G. & Freund, H.J. (2001) "Action Observation Activates Premotor and Parietal Areas in Somatotopic Manner: An fMRI Study", *European Journal of Neuroscience*, 13, 400-404.

Burge, T. (1979) “Indivisualism and The Mental”, *Midwest Studies in Philosophy*, 4, 73-121.

Call, J., & Tomasello, M. (1999) “A Nonverbal False Belief Task: The Performance of Children and Great Apes”, *Child Development*, 70, 381-395.

Call, J., & Tomasello, M. (2008) “Does the Chimpanzee Have a Theory of Mind? 30 Years Later”, *Trends in Cognitive Sciences*, 12(5), 187-192.

Camerer, C. Loewenstein, G. & Weber, M. (1989) “The Curse of Knowledge in Economic Settings: An Experimental Analysis”, *The Journal of Political Economy*, 97, 1232-1254.

Carlson, S.M. & Moses, L.J. (2001) “Individual Differences in Inhibitory Control and Children’s Theory of Mind”, *Child Development*, 72, 1032-1053.

Carrington, S. J., & Bailey, A. J. (2009) “Are There Theory of Mind Regions in The Brain? A Review of the Neuroimaging Literature”, *Human Brain Mapping*, 30, (8), 2313-2335.

Carruthers, P. (1996) "Simulation and Self-Knowledge: A defense of the Theory-Theory" en Carruthers, P. & Smith, P. (eds.) *Theories of Theories of Mind*, Cambridge, Cambridge University Press.

Carruthers, P. (2006) *The Architecture of the Mind*, Oxford, Oxford University Press.

Carruthers, P. (2009) "How We Know Our Own Minds: The Relationship between Mindreading and Metacognition", *Behavioral & Brain Science*, 32, 121-182.

Carruthers, P. (2013) "Mindreading the Self" en Baron-Cohen, S., Tager-Flusgsberg, H. & Lombardo, M. (eds.) *Understanding Other Minds*, Oxford, Oxford University Press.

Carruthers, P. (2011) *The Opacity of Mind: An Integrative Theory of Self-Knowledge*, Oxford, Oxford University Press.

Carruthers, P. & Smith, P. (1996) *Theories of Theories of Mind*, Cambridge, Cambridge University Press.

Chomsky, N. (1965) *Aspects of the Theory of Syntax*, Cambridge, MIT Press.

Churchland, P.M. (1981) "Eliminative Materialism and Propositional Attitudes", *Journal of Philosophy*, 78, 67-90.

Churchland, P.M. (1988) *Matter and Consciousness, Revised Edition: A Contemporary Introduction to the Philosophy Of Mind*, MIT Press.

Crane, T. (1990) "The Language of Thought", *Mind and Language*, 5 (3), 187-212.

Cummins, R. (1983) *The Nature of Psychological Explanation*, Cambirdge, Mass., MIT Press.

Currie, G. & Ravenscroft, I (2002) *Recreative Minds: Imagination in Philosophy and Psychology*, Oxford, Oxford University Press.

Damasio, A. (1994) *El error de Descartes: La razón de las emociones*. Barcelona, Andrés Bello Editor.

Davies, M. (1989) "Tacit Knowledge and Sudoxastic States", en George, A. (comp.) *Reflections on Chomsky*, Blackwell, Oxford.

Davies, M. & Stone, T. (1995a) *Folk Psychology: The Theory of Mind Debate*, Oxford, Blackwell.

Davies, M. & Stone, T. (1995b) *Mental Simulations. Evaluations and Applications*, Oxford, Blackwell.

Davies, M. & Stone, T. (1998) "Folk Psychology and Mental Simulation", en O´Hear, A. (ed.) *Current Issues in Philosophy of Mind*, Cambridge, Cambridge University Press, 53-82.

Dennett, D. (1978) "Belief about Beliefs", *Behavioral and Brain Sciences*, 1, 568-570.

Dennet, D. (1987) *The Intentional Stance*, MA, MIT Press.

Descartes, R. (1641) *Meditaciones metafísicas,* Alfaguara, Madrid, (1993), trad. de Peña García, Vidal.

De Vignemont, F. (2009) "Drawing the Boundary between Low-Level and High-Level Mindreading", *Philosophical Studies* 144, 457-466.

Engelbert, M. & Carruthers, P. (2010) "Introspection", *WIREs Cognitive Science*, 1, 245-253.

Evans, G. (1982) *The Varieties of Reference*, Oxford, Clarendon Press.

Evans, J. St. & Stanovich, K. E. (2013) “Dual-Process Theories of Higher Cognition: Advancing the Debate”, *Perspectives on Psychological Science*, 8 (3), 223-241.

Ekman, P. (1985*) Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage*, NY, W.W. Norton & co.

