CDD: 101

Una critítica al criterio de correción para distinguir entre tipos de procesos subyacentes en una teoría híbrida de *mindreading*

Fernanda Velázquez Coccia

Universidad de Buenos Aires
Instituto de Filosofia
Puán 480 Buenos Aires 1053
Buenos Aires, Argentina
fernandavelaz@gmail.com

Received: 15.03.2016; Revised: 10.08.2016; Accepted: 15.08.2016

DOI: http://dx.doi.org/10.1590/0100-6045.2016.V39N2.FVC

Abstract: Los enfoques híbridos de *mindreading* postulan procesos de "teoría" y "simulación" como subyacentes a esta capacidad. Un problema actual es ofrecer criterios para evaluar las teorías híbridas. Aquí, analizaré el "criterio de corrección" propuesto por Stich & Nichols (2003, Nichols & Stich 2003) para discriminar entre tipos de procesos subyacentes a *mindreading*. Según éste, si el resultado de*mindreading* es correcto es probable que el proceso subyacente sea de tipo simulacional, y si el resultado de *mindreading* es incorrecto es probable que el proceso subyacente sea de bases de información (o teoría). Sostendré que la corrección o incorrección del resultado de *mindreading* no permite distinguir entre tipos de proceso. Intentaré mostrar que los argumentos a favor de un proceso simulacional subyacente a *mindreading* exitoso no descartan explicaciones de bases de información para este mismo fenómeno. A su vez, argumentaré que es posible que un proceso de tipo simulacional subyazca a *mindreading* incorrecto.

Keywords: Enfoques híbridos de la capacidad mentalista , Teoría de la Teoría , Teoría de la Simulación

1. Introducción

Los seres humanos son mentalistas en tanto se atribuyen estados mentales a sí mismos y a las otras personas, y asumen que estos son las causas del comportamiento. A su vez, tales atribuciones se utilizan para explicar y

predecir el pensamiento y el comportamiento, propio y ajeno. Un modo de abordar este fenómeno supone concebirlo como una capacidad cognitiva, o un conjunto de capacidades y estudiar los procesos cognitivos subyacentes. Cuando el fenómeno se aborda desde esta perspectiva, usualmente, se lo denomina *mindreading*. En este sentido, es preciso distinguir entre las autoatribuciones, o *mindreading* de primera persona, y las atribuciones mentalistas a otras personas, o *mindreading* de tercera persona. Algunos prefieren reservar el término "*mindreading*" para la tercera persona y utilizar "metacognición" para las autoatribuciones (Carruthers 2009). Este trabajo se centra en *mindreading* de tercera persona. Cuando sea necesaria la aclaración, distinguiré entre *mindreading* de primera y de tercera persona.

Tradicionalmente, *mindreading* se ha estudiado en el marco del debate teoría-simulación. Este debate se caracteriza por el enfrentamiento entre dos enfoques sobre los procesos subyacentes a *mindreading*, la Teoría de la Teoría (TT) y la Teoría de la Simulación (TS). No obstante, *mindreading* ha mostrado ser un fenómeno complejo y multifacético, y esto ha llevado a muchos teóricos, que antes sostenían posturas puras, a sostener enfoques híbridos de teoría y simulación para dar cuenta del mismo. Un problema actual es ofrecer criterios para evaluar teorías híbridas. Aquí, me ocuparé del "criterio de corrección" propuesto en el enfoque híbrido de teoría y simulación de Nichols & Stich (2003, Stich & Nichols 2003). Usualmente, se asume que el resultado de *mindreading* es correcto cuando su *output*, por ejemplo la predicción del comportamiento ajeno, concuerda con el comportamiento de la persona que es blanco de la predicción. 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. El objetivo de este trabajo es evaluar si este criterio de corrección permite distinguir entre teoría y simulación como procesos subyacentes a un caso de *mindreading*.

Para llevar a cabo esta tarea, en la sección 2 me ocuparé brevemente de la cuestión de los enfoques híbridos. Estos enfoques postulan procesos diferentes subyacentes a una misma capacidad cognitiva. Esto requiere brindar algún criterio para determinar el tipo de proceso que subyace a cierto caso de *mindreading*. Asimismo, presentaré brevemente el enfoque híbrido de Nichols & Stich (2003, Stich & Nichols 2003) y el criterio de corrección propuesto para distinguir entre procesos subyacentes. Para evaluar este criterio lo dividiré en

dos aspectos. En la sección 3, sostendré que los argumentos que fundamentan el primer aspecto del criterio de corrección, *i.e.* si el resultado de *mindreading* es correcto es probable que el proceso subyacente sea de tipo simulacional, tienen problemas. En la sección 4, me ocuparé del segundo aspecto del criterio de corrección, *i.e.* si el resultado de *mindreading* resulta sistemáticamente incorrecto es probable que el proceso subyacente sea de tipo rico en información. Intentaré mostrar, no obstante, que un proceso de tipo simulacional también puede subyacer a *mindreading* con un resultado incorrecto. Concluiré que ambos aspectos del criterio de corrección no son satisfactorios ya que no permiten distinguir entre procesos de teoría o simulación subyacentes a un caso de *mindreading*.

2. Los enfoques híbridos de teoría y simulación

Tradicionalmente, *mindreading* se ha estudiado en el marco del debate teoría-simulación, que se caracteriza por el enfrentamiento entre dos enfoques puros sobre los procesos subyacentes a *mindreading*. Uno postula procesos ricos en información. Otro postula procesos pobres en información. De acuerdo con la TT, *mindreading* se lleva a cabo utilizando una teoría psicológica conformada por un conjunto de conceptos de estados mentales (tales como *creencia* y *deseo*) y ciertos principios o reglas generales que rigen la interacción entre los mismos (por ejemplo, las personas actúan en base a sus creencias para satisfacer sus deseos). Haciendo uso de esta teoría y a partir de información apropiada acerca del agente 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 acerca del comportamiento (Fodor 1987, Stich & Nichols 1992, 1995, 1996, 1997).

Según la TS, en cambio, no es necesario recurrir a generalizaciones psicológicas para que tenga lugar *mindreading*, sino que basta con la simulación mental (Gordon 1992, Heal 1995, Goldman 1995, Harris 1992)[1]. En líneas

[1] La noción de "simulación mental" debe distinguirse de otros usos de "simulación", particularmente, de la noción proveniente de la ciencia de la computación. En la simulación computacional se intenta predecir el comportamiento de un sistema utilizando un modelo computacional del mismo. Tales modelos pueden caracterizarse

generales, ésta puede entenderse como la utilización de los recursos mentales propios como "modelo" de la mente del otro, para realizar predicciones acerca del comportamiento ajeno. Esto, en virtud de que la mente del otro es similar a la propia (Stich & Nichols 1992). En este sentido, no resulta necesario recurrir a inferencias basadas en generalizaciones psicológicas para que tenga lugar *mindreading*.

En este debate, se han tratado de identificar, desde una perspectiva conceptual, las virtudes explicativas de las teorías con el objetivo de elegir la mejor. Para evaluar teorías, usualmente, se utilizan criterios generales tales como la precisión, el alcance, la simplicidad, entre otros. En concordancia con estos criterios, se ha señalado, por ejemplo, que la TT resulta menos simple que la TS en la medida en que postula la posesión de generalizaciones específicas acerca de estados mentales (Goldman 1995). No obstante, también se ha cuestionado la simplicidad de la TS en tanto que postula entidades novedosas como los mecanismos generadores de *inputs* ficticios (Stich & Nichols 1997). Desde una perspectiva empírica, se asumió que de la TT y la TS se desprenderían hipótesis que al ser puestas a prueba empíricamente permitirían elegir entre las mismas. Sin embargo, tampoco ha sido posible elegir de este modo entre los enfoques. Al parecer, la mayor dificultad reside en que no se trata de una oposición simple entre "teoría" y "simulación", sino entre las distintas versiones de cada enfoque. En suma, los esfuerzos para tratar de elegir entre enfoques sopesando evidencia empírica y argumentos conceptuales no han sido fructíferos.

Esta situación ha generado, en la literatura de *mindreading*, una tendencia a postular teorías híbridas, incluso entre quienes antes sostuvieron teorías puras (Nichols & Stich 2003, Goldman 2006). Esto está en concordancia, por un lado, con el acuerdo creciente respecto de que *mindreading* es un fenómeno tan

---

por ecuaciones matemáticas o reglas. Justamente, estos modelos suponen cuerpos de información acerca del sistema a ser simulado, por ejemplo, la simulación computacional de un huracán se lleva a cabo mediante un programa computacional que tiene información acerca de las leyes de aerodinámica e hidrodinámica que gobiernan los huracanes (Haugeland 1985). En este sentido, se trata de simulaciones basadas en información, mientras que la simulación mental postula la posibilidad de realizar predicciones del comportamiento de un sistema sin recurrir a cuerpos de información acerca de su funcionamiento.

complejo y multifacético que no puede explicarse por enfoques puros de teoría o simulación y que se precisan enfoques mixtos o híbridos (Stich & Nichols 1995, Nichols *et al.* 1996, Perner 1996, Davies & Stone 1998) y, por el otro, con la aceptación de que cada propuesta teórica tiene casos a favor (véase Stich & Nichols 2003, Goldman 2006, Apperly 2008, entre otros).

Así, en los enfoques híbridos de *mindreading* se asume que teoría y simulación no están en competencia sino que, más bien, ambos procesos contribuyen a la capacidad. De modo que, al menos en este contexto, el término "híbrido" puede entenderse en el sentido de que dos tipos de procesos que anteriormente eran considerados excluyentes, ahora se consideran ambos subyacentes a *mindreading*. Ahora bien, la cuestión es cómo dos tipos distintos de procesos pueden subyacer a una misma capacidad cognitiva. En principio, se pueden concebir varias relaciones entre los mismos. Los procesos subyacentes pueden complementarse, superponerse o alternarse para dar lugar a la capacidad. En este sentido, se puede sostener, por ejemplo, que los procesos se alternan de modo tal que, en algunas condiciones interviene uno y, en otras condiciones, el otro.

Las propuestas híbridas de *mindreading* disponibles hasta el momento no describen satisfactoriamente la relación entre los procesos subyacentes (Nichols & Stich 2003, Goldman 2006). Particularmente, Nichols & Stich (2003) proponen un enfoque híbrido de *mindreading* en el que intervienen procesos de tipo simulacional y ricos en información, y dos relaciones entre estos que pueden caracterizarse como cooperación y alternancia[2]. Nichols & Stich (2003) postulan dos sistemas diferenciables filogenética y ontogenéticamente subyacentes a *mindreading*. El sistema temprano, y más antiguo, permite predecir

[2] Stich & Nichols (2003) no explicitan el vínculo entre las relaciones de cooperación y alternancia para los distintos tipos de procesos subyacentes a *mindreading*. En principio, es pensable que constituyen dos aspectos de una misma teoría híbrida, aunque no en un sentido de superposición sino, más bien, como una división de tareas. Estas dos relaciones entre procesos se postulan al desarrollar distintos aspectos de *mindreading*. A saber, la cooperación entre simulación y teoría subyace a *mindreading* cuando se trata de la anticipación del comportamiento ajeno, mientras que la alternancia entre teoría y simulación subyace a *mindreading* cuando se trata de la anticipación del pensamiento ajeno, particularmente, de las inferencias ajenas.

el comportamiento de los otros atribuyendo metas o deseos, y definiendo la mejor estrategia para alcanzarlos. Este sistema está compuesto por ciertos mecanismos (los detectores de deseo, el planificador y el coordinador de *mindreading*) que involucran procesos de tipo ricos en información. Esto se asume en virtud de la presencia de errores sistemáticos asociados a los mismos (*e.g.* la presencia de errores sistemáticos en la detección de los deseos ajenos), que se conciben como predicciones erróneas generadas por cuerpos de información subyacentes parcialmente incompletos o erróneos.

El sistema tardío utiliza los mecanismos del sistema temprano pero recluta, además, la caja de mundos posibles, un componente perteneciente a otra capacidad cognitiva denominada "ficción". Este sistema posibilita generar un modelo de las creencias del blanco y, así, se pueden contemplar creencias que el *mindreader* no posee al llevar a cabo *mindreading*. Esto otorga un mayor poder predictivo en comparación con el sistema temprano, que sólo permite predecir el comportamiento ajeno a partir de las creencias del *mindreader*. El aspecto simulacional de la propuesta híbrida consiste en utilizar las operaciones de los recursos mentales propios (*e.g.* el sistema de toma de decisiones propio) sobre el modelo de las creencias del blanco para llevar a cabo *mindreading* (*e.g.* la predicción de las decisiones ajenas). Hasta aquí, la relación entre procesos sugerida resulta de tipo cooperativo en tanto que los distintos procesos están involucrados en la predicción del comportamiento ajeno. Mediante los procesos ricos en información del sistema temprano, se detectan las metas del blanco, y el sistema tardío genera un modelo de las creencias del blanco, sobre el cual operan los recursos cognitivos propios del *mindreader* (el aspecto simulacional).