Fadiga, L., Fogassi, L., Pavesi, G. & Rizzolatti, G. (1995) "Motor Facilitation during Action Observation: A Magnetic Stimulation Study", *Journal of Neurophysiology*, 73, 2608-2611.

Ferrari, P.F., Gallese, V., Rizzolatti, G. & Fogassi, L. (2003) "Mirror Neurons Responding to the Observation of Ingestive and Communicative Mouth Actions in the Monkey Ventral Premotor Cortex", *European Journal Of Neuroscience*, 17, 1703-1714.

Flavell, J. H. (1988) The Development of Children’s Knowledge about the Mind: From Cognitive Connections to Mental Representations, en J. Astington, P. Harris & D. Olson (eds.) *Developing theories of Mind*, NY, Cambridge University Press, 244-267.

Flavell, J. H., Flavell, E. R. & Green, F. L. (1987) “Young Children’s Knowledge about the Apparent-Real and Pretend-Real Distinctions”, *Developmental Psychology, 23*, 99-103.

Fletcher, P.C., Happé, F., Frith, U., Baker, S.C., Dolan, R.J., Frackowiak, R.S.J. & Frith, C. D. (1995) “Other Minds in the Brain: A Functional Imaging Study of ‘Theory of Mind’ in Story Comprehension”, *Cognition*, 57, 109-128.

Fodor, J. (1975) *The Language Of Thought*, Thomas Y. Crowell Co., Harvard University Press.

Fodor, J. (1983) *Modularity of mind*, Cambridge, MIT Press.

Fodor, J. (1987) *Psychosemantics*, Cambridge, MIT Press.

Fodor, J. (1992) “A Theory of the Child’s Theory of Mind”, *Cognition*, 44, 283-296.

Fodor, J. (1998) *Concepts Where Cognitive Science Went Wrong*, Oxford, Clarendon Press.

Fodor, J. (2000) *The Mind Doesn’t Work that Way: The Scope and Limits of Computational Psychology*, Cambridge, MIT Press.

Frith, C. D. & Frith, U. (1999) “Interacting Minds: A Biological Basis”, *Science*, 286, 1692-95.

Frith, U. & Frith, C.D. (2003) “Development and Neurophysiology of Mentalizing”, en C.D. Frith & D. Wolpert (eds.), *The Neuroscience of Social Interaction*, Oxford, Oxford University Press, 45-75.

Frith, C. D. (2007) “The social brain?”, *Philosophical Transactions of the Royal Society B-Biological Sciences*, 362, (1480), 671-678.

Gallagher, S. (2001) “The Practice of Mind: Theory, Simulation or Interaction?”, *Journal of Consciousness Studies,* 8, *83-107.*

Gallagher, S. (2006) *How the Body Shapes the Mind*, Oxford, Claredon Press.

Gallagher, S. & Hutto, D. (2007) “Understanding Others Through Primary Interaction and Narrative Practice”, en Zlatev, J. *et al*. (eds.), *The shared mind: Perspectives in intersubjetivity*, Amsterdam, John Benjamin Publishing Company, 17-38.

Gallagher, H. L., Happe, F., Brunswick, N., Fletcher, P. C., Frith, U., & Frith, C.D. (2000)"Reading the Mind in Cartoons and Stories: An fMRI Study of 'Theory Of Mind' in Verbal and Nonverbal Tasks, *Neuropsychologia*, 38, 11-21.

Gallagher, H. L., & Frith, C. D. (2003) "Functional Imaging of 'Theory Of Mind'", *Trends in Cognitive Science*, 7, 77-83.

Gallese, V. Fadiga, L., Fogassi, L. & Rizzolatti, G. (1996) "Action Recognition in the Premotor Cortex", *Brain*, 119, 593-609.

Gallese, V. & Goldman, A. (1998) "Mirror Neurons and the Simulation Theory of Mind-Reading", *Trends in Cognitive Sciences*, 3, 493-501.

Gallese, V., Fogassi, L., Fadiga, L. & Rizzolatti, G. (2002) "Action Representation and The Inferior Parietal Lobule", en Prinz, W. & Hommel, B. (eds.) *Attention and Performance XIX*, Oxford, Oxford University Press, 247-266.

Gazzaniga, M. (1995) "Consciousness and the Cerebral Hemispheres. In The Cognitive Neurosciences", en Gazzaniga, M. (ed) *The New Cognitive Neurosciences*, MIT Press, 1391–1400.

Gazzaniga, M. & Baynes, K. (2000) "Consciousness, Introspection,and the Split-Brain:The Two Minds/One Body Problem", en Gazzaniga, M. (ed)*The New Cognitive Neurosciences*, Segunda Edición, 1355-1364.

Gegerly, G., Nadasdy, Z., Csibra, G. & Biro, S. (1995) "Taking the Intentional Stance at 12-month-of age", *Cognition*, 56, 165-193.

Grezes, J., Frith, C.D. & Passingham, R.E. (2004) "Inferring False Beliefs from the Actions of Oneself and Others: An fMRI Study", *NeuroImage*, 21, 744-750.

Gilbert, S. J., Frith, C. D., & Burgess, P. W. (2005) **"**Involvement of Rostral Prefrontal Cortex in Selection Between Stimulus-Oriented and Stimulus-Independent Thought", *European Journal of Neuroscience*, 21, 1423-1431.

Goel, V., Grafman, J., Sadato, N., & Hallett, M. (1995) "Modeling Other Minds", *Neuroreport*, 6, 1741-1746.

Goldman, A. (1993a) "The Psychology of Folk Psychology", *Behavioral and Brain Sciences*, 16, 15-28.

Goldman, A. (1993b) "Consciousness, Folk Psychology, and Cognitive Science", *Consciousness and Cognition*, 2, 364-382.

Goldman, A. (1995a) "Interpretation Psychologized", en M. Davies & T. Stone (eds.) *Folk Psychology. The Theory of Mind Debate*, Oxford, Blackwell, 60-73.

Goldman, A. (1995b) "In Defense of the Simulation Theory", en M. Davies & T. Stone (eds.) *Folk Psychology. The Theory of Mind Debate*, Oxford, Blackwell, 191-206.

Goldman, A. (1995c) "Empathy, Mind, and Morals", en M. Davies & T. Stone (eds.) *Mental Simulations. Evaluations and Applications*, Oxford, Blackwell, 185-208.

Goldman, A. (2000) "The Mentalizing Folk", en Sperber, D. (ed.) *Metarepresentations: A multidisciplinary perspective*, Oxford, Oxford University Press, 171-196.

Goldman, A. (2006) *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*, Oxford, Oxford University Press.

Goldman, A. & de Vignemont, F. (2009) "Is Social Cognition Embodied?", *Trends in Cognitive Sciences*, 30 (10), 1-6.

Goldman, A. & Sripada, C.S. (2005) “Simulationist Models of Face-Based Emotion Recognition”, *Cognition*, 94, 193-213.

Gopnik, A. (1993) “How We Know Our Own Minds: The Illusion of First-Person Knowledge of Intentionality”, *Behavioral and Brain Sciences*, 16, 1-14.

Gopnik, A. & Astington, J.W. (1988) “Children's Understanding of Representational Change and Its Relation to the Understanding of False Belief and the Appearance-Reality Distinction”, Child Development, 59(1), 26-37.

Gopnik, A. & Wellman, H. (1992) “Why the Child’s Theory of Mind Really Is a Theory”, *Mind & Language*, 7 (1-2), 145-71.

Gopnik, A. & Meltzoff, A. (1994) “Minds, Bodies, and Persons: Young Children’s Understanding of the Self and Others as reflected in Imitation and Theory of Mind Research”, en S. Parker, R. Mitchell & M. Boccia (eds.) *Self- Awareness in animals and Humans*, NY, Cambridge University Press, 166-186.

Gopnik, A. & Wellman, H. (1994) “The Theory Theory”, en L. Hirschfeld & S. Gelman (eds.) *Mapping the Mind: Domain Specificity in Cognition and Culture*, New York, Cambridge University Press, 257-293.

Gopnik, A. & Meltzoff, A. (1997) *Words, Thoughts and Theories*, Cambridge, Cambridge University Press.

Gordon, R. (1995a) “Folk Psychology as Simulation”, en T. Stone & M. Davies (eds.) *Folk Psychology. The Theory of Mind Debate*, Oxford, Blackwell, 60-73.

Gordon, R. (1995b) “The Simulation Theory: Objections and Misconceptions”, en T. Stone & M. Davies (eds.) *Folk Psychology. The Theory of Mind Debate*, Oxford, Blackwell, 100-122.

Gordon, R. (1995c) “Reply to Stich and Nichols”, en T. Stone & M. Davies (eds.) *Folk Psychology. The Theory of Mind Debate*, Oxford, Blackwell, 174-184.

Gordon, R. (1995d) “Simulation Without Introspection or Inference from Me to You”, en T. Stone & M. Davies (eds.) *Mental Simulation: Evaluations and Applications*, Oxford, Blackwell, 53-67.

Gordon, R. (1995e) “Sympathy, Simulation, and the Impartial Spectator”, *Ethics*, 105, 727-742.

Gordon, R. (1996) “Radical Simulation”, en P. Carruthers & P. Smith (eds.) *Theories of theories of Mind*, Cambridge, Cambridge University Press, 11-21.

Gordon, R. (2007) “Ascent Routines for Propositional Attitudes”, Synthese, 159, (2), 151-165.

Gordon, R. (2009) “Folk Psychology as Mental Simulation”, en *The Stanford Encyclopedia of Philosophy (Fall 2009 Edition)*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2009/entries/folkpsych-simulation/>.