En esta propuesta híbrida se postula, también, otro tipo de relación entre los distintos procesos subyacentes, que no está tan descripta y que se puede considerar como una relación de alternancia. Al dar cuenta de la capacidad de predecir exitosamente las inferencias ajenas, los autores sostienen que es probable que un proceso simulacional subyazca a este aspecto de *mindreading*, en la medida en que resulta más plausible que se recurra a los mecanismos inferenciales propios que a un conjunto de generalizaciones sobre cómo las personas realizan inferencias (Nichols & Stich 2003). Se sostiene, además, que 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. Y 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, p. 245-246). Así, según estas afirmaciones, se sugiere el siguiente criterio de corrección para determinar el proceso subyacente a un caso de *mindreading*: 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. De este modo, se postula una relación de alternancia entre los procesos subyacentes a *mindreading* basada en el criterio de corrección.

Ahora bien, si ambos procesos contribuyen a la capacidad alternándose entre sí se requiere establecer bajo qué condiciones lo hacen. De lo contrario, un enfoque híbrido apenas estaría postulando que, a *mindreading*, a veces le subyace un proceso de tipo simulacional y a veces, un proceso de tipo rico en información. Pero este no parece ser el propósito de un enfoque híbrido, ya que esto no permitiría realizar predicciones de ningún tipo. De modo que para que un enfoque híbrido de teoría y simulación pueda generar predicciones adecuadas parece necesario que ofrezca un criterio claro para determinar qué tipo de proceso subyace a una instancia de *mindreading*.

No obstante, alguien podría poner en duda la necesidad de un criterio para distinguir entre procesos en las teorías híbridas. Usualmente, en las teorías de la ciencia cognitiva que postulan más de un proceso subyacente a una capacidad, se asume que los mismos están en competencia (por ejemplo, la propuesta de "Teorías Multiproceso", Machery 2009). Sin embargo, ésta no es la intuición de quien ofrece un enfoque híbrido de teoría y simulación para *mindreading*. En éste se asume que los procesos subyacentes no están en competencia, sino que, más bien, colaboran de algún modo. Por esta razón, una de las cuestiones fundamentales de las teorías híbridas consiste en ofrecer un enfoque acerca de la naturaleza de esta colaboración.

En la propuesta de Nichols & Stich (2003, Stich & Nichols 2003) se sugieren dos formas de colaboración entre procesos subyacentes. Por un lado, la relación de cooperación en la que los dos tipos de procesos participan a la vez para que tenga lugar *mindreading*. Esta relación está descripta y bastante desarrollada. Por otro lado, se sugiere una relación de alternancia entre los distintos procesos subyacentes regida por un criterio de corrección. Esta relación no está suficientemente descripta ni, por ende, desarrollada. A

continuación, me ocuparé de esta relación de alternancia y evaluaré si el criterio de corrección en que se basa permite distinguir efectivamente entre tipos de procesos en una teoría híbrida de *mindreading*. Analizaré el criterio de corrección en función de dos aspectos. Según el primer aspecto, si el resultado de *mindreading* es correcto es probable que el proceso subyacente sea de tipo simulacional. El segundo aspecto del criterio afirma que si sistemáticamente nos equivocamos en *mindreading* es probable que el proceso subyacente sea de tipo rico en información.

3. El criterio de *mindreading* exitoso

Según el primer aspecto del criterio de corrección, si el resultado de *mindreading* es correcto es probable que el proceso subyacente sea de tipo simulacional. Stich and Nichols (2003) afirman esto luego de mencionar un hecho notorio en relación con las capacidades de *mindreading* de un adulto normal. Los seres humanos somos particularmente buenos para predecir las inferencias de las otras personas. Y, al sostener esto, aluden a fenómenos en los que el *mindreader* se genera creencias acerca de las inferencias que otras personas podrían llevar a cabo y, aún cuando tales inferencias podrían ser falsas, se predicen de igual manera. Por ejemplo, luego de escuchar junto a Juan la noticia de que el presidente ha renunciado, el *mindreader* se forma la creencia de que Juan cree que el próximo presidente será el vicepresidente. Se realiza la predicción acerca de las inferencias de Juan, aún cuando tal inferencia podría ser incorrecta en varios sentidos, por ejemplo, si fuera el caso que el vicepresidente hubiera renunciado. No obstante, según Stich & Nichols (2003), la predicción se lleva a cabo y además, usualmente, resulta correcta. Este fenómeno tan notorio es algo que un enfoque adecuado de *mindreading* tiene que poder explicar.

En relación con esto, se ofrecen dos argumentos: el argumento de la simplicidad y el argumento del acierto en *mindreading*. En la sección 3.1 discutiré el argumento de la simplicidad según el cual debe preferirse la explicación de la predicción exitosa de las inferencias ajenas que ofrece un enfoque simulacional. Esta afirmación se basa en que, al parecer, un defensor del enfoque de procesos ricos en información se ve en la necesidad de postular teorías reduplicadas para dar cuenta de la predicción de las inferencias ajenas, y esto resulta una

explicación derrochadora, tal como sugiere Harris (1992). Intentaré mostrar que adoptar un enfoque de TT no implica necesariamente postular teorías reduplicadas de modo que este argumento no es suficiente para sostener que un proceso simulacional subyace a la predicción exitosa de las inferencias ajenas. En la sección 3.2 discutiré el argumento del acierto en *mindreading*. Este argumento tiene la forma de una inferencia a la mejor explicación para la predicción exitosa de inferencias ajenas. Sin embargo, sostendré que en tanto inferencia a la mejor explicación falla puesto que un enfoque de procesos ricos en información puede explicar de una mejor manera el resultado sistemáticamente correcto de *mindreading* en determinados casos, por ejemplo, la predicción del comportamiento ajeno. En este sentido, ambos argumentos tienen problemas para brindar fundamento al primer aspecto del criterio.

3.1. El argumento de la simplicidad

Según Nichols & Stich (2003), las diferentes versiones de la TT no se han ocupado del fenómeno de la predicción exitosa de las inferencias ajenas. Esto les resulta llamativo y proponen una razón para esta omisión. Sostienen que la única explicación que un defensor de un enfoque rico en información tiene a la mano para dar cuenta de este fenómeno, implica postular que poseemos una buena teoría al respecto. En principio, esto implicaría postular una reduplicación de teorías o cuerpos de información en tanto que a los principios que ya poseemos para realizar las inferencias propias, hay que agregar el conjunto de generalizaciones sobre cómo las personas llevan a cabo inferencias. Y, esto *resulta un enfoque derrochador* (Stich & Nichols 2003). Resulta más parsimonioso, en cambio, postular que utilizamos el sistema inferencial propio para predecir las inferencias ajenas, tal como se postula en el enfoque simulacional. Así, los teóricos de la TT deben haber omitido este fenómeno en virtud de que la única explicación a su alcance resulta poco parsimoniosa (Stich & Nichols 2003, Nichols & Stich 2003).

La necesidad de reduplicar teorías está relacionada, al parecer, con aquello que implica adoptar un enfoque rico en información. En la versión de tendencia científico-cognitiva de la TT, como la que defienden Stich y sus colegas, se postulan cuerpos de información subyacentes a *mindreading*, que guían la ejecución de esta capacidad (Stich & Nichols 1992, 1995; Nichols, Stich

& Klein 1996). En este sentido, se asume que un defensor de esta sólo puede dar cuenta de capacidades cognitivas postulando cuerpos de información. Más específicamente, en el caso de la predicción exitosa de las inferencias ajenas, se requiere postular *más* información en tanto que se requiere postular ciertas generalizaciones sobre cómo las personas llevan a cabo inferencias, y esto constituye un cuerpo de información distinto del cuerpo de principios que ya poseemos para realizar las inferencias propias (Stich & Nichols 2003). En este sentido, se asume que un enfoque rico en información requiere postular una reduplicación de teorías o cuerpos de información.

Para mostrar por qué esto resulta derrochador, Stich & Nichols ofrecen una analogía con un argumento ofrecido por Harris (1992), un defensor de la simulación.

> 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 de 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, p. 244)

De modo que Stich & Nichols (2003) afirman que la explicación de la predicción de las inferencias ajenas que ofrece un enfoque de TT resulta derrochadora por razones similares a las que se brindan en el argumento de Harris (1992). Éste muestra lo absurdo que resultaría poseer una teoría acerca de las intuiciones gramaticales ajenas a la vez que poseemos una gramática para llevar a cabo los juicios gramaticales propios. La explicación del enfoque simulacional, que postula que usamos el mismo mecanismo para realizar y predecir juicios gramaticales, resulta más parsimoniosa.

Harris propone pensar en un estudio psicolingüístico hipotético. Supóngase que se le han presentado a un grupo de hablantes de un idioma, *e.g.* español, una serie de oraciones gramaticales y agramaticales para que estos lleven a cabo juicios de gramaticalidad. Si se le pide a otro hablante del idioma español que prediga las decisiones de los participantes respecto de cada una de

las oraciones, el éxito en esta tarea será notable. Es más, si se le pidiera al participante de la segunda tarea que justificara sus decisiones, señalaría las mismas construcciones y morfemas en las oraciones agramaticales que los participantes de la primera tarea. De modo que hay que explicar cómo estas predicciones son tan acertadas (Harris 1992).

Una respuesta posible, según Harris la más plausible, es que el participante de la segunda tarea leyó cada oración y se preguntó si le sonaba gramatical o no, asumiendo que los otros hablantes del español llevarían a cabo los mismos juicios de gramaticalidad basados en las mismas razones. Una explicación alternativa es que se poseen dos cuerpos de representaciones de la gramática española. Uno de primer orden, que se usa para llevar a cabo los juicios propios. Otro, que representa las representaciones de las otras personas y produce juicios equivalentes al primero. Éste, se usa para predecir los juicios que los otros llevan a cabo. Sin embargo, según Harris, esta alternativa resulta poco creíble y poco parsimoniosa. (Harris 1992). Así, Harris muestra lo absurdo que resultaría poseer una teoría sobre las intuiciones gramaticales ajenas a la vez que poseemos una teoría para llevar a cabo juicios gramaticales propios (la gramática). De modo que la explicación del enfoque simulacional, que postula un único mecanismo, resulta más parsimoniosa.

Sin embargo, los argumentos de Harris (1992) y de Nichols & Stich (2003), que abogan por una explicación simulacional para la predicción exitosa de estados mentales ajenos, no son buenos argumentos. Es cuestionable la necesidad de la reduplicación de teorías que le atribuyen al enfoque de procesos ricos en información. En primer lugar, el argumento depende de una analogía entre la sintaxis y la teoría subyacente a *mindreading* que no se sostiene. El punto de partida del argumento es que un teórico de la teoría está comprometido a apelar a cuerpos de información para explicar capacidades cognitivas. Si éste sostiene que los juicios de gramaticalidad se llevan a cabo mediante una gramática subyacente, se encuentra con el problema de que la gramática no puede considerarse un cuerpo de información acerca de la psicología humana. Ésta versa sobre la sintaxis y no sobre las "creencias" acerca de la sintaxis. Puesto que la gramática y las creencias acerca de la gramática son dominios diferentes, les corresponden teorías diferentes (Goldman 2006). Dado que un teórico de la teoría no puede recurrir a la gramática para explicar la predicción de las intuiciones lingüísticas ajenas, el único recurso que tiene a la mano consiste en postular teorías reduplicadas. La gramática y una teoría que abarca

un conjunto de generalizaciones sobre creencias acerca de la gramática o cómo otros hablantes usan la gramática para realizar juicios lingüísticos.

Ahora bien, en el caso de la predicción de las intuiciones gramaticales, la necesidad de reduplicar teorías parece, en principio, pertinente puesto que se trata de dominios diferentes. Sin embargo, la afirmación de la necesidad de reduplicación de teorías no puede extenderse simplemente al caso de *mindreading*. Las atribuciones de estados mentales a otras personas y a uno mismo no pertenecen a dominios distintos como sucede en el caso de la sintaxis y las creencias acerca de cómo los otros usan la sintaxis. En principio, no hay nada que impida que el cuerpo de información psicológica que subyace a la atribución de estados mentales a otras personas, también pueda subyacer a la autoatribución, tal como han sostenido algunos filósofos y psicólogos (véase Sellars 1956 y Gopnik 1993, entre otros).

En segundo lugar, es discutible que la necesidad de reduplicar teorías se sostenga para el caso específico de la predicción de las inferencias ajenas. En el argumento de Stich & Nichols (2003, Nichols & Stich 2003), parece asumirse algo similar a lo que se asume para las intuiciones lingüísticas. Las inferencias y las creencias acerca de las inferencias son dominios distintos y, por esto, les corresponden teorías distintas (Stich & Nichols 2003, Nichols & Stich 2003, Goldman 2006). En este sentido, poseer un conjunto de principios que guían las inferencias no implica poseer principios sobre cómo las personas llevan a cabo inferencias en general. Tales principios sólo indican qué conclusiones se pueden inferir dadas ciertas premisas, pero no versan sobre qué es lo que infieren las personas a partir de ciertas premisas (Goldman 2006). De modo que se asume que un teórico de la teoría se vería obligado a postular una reduplicación de teorías para dar cuenta de la predicción de las inferencias ajenas. Algo así como una teoría psicológica sobre cómo las personas llevan a cabo inferencias, que es distinta de los principios que las personas poseen para realizar inferencias.