Grafton, S.T., Arbib, M.A., Fadiga, L. & Rizzolatti, G. (1996) "Localization of Grasp Representations in Humans by PET: II. Observation compared with Imagination", *Experimental Brain Research*, 112, 103-111.

Greene, J. D. & Haidt J. (2002) “How and Where does Moral Judgment Work?”, *Trends in Cognitive Science*, 6, 517-523.

Grezes, J., Frith, C.D. & Passingham, R.E. (2004) “Inferring False Beliefs from the Actions of Oneself and Others: An fMRI Study”, *NeuroImage*, 21, 744-750.

Gusnard, D. A., Akbudak, E., Shulman, G. L., & Raichle, M. E. (2001) “Medial Prefrontal Cortex and Self-Referential Mental Activity: Relation to A Default Mode of Brain Function”, *Proceedings of the National Academy of Sciences*, U.S.A., 98, 4259-4264.

Harman, G. (1978) “Studying the Chimpanzee’s Theory of Mind”, *Behavioral and Brain Sciences*, 1, 515-526.

Harris, P. (1995) “From Simulation to Folk Psychology: The Case for Development”, en M. Davies & T. Stone (eds.) *Folk Psyhcology. Readings in Mind & Language*., Oxford, Blackwell, 207-231.

Heal, J. (1995a) “Replication and Functionalism”, en M. Davies & T. Stone (eds.) *Folk Psychology. Readings in Mind & Language*, Oxford, Blackwell, 45-59.

Heal, J. (1995b) “How to Think about Thinking”, en T. Stone & M. Davies (eds.) *Mental Simulation: Evaluations and Applications*, Oxford, Blackwell, 33-52.

Heal J. (2003) *Mind, Reason and Imagination*, Cambridge, Cambridge University Press.

Hughes, C. (1998) “Finding Your Marbles: Does Preschoolers' Strategic Behavior Predict Later Understanding of Mind?”, *Developmental Psychology*, 34, 1326-1339.

Hurlburt, R. (1990) *Sampling normal and schizophrenic inner experience*, New York, Plenum Press.

Hurlburt, R. T. & Akhter, S. A. (2006) “The Descriptive Experience Sampling Method”, *Phenomenology and the Cognitive Sciences*, 5, 271–301.

Hurlburt, R. T. & Heavey, C. L. (2001) “Telling What We Know: Describing Inner Experience”, *Trends in Cognitive Sciences*, 5, 400-403.

Hutto, D. (2008) *Folk Psychological Narratives: The Sociocultural Basis of Understanding Reasons*, NY, Bradford Books.

Hutto, D. & Radcliffe, M. (2008) *Folk Psychology Re-Assesed*, Dodretch, Springer.

Jackson, F. (1999) “All that Can Beat Issue in the Thoery-Theory Simulation Debate”, *Philosophical Papers*, 28, (2), 77-96.

Jenkins, A. C. & Mitchell, J. P. (2009) “Mentalizing Under Uncertainty: Dissociated Neural Responses to Ambiguous and Unambiguous Mental State Inferences”, *Cerebral Cortex*, 16, 9-39.

Johnson-Laird, P. (1983) *Mental Models: Towards a Cognitive Science of Language, Inference and Consciousness*, Cambridge, Harvard University Press.

Kavanaugh RD. (2006) “Pretend Play and Theory of Mind”, en L. Balter & C.S. Tamis-LeMonda (eds.) *Child psychology: A handbook of contemporary issues*, New York, Psychology Press, 153-166.

Keysers, C. & Gazzola, V. (2007) “Integrating Simulation and Theory of Mind: From Self to Social Cognition”, *Trends in Cognitive Sciences*, 11, (5),194-196.

Keysers, C. and Gazzola, V. (2006) “Towards a Unifying Neural Theory of Social Cognition”, *Progress in Brain Research*, 156, 383–406.

Kohler, E., Keysers, C., Umilitá, M.A., Fogassi, L., Gallese, V. & Rizzolatti, G (2002) "Hearing Sounds, Understanding Actions: Action Representation in Mirror Neurons", *Science*, 297, 846-848.

Kosslyn, S. (1994) Image and brain, Cambridge, MIT Press.

Lawrence, A., Calder, A., McGowan, S. &Grasby, P. (2002) "Selective Disruption of the recognition of Facial Expressions of Anger", *NeuroReport*, 13, 881-884.

Leslie, A. (1987) "Pretense and Representation: The Origins of 'Theory of Mind'", *Psychological Review*, 94, (4), 412-426.

Leslie, A. (1991) "The Theory of Mind Impairment in Autism: Evidence for a Modular Mechanism of Development?", en A. Whiten (ed.) *Natural Theories of Mind: Evolution, Development and Simulation of Everyday Mindreading*, Oxford, Basis Blackwell, 63-78.

Leslie, A. (1994) "Pretending and Believing: Issues in the TOMM", *Cognition*, 50, 211-238.

Leslie, A. (2000) "'Theory of Mind' as a mechanism of Selective Attention", en M. Gazzaniga (ed.) *The New Cognitive Neurosciences*, Cambridge, Cambridge University Press, 1235-1248.

Leslie, A. (2005) "Developmental Parallels in Understanding Minds and Bodies", *Trends in Cognitive Sciences*, 9 (10), 459-462.

Leslie, A. M., Friedman, O., & German, T. P. (2004) "Core Mechanisms in 'Theory of Mind'", *Trends in Cognitive Sciences*, 8(12), 528-533.

Leslie, A. M., German, T. P., & Polizzi, P. (2005) "Belief-desire Reasoning as A Process of Selection", *Cognitive Psychology*, 50, 45-85.

Leslie, A. M., & Polizzi, P. (1998) "Inhibitory Processing in the False Belief Task: Two Conjectures", *Developmental Science*, 1, 247-253.

Leslie, A. & Scholl, B.J. (1999) "Modularity, Development and 'Theory of Mind'", *Mind & Language*, 14, (1), 131-153.

Leslie, A. M., & Thaiss, L. (1992) "Domain Specificity in Conceptual Development: Neuropsychological Evidence from Autism", *Cognition*, 43, 225-251.

Lewis, D. (1970) "How to Define Theoretical Terms", *The Journal of Philosophy*, 67, 13, 427-446.

Lewis, D. (1973/1999) "Psychophysical and Theoretical Identifications", en Lewis, D. (ed.) *Papers in metaphysics and Epistemology*, Cambridge, Cambridge University Press.

Linder, E. Cooper J. & Jones E. J. (1967) "Decision Freedon As a Determinant of the Role of Incentive Magnitude in Attitude Change", Journal of Personality and Social Psychology, 6, (3), 245-254.

Lycan, W. (1994) "Functionalism (1)", en Guttenplan, S. (ed.) *A Companion to the Philosophy of Mind*, Oxford, Blackwell.

Legrand, D., & Ruby, P. (2009) "What Is Self-Specific? Theoretical Investigation And Critical Review oOf Neuroimaging Results", *Psychological Review*, 116, (1), 252-282.

Loewenstein, G. & Adler, D. (1995) "A Bias in the Perception of Tastes", *Economic Journal: The Quaterly Journal of the Royal Economic Society*, 105, 929-937.

Machery, E. (2009) *Doing without concepts*, Oxford, University Press.

Meltzoff, A. (1995) "Understanding the Intentions of Others: Re-Enactment of Intended Actions by 18-onth-Old-Children", *Developmental Psychology*, 31, 838-850.

Meltzoff, A., Gopnik, A. & Repacholi, B. (1999). “Toddlers Understanding of Intentions, Desires and Emotions: Explorations of the Dark Ages”, en Zelazo, P. (ed.) *Developing Theories of Intention*, New Jersey, Erlbaum.

Milgram, S. (1963) “Behavioral Study of Obedience”, *Journal of Abnormal and Social Psychology*, 67, 371-378.

Mitchell, J.P. (2009) “Inferences about Mental States”, *Philosophical Transactions of the Royal Society*, 364, 1309-1316.

Mitchell, P. & Lacohee, H. (1991) “Children’s Early Understanding of False Belief”, *Cognition*, 39, 107-127.

Mitchell, P., Robinson, E. J., Isaacs, J. E., & Nye, R. M. (1996) “Contamination in Reasoning about False Belief: An Instance of Realist Bias in Adults But Not Children”, *Cognition*, 59, 1-21.

Morton, A. (1980) *Frames of Mind: Constraints on the Common-Sense Conception of the Mental*, Oxford, Clarendon Press.

Nichols, S., Stich, S., Leslie, A. & Klein, D. (1996) “Varieties of Off-Line Simulation”, en Carruthers, P. & Smith, P. (eds.) *Theories of Theories of Mind*, Cambridge, Cambridge University Press.

Nichols, S. (en prensa) “Mindreading and the Philosophy of Mind”, en Prinz, J. (ed.) *The Oxford Handbook on Philosophy of Psychology*, New York, Oxford University Press.

Nichols, S. & Stich, S. (2003) *Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding of Other Minds*, Oxford, Oxford University Press.

Nickerson, R. S. (1999) “How We Know ± and Sometimes Misjudge ± What Others Know: Imputing One's Own Knowledge to Others”, *Psychological Bulletin*, 125, 737-759.

Nisbett, R. & Ross, L. (1980) *Human Inference: Strategies and Shortcomings of Social Judgment*, Prentice-Hall, 1980

Nissbet, R. & Wilson T. (1977) “Telling More Than We Can Know: Verbal Reports on Mental Processes”, Psychological Review, 84, (3), 231-259.