Vale la pena señalar que cuando los defensores de la simulación y de los enfoques híbridos, como Stich & Nichols, se ocupan de la capacidad inferencial y de lo que ésta implica desde el punto de vista de un enfoque de procesos ricos en información, no queda muy claro qué tienen en mente. Simplemente, se afirma que poseemos principios para llevar a cabo inferencias. Ahora bien, si lo que quieren decir es que las personas tienen una "lógica en la cabeza", en el sentido de una teoría formal de la demostración y las inferencias

válidas que guía la capacidad inferencial, surge un problema. La capacidad de realizar inferencias válidas no implica que los principios según los cuales éstas se rigen estén representados internamente. De modo que, aún cuando se concediera la necesidad de postular un cuerpo de información acerca de cómo las personas llevan a cabo inferencias para dar cuenta de la capacidad exitosa de predecir las inferencias ajenas, no está claro en qué consiste el otro aspecto de la reduplicación, *i.e.* los principios para realizar las inferencias propias. En virtud de que no está claro qué es aquello que está representado internamente y que guía la capacidad de realizar inferencias, no parece pertinente afirmar la necesidad de una reduplicación para el caso de la predicción de inferencias ajenas.

En tercer lugar, hasta donde sé, no hay ninguna versión de la TT que sostenga la reduplicación de teorías. En las distintas versiones de la "teoría de la teoría" (*i.e.* el enfoque causal, el enfoque de niño científico, el enfoque modular) se asume que utilizamos una misma teoría para atribuir estados mentales, y explicar y predecir el comportamiento, propio y ajeno (véase Sellars 1956, Gopnik & Meltzoff 1997, Leslie 2000, entre otros). En el enfoque causal se propone que, al adscribir estados mentales a otros para explicar y predecir su comportamiento, se postulan entidades teóricas para explicar y predecir fenómenos observables (Sellars 1956). En tanto se postulan estados mentales que no son observables, como las creencias, el conocimiento de sentido común sobre la mente es una "teoría" psicológica. Lewis (1970, 1973) postula que esta teoría psicológica ordinaria, que introduce términos teóricos al modo de las teorías científicas, 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 1973). Es preciso señalar que en este enfoque se asume, particularmente, que los términos teóricos que se utilizan para interpretar el comportamiento de las otras personas también se utilizan en la autodescripción del comportamiento propio (Sellars 1956, p. 58).

Según el enfoque del niño científico, la comprensión de la mente se constituye como una teoría representacional implícita que es análoga a las teorías científicas, los cambios en esta comprensión pueden asimilarse a los cambios conceptuales en las teorías empíricas (Gopnik & Meltzoff 1997). Además, según este enfoque, la teoría se aplica por igual al comportamiento

propio y ajeno, de modo que si uno posee una concepción errónea o incompleta de un concepto mental, por ejemplo *creencia*, entonces se atribuirán incorrectamente creencias no sólo a los otros sino, también, a uno mismo (Gopnik 1993, Gopnik & Meltzoff 1994). Este enfoque encuentra apoyo en la evidencia empírica que sugiere que los niños cometen errores en la comprensión de las creencias ajenas y, también, de las propias. Particularmente, se ha observado que la capacidad de los niños para responder correctamente a la pregunta por sus propias creencias falsas se correlaciona con la capacidad para responder exitosamente la pregunta sobre las creencias falsas ajenas (Gopnik & Astington 1988).

El enfoque modular postula la existencia de un mecanismo específico subyacente a *mindreading*, el mecanismo de "Teoría de la Mente" (MTdM). En términos generales, éste permite atender a los comportamientos e inferir los estados mentales que los causan (Leslie 1987). Más específicamente, el MTdM pone a disposición del niño pequeño conceptos innatos de estados mentales, antes de que éste haya adquirido otros conceptos abstractos mediante la construcción general de teorías, y le provee de un *insight* intencional respecto del comportamiento de los otros (Scholl & Leslie 2001). Para esto, se asume que los mecanismos de procesamiento apropiados usan un sistema de representación capaz de representar estados mentales, denominado sistema "m-representación". Éste provee descripciones del comportamiento centradas en un agente, que hacen explícita cuatro tipos de información. El agente implicado (*e.g.* mamá), la actitud del agente (*e.g.* finge que), el aspecto del mundo que ancla la actitud del agente (*e.g. esta* banana) y el contenido de la actitud del agente (*e.g.* es un teléfono). Se asume que este sistema m-representación es muy flexible de modo que puede identificar a uno mismo como el agente y describir un comportamiento tal como *yo* finjo que la banana es un teléfono (Leslie 2000).

Así, según las distintas versiones de la TT, un mismo cuerpo de información subyace a las atribuciones mentalistas de primera y tercera persona. No se trata de dominios diferentes. En este sentido, puede apreciarse que, del compromiso con la postulación de cuerpos de información subyacentes a una capacidad, no se sigue necesariamente la postulación de teorías reduplicadas, en el sentido de principios que se apliquen para las atribuciones de primera persona y otros para las de tercera persona. En consecuencia, la explicación de *mindreading* que ofrecen las distintas versiones de la TT que apela a un único cuerpo de información subyacente, no resulta afectada por consideraciones de

parsimonia. De modo que no sólo el enfoque simulacional postula un único tipo de mecanismo, sino que las distintas versiones de la TT también lo hacen.

3.2. El argumento del acierto en *mindreading*

El argumento de la simplicidad no es la única razón que se esgrime para preferir la explicación de un enfoque simulacional de la predicción exitosa de inferencias ajenas en lugar de la explicación propuesta por un enfoque rico en información. Stich & Nichols (2003) proponen el argumento del acierto en *mindreading* (*the argument from accuracy*) y consideran que este segundo argumento puede servir como una heurística para decidir a qué aspectos de *mindreading* plausiblemente les subyace un proceso simulacional (Stich & Nichols 2003). De este modo, la afirmación de que un proceso simulacional subyace a la capacidad notoriamente exitosa de la predicción de las inferencias ajenas se extiende a todos aquellos casos de predicción y atribución de estados mentales ajenos en los que somos particularmente exitosos.

El argumento del acierto en *mindreading* comienza con la observación de que los seres humanos somos muy buenos para ciertas tareas de *mindreading*, *e.g.* para predecir las inferencias ajenas, pero muy malos para otras, *e.g.* para predecir los deseos ajenos. Según los autores, esto implica un problema para un enfoque rico en información en tanto que tiene que dar cuenta de:

> ¿Cómo los *mindreaders* se las arreglan para tener una teoría tan *acertada* acerca de cómo las personas realizan inferencias – una teoría que permite predicciones correctas incluso en inferencias de tipo novedosas? El problema se vuelve más grave por el hecho de que hay otras clases de tareas de *mindreading* que las personas realizan muy mal. ¿Por qué las personas adquieren una teoría correcta acerca de las inferencias y una teoría *incorrecta* acerca de otros procesos mentales? Contrariamente, un enfoque simulacional de la predicción de inferencias tiene una explicación de nuestros aciertos. Según el enfoque simulacional, utilizamos el mismo mecanismo inferencial en ambos casos, al *realizar* y al *predecir* inferencias, entonces se espera que podamos predecir que las otras personas llevan a cabo las mismas inferencias que nosotros realizamos. (Stich & Nichols 2003, p. 245)

Así, Stich & Nichols (2003) ofrecen una inferencia a la mejor explicación. Según los autores, un enfoque rico en información puede ofrecer

una explicación de los errores sistemáticos en *mindreading*, puesto que tales errores se explican usualmente como predicciones incorrectas producto de teorías parcialmente incompletas o erróneas. Sin embargo, un enfoque rico en información no está en condiciones de ofrecer una buena explicación del éxito sistemático en la predicción de las inferencias ajenas, en virtud del argumento que reconstruyo a continuación.