Onishi, K. & Baillargeon, R. (2005) “Do 15.Old.Month Infants Understand False Beliefs?”, *Science*, 308(5719), 255-258.

Premack, D. & Woodruff, G. (1978) “Does Chimpanzee Have a Theory of Mind?”, *Behavioral and Brain Sciences*, 4, 515-526.

Perner, J. (1991) *Understanding the Representational Mind*, Cambridge, MIT Press.

Perner, J. (1996) “Simulation as Explicitation of Predication-Implicit Knowledge about the Mind: Arguments for a Simulation-Theory Mix”, en Carruthers, P. & Smith, P. (eds.) *Theories of Theories of Mind*, Cambridge, Cambridge University Press.

Perner, J. & López, A. (1997) “Children’s Understanding of Belief and Disconfirming Visual Evidence”, *Cognitive Development*, 12, 463-476.

Perner, J. Leekman, S. & Wimmer, H. (1987) “Three-Year-Olds' Difficulty with False Belief: The Case for A Conceptual Deficit”, *British Journal of Developmental Psychology*, 5(2), 125-137.

Pineda, J. & Hetch, E. (2008) “Mirroring and mu rhythm involvement in social cognition: Are there dissociable subcomponents of theory of mind?”, Biological Psychology, en prenss.

Putnam, H. (1975) “The Meaning of ‘Meaning’”, en *Mind, Language and Reality: Philosophical Papers*, 2, Cambridge, Cambridge University Press, 125-271.

Radcliffe, M. (2007) *Rethinking Commonsense Psychology: A Critique of Folk Psychology*, *Theory of Mind and Simulation*, NY, Macmillan.

Ravenscroft, I. (2009) “Is Folk Psychology a Theory?”, en Symons, J. & Calvo, P. (eds.) *The Routledge Companion to Philosophy of Psyhcology*, London & NY, Routledge, 131-147.

Ravenscroft, I. (2010) “Folk Psychology as Theory”, en *Stanford Encyclopedia of Philosophy* http://plato.stanford.edu/entries/folkpsych-theory/

Ramnani, N. & Miall, RC. (2004) “A System in the Human Brain for Predicting the Actions of Others”, *Nature Neuroscience*, 2004, 7(1), 85-90.

Repacholi, B. & Gopnik, A. (1997) “Early Understanding of Desires: Evidence from 14 and 16 Month Olds”, *Developmental Psychology*, 33, 12-21.

Rey, G. (2008) “(Even Higher-Order) Intentionality without Consciousness”, *Revue Internationale de PhilosophieK,* 62, 51–78.

Rizzolatti, G., Arbib, M.A., Fadiga, L. & Rizzolatti, G. (1996) "Localization of Grasp Representations in Humans by PET: I. Observation versus Execution", *Experimental Brain Research*, 111, 246-252.

Rizzolati, G., Fadiga, L., Fogassi, L. & Gallese, V. (2002) “From Mirror Neurons to Imitation: Facts and Speculations”, en Meltzoff, A. & Prinz W. (eds) *The Imintative Mind: Development, Evolution and Brain Basis*”, Cambridge, Cambridge University Press.

Rizzolati, G., Fogassi, L. & Gallese, V. (2001) "Neurophysiological Mechanisms Underlying the Understanding and Imitation ff Action", *Nature Neuroscience Reviews*, 2, 661-670.

Rizzolatti, G. & Craighero, L. (2004) "The Mirror Neuron System", *Annual Review of Neuroscience*, 27, 169-192.

Ross, L., Lepper, M. R. & Hubbard, M. (1975) "Perseverance in Self-Perception and Social Perception: Biased Attributional Processes in the Debriefing Paradigm", *Journal of Personality and Social Psychology*, 32(5), 880-892.

Roth, D. & Leslie, A. (1991) “The Recognition of Attitude Conveyed by Utterance: A Study of Preschool and Autistic Children”, *British Journal of Developmental Psychology*, 9, 315-330.

Royzman, E. B., Cassidy, K. W., & Baron, J. (2003) “‘I Know, You Know’: Epistemic Egocentrism in Children and Adults”, *Review of General Psychology*, 7, (1), 38-65.

Ryle, G. (1949) *The Concept of Mind*, London, Hutchinson.

Samson, D. Apperly, I., *et al.* (2004) “Left Temporo-parietal Junction is Necessary for Representing Someone Else’s Belief”, *Nature Neuroscience*, 7(5), 499-500.

Samson, D. Apperly, I., *et al.* (2005) “Seeing It My Way: A Case of Selective Deficit in Inhibiting Self-Perspective”, *Brain*, 128, 1102-1111.

Samson, D., Apperly, I. A., Kathirgamanathan, U., & Humphreys, G. W. (2005) “Seeing It My Way: A Case of Selective Deficit in Inhibiting Self-perspective”, *Brain*, 128, 1102-1111.

Samson, D. Apperly, I. (2007) “Error Analyses Reveal Contrasting Deficits in ‘Theory of Mind’: Neuropsychological Evidence from a 3-Option False Belief Task”, *Neuropsychologia*, 45(11), 2561-2569.

Saxe, R. & Powell, L.J. (2006) “It’s the Thought that Counts. Specific Brain Regions for One Component of the Theory of Mind”, *Psychological Science*, 17, (8), 692-699.

Saxe, R. & Kanwisher, N. (2003) "People Thinking about Thinking People: The Role Of The Temporo-Parietal Junction In ‘Theory of Mind’”, *NeuroImage*, 19, 1835-1842.

Scholl, B.J. & Leslie, A. (1999) “Modularity, Development and ‘Theory of Mind’”, *Mind & Language*, 14 (1), 131-153.

Scholl, B.J. & Leslie, A. (2001) “Mind, Modules & Meta-Analysis“, *Child Development*, 72, (3), 696-701.

Schwitzgebel, E. (2011) “Knowing Your Own Beliefs”, *Canadian Journal of Philosophy*, 35, 41-62.

Schwitzgebel, E. "Introspection", *The Stanford Encyclopedia of Philosophy* (Summer 2014 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/sum2014/entries/introspection/>.

Scott, R. M. & Baillargeon, R. (2009) “Which penguin is this? Attributing false beliefs about object identity at 18 months”, Child Development, 80, 1172-1196.

Searle, J. (1983) *Intentionality. An essay in the philosophy of mind*, Cambridge, Cambridge University Press.

Sellars, W. (1956) “Empiricism and the Philosophy of Mind”, en Feigl, H. & Scriven, M. (eds.) *The Foundations of Science and the Concepts of Psychology and Psychoanalysis: Minnesota Studies in the Philosophy of Science*, vol.1, Minneapolis: University of Minnesota Press, 253-329.

Shoemaker, S. (1996) *The First Person Perspective and Other Essays*, NY, Cambridge University Press.

Shoemaker, S. (2009) “Self-Intimation and Second Order Belief”, *Erkenntnis*, 71, 35-51.

Skidelsky, L. (2007) “La distinción doxástico-subdoxástico”, *Crítica*, 39, (115), 31-60.

Spaulding, S. (2010) “Embodied Cognition and Mindreading”, *Mind and Language*, 25, (1), 119-140.

Sprengelmeyer, R., Young, A., Schroeder, U., Grossenbacher, P., Federlein, J., Buttner, T. & Przuntek, H. (1999) “Knowing no Fear”, *Proceedings of the Royal Society*, series Biology, 266, 2451-2456.

Stich, S. (1978) “Beliefs and Subdoxastic States”, *Philosophy of Science*, 45, 499-518

Stich, S. (1992) “Folk Psychology: Simulation or Tacit Theory?”, *Mind & Language*, 7 (1-2), 35-71.

Stich, S. & Nichols, S. (1995) “Second Thoughts on Simulation”, en Davies M. & Stone, T. (eds.), *Mental Simulation: evaluations and applications*, Oxford, Blackwell, 87-108.

Stich, S. & Nichols, S. (1996) “How Do Minds Understand Minds? Mental Simulation Versus Tacit Theory”, en Stich S. *Deconstructing the Mind*, Oxford, Oxford University Press.

Stich, S. & Nichols, S. (1997) “Cognitive Penetrability, Rationality and Restricted Simulation”, *Mind & Language,* 12, 297-326.

Stich, S. & Nichols, S. (2003) “Folk Psychology”, en Stich, S. & Warfield, T.A. (eds.) *The Blackwell Guide to the Philosophy of Mind*, Oxford, Basil Blackwell, 235-255.

Stich, S. & Ravenscroft, I. (1994) “What is Folk Psychology?”, *Cognition*, 50, 447-468.

Stone, T. & Davies, M. (1996) “The Mental Simulation Debate: A Progress Report”, en Carruthers, P. & Smith, P. (eds.) *Theories of Theories of Mind*, Cambridge, Cambridge University Press.

Sommerville, J. A., Woodward, A. L. & Needham, A. (2005) “Action Experience Alters Three-Month-Old Infants’ Perceptions of Others’ Actions”, *Cognition*, 96, B1-B11.

Southgate, V., Senju, A. and Csibra, G. (2007) “Action Anticipation through Attribution of False Belief by 2-year-olds”, Psychological Science, 18, 587–92.

Spivey, M., Tyler, M., Richardsn, D, & Young, E. (2000) “Eye Movements During Comprehension of Spoken Scene Descriptions”, *Proceesings of the 22nd Annual Conference of th Cognitive Science Society*, 487-492.