Según la literatura en psicología social cognitiva, somos malos para realizar inferencias. Se ha reportado la existencia de una serie de sesgos y heurísticos a los que los agentes están sujetos al resolver problemas que requieren inferencias deductivas e inductivas (*e.g.* Kahneman, Slovic & Tversky 1982). Dado que nos equivocamos sistemáticamente al realizar inferencias, si una teoría subyace a esta capacidad inferencial, debe tratarse de una teoría parcialmente errónea o incompleta. Ahora bien, si se usara esta teoría para predecir las inferencias ajenas, las predicciones tendrían que resultar erróneas. Sin embargo, nuestra capacidad para realizar predicciones sobre las inferencias ajenas es muy buena. Acertamos sistemáticamente, sean éstas malas o buenas inferencias (Stich & Nichols 2003). Puesto que una teoría parcialmente errónea no puede generar inferencias sistemáticamente acertadas, la teoría que subyace a nuestra capacidad de realizar inferencias no puede subyacer a nuestra capacidad de predecir las inferencias ajenas.

La explicación del enfoque simulacional, que postula el recurso a un mismo mecanismo para realizar y predecir inferencias, resulta una mejor explicación de la predicción exitosa de las inferencias ajenas. Se asume que 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. Este argumento se considera, a su vez, "una espada de doble filo" en tanto implica también que 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, pp-245-246). De este modo, la afirmación de que el enfoque simulacional tiene una buena explicación del acierto en la predicción de inferencias ajenas se extiende a *mindreading* exitoso en general. Así quedan formulados el primer aspecto del criterio de corrección: a *mindreading* con un resultado correcto le subyace un proceso de tipo simulacional, y el segundo

aspecto del criterio de corrección: a *mindreading* con un resultado incorrecto le subyace un proceso de tipo rico en información (me ocuparé del segundo aspecto de este criterio en la sección 4).

El argumento del acierto en *mindreading* ofrecido por Stich & Nichols, sin embargo, no resulta una buena inferencia a la mejor explicación respecto de *mindreading* exitoso. En principio, que el enfoque simulacional explique los casos de acierto en la predicción de inferencias ajenas, no descarta la posibilidad de que un enfoque de TT pueda explicar el acierto en otros aspectos de *mindreading*. En primer lugar, es necesario mencionar que el éxito sistemático en la predicción de inferencias ajenas, que es el punto de partida de este argumento así como del argumento de la simplicidad (Stich & Nichols 2003), es una especulación. Los autores están convencidos de que es posible que esto sea así en virtud de que, en la extensa literatura de psicología social cognitiva sobre los sesgos y heurísticos en el razonamiento deductivo e inductivo, no se menciona que se cometan errores al predecir las inferencias ajenas. Sin embargo, como ellos mismos advierten, esto no se ha estudiado sistemáticamente (Stich & Nichols 2003).

En segundo lugar, puede concederse que la explicación simulacional de la predicción de las inferencias ajenas sea la mejor explicación, y que una muestra de esto es que los enfoques de TT no se han ocupado de este aspecto de *mindreading* (Stich & Nichols 2003). Sin embargo, aún así resulta problemático extender a todos los casos de acierto en *mindreading* la afirmación de que un proceso simulacional subyace a los mismos. El enfoque de la TT está en mejores condiciones de explicar otros casos exitosos de *mindreading*, por ejemplo, casos en los que subyace una buena teoría. A continuación mostraré que la TT puede explicar satisfactoriamente el acierto en la predicción exitosa del comportamiento ajeno.

En las distintas versiones de la TT se postula que la capacidad de predecir el comportamiento ajeno está guiada por una psicología de deseos y creencias subyacente. Según el enfoque del niño científico, por ejemplo, la teoría que subyace a la predicción del comportamiento ajeno consiste en una concepción representacional de la mente, que se desarrolla paulatinamente durante la infancia. Se asume que la teoría subyacente a *mindreading* es análoga a las teorías científicas en tanto los conceptos mentalistas están relacionados entre sí de manera legaliforme de modo que permiten explicaciones causales, facilitan predicciones y otorgan capacidad de interpretación de la evidencia.

Además, estos están sujetos al cambio conceptual equiparable al que ocurre en las teorías científicas (Gopnik & Wellman 1992). 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. Esta caracterización de *mindreading* permite explicar ciertos fenómenos en el desarrollo de la comprensión mentalista.

Particularmente, este enfoque da cuenta de un hallazgo fundamental en el desarrollo de *mindreading*. Los niños menores de 3 años fallan sistemáticamente en la tarea de falsa creencia (TFC), mientras que alrededor de los 4 años comienzan a desempeñarse exitosamente (Wimmer & Perner 1983, Wellamn *et al*. 2001). En esta tarea, se narra una historia en la que un personaje coloca un objeto en cierta ubicación y, en su ausencia, se mueve de lugar el objeto. Llegada esta instancia, se le pregunta al participante dónde buscará el personaje el objeto cuando regrese. Para desempeñarse exitosamente en la TFC, el participante debe entender que el personaje cree que el objeto está aún donde él lo dejo. Esto es, el participante debe entender que el personaje tiene una creencia errónea, es decir, una "falsa creencia".

Así, se asume que para desempeñarse exitosamente en la TFC es preciso comprender que las personas poseen creencias falsas. Según el enfoque de niño científico, la teoría psicológica representacional que permite la comprensión de las creencias falsas se adquiere alrededor de los 4 años. Antes, los niños poseen una teoría psicológica rudimentaria que les permite comprender que los estados internos causan el comportamiento, pero estos no se conciben aún como estados representacionales. Alrededor de los 2 años, los deseos se entienden como un impulso hacia los objetos (Wellman & Woolley 1990) y la percepción como "darse cuenta" de los objetos (Flavell 1988). De este modo, los niños tratan a las percepciones y los deseos como vínculos causales simples con el mundo. Si alguien desea un objeto, actuará para obtenerlo. Si un objeto está en su campo visual, el agente lo ve. Estos constructos causales simples tienen poder predictivo en tanto 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. Sin embargo, no es suficiente para pasar la TFC.