Strawson, G. (1994) Mental Reallity, Cambridge, MIT press.

Surian, L., Caldi, S. and Sperber, D. (2007) “Attribution of beliefs by 13-month-old infants”,*Psychological Science*, 18, 580–6.

Tager-Flusberg, H., & Sullivan, K. (2000) “A Componential View of Theory of Mind: Evidence from Williams Syndrome”, *Cognition*, 76, 59-89.

Trevarthen, C. (1979) “Communication and Cooperation in Early Infancy: A Description of Primary Intersubjetivity”, en Bulova, M. (ed) *Before speech*, Cambridge, Cambridge University Press.

Tye, M. (1995) *Ten Problems of Consciousness: A Representtional Theory of the Phenomenal Mind*. Cambridge, Bradford Books/MIT Press.

Tye, M. "Qualia", *The Stanford Encyclopedia of Philosophy* (Fall 2013 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2013/entries/qualia/>.

Turing, A. (1950) “Computing Machinery and Intelligence”, en B. Copeland (ed.) *The Essential Turing*, 2004, Clarendon Press, Oxford, 433-464.

Uddin, L., Iacobon, M., Lange, C. & Keenan J.P. (2007) “The Self and Social Cognition: The Role of Cortical Midline Structures and Mirror Neurons”, *Trends in Cognitive Sciences*, 11, (4), 153-157.

Umilitá, M.A., Kohler, E., Gallese, V., Fogassi, L., Fadiga, L., Keysers, C., & Rizzolatti, G. (2001) "'I Know What You Are Doing': A Neurophysyological Study ", *Neuron*, 32, 91-101.

Van Boven, L., Dunning, D. & Loewenstein, G. (2000) “Egocentric Empathy Gaps between Owners and Buyers: Misperceptions of the Endowment Effect”, *Journal of Personality and Social Psychology*, 79(1), 66-76.

Van Boven, L. & Loewenstein, G. (2003) “Social Projection of Transient Drive States“, *Personality and Social Psychology Bulletin*, 29(9), 1159-1168.

Van Orden, GC. & Paap, KR (1997) “Functional Neuroimages Fail to Discover Pieces of Mind in the Parts of the Brain”, *Philosophy of Science*, 64, S85-S94.

Vogeley, K. Bussfeld, A., Newen, S., Herrman, F., Happé, F., Flakai, P., Maier, W., Shah, N., Fink, G. and Zilles, K. (2001) “Mind Reading: Neural Mechanisms of Theory of Mind and Self-Perspective”, *NeuroImage*, 14, 170-181.

Watson, J.B (1913) "Psychology as the Behaviorist Views It", Psychological Review, 20, 158-177.

Wellman, H. (1990) *The Child’s theory of Mind*, Cambridge, MIT Press.

Wellman, H. & Banerjee, M. (1991)”Mind and Emotion: Children's Understanding of the Emotional Consequences of Beliefs and Desires”, *British Journal of Developmental Psychology*, 9, 191-214.

Wellman, H. & Wooley, J. (1990) “From Simple Desires to Ordinary Beliefs: The Early Development of Everyday Psychology”, *Cognition*, 35, 245-273.

Wellman, H., Cross, D. & Watson, J. (2001) “Meta-Analysis of Theory-of-Mind Development: The Truth about False Belief”, *Child Development*, 72, (3), 655-684.

Wells, G. & Petty, R. (1980) “The effects of overt head movements on persuasion: Compatibility and incompatibility of responses”, *Basic and Applied Social Psychology*, 1, 219–30.

Wegner, D. (2002) *The illusion of conscious will*, Cambridge, MIT Press.

Wegner, D. & Wheatley, T. (1999) “Apparent Mental Causation: Sources of the Experience of the Will”, *American Psychologist*, 54, 480–91.

Wicker, B., Keysers, C., Plailly, J., Royet, J.P., Gallese, V. & Rizzolatti, G. (2003) “Both of Us Disgusted in My Insula:The Common Neural Basis of Seeing and Feeling Disgust”, Neuron, 40, 655-664.

Wilson, T. (2002) *Strangers to ourselves*, Cambridge, Harvard University Press.

Wimmer, H. & Perner, J. (1983) “Beliefs about Beliefs”, *Cognition*, 13, 103-128.

Woodward, A. L. (2009) “Infants’ Grasp of Others’ Intentions”, *Current Directions in Psychological Science*, 18(1), 53-57.

Zahavi, D. (2008) “Simulation, Projection and Empathy”, *Consciousness & Cognition*, 17, 514-522.

Zaitchik, D. (1990) “When Representations Conflict with Reality: The Preschooler's Problem with False Beliefs and ‘False’ Photographs”, *Cognition*, 35, 41-68.