Si bien en este momento del desarrollo los niños son capaces de advertir que hay una relación entre los estados internos y el mundo, esta

relación se considera directa o simple en el sentido de que el estado interno y el mundo concuerdan. De modo que aún no es posible advertir que un estado interno puede representar el mundo incorrectamente. En la TFC, los participantes menores de 3 años ven que el objeto es movido de lugar en ausencia del personaje, sin embargo, cuando reportan dónde buscará el personaje el objeto, fallan sistemáticamente y reportan el lugar donde el objeto está de hecho. A partir de 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, entienden que las personas pueden tener deseos diferentes. Sin embargo, aún no son capaces de comprender que las creencias pueden representar el mundo de manera incorrecta. Recién a los 4 años, son capaces de advertir que las acciones no están determinadas por el mundo, sino por la representación que el agente tiene del mundo. En este momento del desarrollo, se asume que la teoría central de los niños se reorganiza como una psicología representacional de los estados mentales en general, que posibilita el desempeño exitoso en TFC. Posteriormente, esta teoría se conserva y se seguirá sofisticando en la adultez.

De este modo, postular una teoría psicológica subyacente a *mindreading* permite dar cuenta del desarrollo de la comprensión mentalista en términos de una teoría rudimentaria subyacente, que está sujeta al cambio conceptual y que cambia hasta llegar a ser una teoría completamente desarrollada. Esto da cuenta de un hallazgo en el desarrollo de *mindreading*. Alrededor de los 4 años, cuando la teoría se ha desarrollado completamente, los niños comienzan a desempeñarse exitosamente en la TFC. Esto es, a realizar predicciones exitosas del comportamiento ajeno. Ahora bien, puesto que el enfoque del niño científico da cuenta del acierto en la predicción de comportamiento ajeno como una capacidad guiada por una buena teoría, no parece ser el caso que la explicación simulacional sea la mejor explicación del acierto en *mindreading*.

4. El criterio de *mindreading* incorrecto

A partir del argumento de la penetrabilidad cognitiva, formulado en los comienzos del debate teoría-simulación, se asume que las predicciones sistemáticamente fallidas son el producto de la posesión de una teoría psicológica parcialmente equivocada o incompleta, que conduce a predicciones

erróneas (Stich & Nichols 1992, 1995, 1996). 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.

Así, según el argumento de la penetrabilidad cognitiva, los errores sistemáticos en *mindreading* se explican como predicciones incorrectas generadas por una teoría incompleta o parcialmente errónea. En cambio, un enfoque simulacional no dispone de una explicación para los mismos, en tanto la información psicológica resulta irrelevante para llevar a cabo *mindreading*. Así, se asume que es probable que el proceso subyacente a *mindreading* con un resultado erróneo sea de tipo rico en información, en virtud de que no parece haber otro modo de explicar las fallas sistemáticas en *mindreading*. En este sentido, el segundo aspecto del argumento de corrección, *i.e.* si sistemáticamente nos equivocamos en *mindreading* es probable que el proceso subyacente sea de tipo de bases de información, concuerda con lo que ya se afirma en el argumento de la penetrabilidad cognitiva. No obstante, aquí sostendré que es posible que un proceso de tipo simulacional subyazca a ciertas fallas sistemáticas en las atribuciones mentalistas, de modo que éstas no sólo se explican por el recurso a cuerpos de información erróneos o incompletos.

Usualmente, se asume que si *mindreading* se lleva a cabo por simulación esto sólo puede conducir a *mindreading* correcto en virtud de estar utilizando un único tipo de mecanismo. Más precisamente, al tratarse de un sistema relevantemente similar al del blanco que es alimentado por *inputs* relevantemente similares a los de éste, es esperable que el resultado de *mindreading* sea correcto. No obstante, considero que es posible que la simulación conduzca a *mindreading* con un resultado incorrecto. El recurso a un sistema relevantemente similar al del blanco no garantiza la corrección de *mindreading*. Particularmente, dado el funcionamiento no estándar del sistema cognitivo en la simulación, en principio, es posible que los sesgos que operan durante el funcionamiento normal (*online*) del sistema de toma de decisiones no se recluten en la simulación mental. Esto podría tener lugar de la siguiente manera.

Se asume que el sistema cognitivo reclutado en la simulación *off-line* funciona de manera no estándar (Stich & Nichols 1992, 1995, 1996, 2003). Más precisamente, el sistema cognitivo reclutado opera desacoplado de los sistemas motores de modo que el *output* correspondiente se utiliza para realizar una predicción del comportamiento ajeno, y no para llevar a cabo una acción. A su vez, el sistema reclutado en la simulación es alimentado por *inputs* ficticios, que son relevantemente similares a los del blanco. En este sentido, ambos elementos, el desacople y los *inputs* ficticios, contribuyen al funcionamiento no estándar del sistema cognitivo cuando es reclutado para *mindreading*.

En principio, 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, la predicción del comportamiento ajeno, que recluta al sistema de toma de decisiones (STD) para funcionar de manera *off-line*, resulta un ejemplo paradigmático de *mindreading* simulacional (Goldman 1992, 2006, Stich & Nichols 1992, 2003, 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, que a menudo las personas toman decisiones inesperadas o irracionales. En relación a esto, los teóricos de *mindreading* sostienen que no es algo preocupante para la propuesta simulacional. 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 propias y para predecir las ajenas (Stich & Nichols 1992). Estas peculiaridades son los sesgos irracionales que, según la literatura de psicología social cognitiva, operan usualmente sobre la toma de decisiones, aunque también sobre las inferencias y las atribuciones de estados mentales, y dan lugar a efectos como, por ejemplo, el efecto de posición (Nisbett & Ross 1980)[3].

[3] El efecto de posición alude al siguiente fenómeno. En lo que parece ser una encuesta de opinión del consumidor, se invita a los participantes del experimento a examinar una serie de productos, por ejemplo, pijamas o medias. Se ofrece una recompensa por la participación, pueden conservar la prenda que elijan. Al examinar los productos, no parecen encontrarse diferencias significativas entre los mismos. Los ítems son todos iguales pero los participantes desconocen esto. Cuando se les pregunta cuál eligen, se

Sin embargo, no es claro que del reclutamiento del STD para simular una decisión ajena se siga que, si el proceso decisorio implica la operación de algún sesgo en su modo de funcionamiento normal, este sesgo también intervenga en la decisión que se genera vía la simulación. En principio, no disponemos de modelos acerca de 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 que el STD funciona de manera no estándar cuando es reclutado para la simulación *off-line*, en principio, hay razones para suponer que los sesgos pueden no reclutarse en este caso. Como consecuencia, el recurso a un mismo sistema para tomar decisiones y para predecir decisiones no garantiza que el resultado de *mindreading* sea correcto. La corrección de *mindreading* implica la coincidencia entre la decisión generada en el blanco y la predicción de la acción basada en la decisión generada por el STD del *mindreader* reclutado en la simulación. Si bien menciono aquí el caso específico de la toma de decisiones, que es el caso paradigmático de la simulación mental, esto quizás pueda extenderse a las otras capacidades cognitivas consideradas como reclutables en la simulación mental, a saber, los sistemas inferenciales y de atribución, que también se ven afectados, por ejemplo, por sesgos egocéntricos[4].

El diagrama de flujo utilizado por los teóricos de *mindreading* para caracterizar la toma de decisiones ordinaria, sobre el cual se modela *mindreading* por simulación (ver la figura 1), es muy esquemático (Stich & Nichols 1992, 1995b, 1997, Gallese & Goldman 1998, Goldman 2006)[5]. En virtud de la

---

espera que la mayoría de los participantes reporten que son todos similares y elijan al azar. Sin embargo, sorprendentemente, en el experimento se observa un efecto de posición en las evaluaciones. Las prendas ubicadas más a la derecha son mucho más preferidas que las ubicadas a la izquierda. No obstante, los participantes no advierten este efecto de posición en sus decisiones.

[4] Cierta evidencia sugiere que al realizar juicios sobre lo que otras personas creen o conocen, a menudo, se parte del conocimiento y las creencias propias para luego ajustar esto, con esfuerzo, al conocimiento y las creencias del blanco (*e.g.* Nickerson 1999). En este sentido, se sugiere la presencia de "sesgos egocéntricos" en las atribuciones a otras personas.

[5] Stich & Nichols son claros en este respecto cuando sostienen que "Los diagramas son considerados esquemas rudimentarios de algunos de los mecanismos y procesos

generalidad con la que los teóricos de *mindreading* describen el proceso de toma de decisiones y en la medida en que no se dispone de propuestas acerca de en qué momento del proceso de toma de decisiones operan los sesgos, en principio, todas las posibilidades quedan abiertas. En este sentido, puede ser el caso que los sesgos operen (i) a nivel del *input* del STD, (ii) a nivel del *output* del STD, o bien, (iii) al interior del STD. Ahora bien, como señalé, se postula el funcionamiento no estándar del sistema cognitivo cuando es reclutado en la simulación, en la medida en que:

(A) el sistema cognitivo funcionará desacoplado (modo *off-line*)
(B) el sistema es alimentado por *inputs* ficticios.

En base a esto, si es el caso que (iii) los sesgos operan al interior del sistema cognitivo, puesto que en la simulación mental se recluta el sistema cognitivo, es plausible que los sesgos también sean reclutados en la simulación y que el resultado de *mindreading* no se vea afectado. Esto conducirá a la concordancia entre el *output* de *mindreading* y el *output* del blanco. De este modo, de acuerdo con Stich & Nichols (1992), puede asumirse que la simulación conducirá a *mindreading* con un resultado correcto. Sin embargo, los casos (i) y (ii) resultan problemáticos para afirmar la concordancia entre el *output* generado por el STD en el blanco y el *output* del STD reclutado en el *mindreader*, porque los modos (i) y (ii) pueden considerarse externos al STD.

Si es el caso (i) el sesgo operaría a nivel del *input*. Es preciso notar que, por definición de la simulación *off-line*, (B) el sistema cognitivo es alimentado por *inputs* ficticios que son generados por algún generador de *inputs* ficticios (Stich & Nichols 1997). Si bien no está claro qué implica esto último, al menos parece que el *input* en la simulación no se genera por la misma vía que los *inputs* del funcionamiento *online*. En este sentido, el *input* no sería estándar. Ahora bien, puesto que no disponemos de modelos acerca del momento en el que los sesgos operan en el proceso decisorio, si fuera el caso que algún sesgo operara a nivel del *input* y dado que los *inputs* no son estándar sino ficticios, en principio, hay razones para afirmar que el resultado de *mindreading* se verá afectado. A mi

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, p. 305).

entender, si el sesgo opera a nivel del *input* en el funcionamiento *online*, en principio no hay razones para afirmar que el sesgo también operaría sobre el *input* no estándar de la simulación, ya que al ser un *input* no estándar no se trataría del mismo tipo de *input* que alimenta al sistema en su funcionamiento *online*. Así, puede ser el caso que el sesgo que opera a nivel del *input* en el modo de funcionamiento *online*, no opere sobre los *inputs* ficticios. Si esto es así, entonces el sesgo no sería reclutado para la simulación y no habría concordancia entre el *output* de *mindreading* y el *output* del blanco. De modo que, de darse este caso, el resultado de *mindreading* no sería correcto.

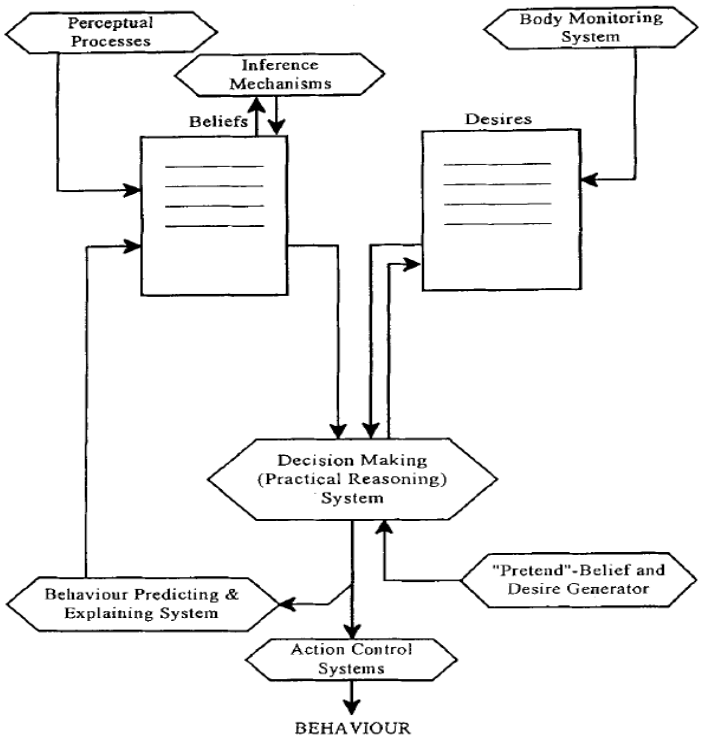

Figura 1. Diagrama de flujo que caracteriza la toma de decisiones ordinaria. Extraído de Stich & Nichols (1997), p. 303.

De manera similar, surgiría un problema si se diera el caso (ii). Los sesgos operarían a nivel del *output*, esto es, una vez que el STD arrojó su resultado. Dado que el sistema cognitivo reclutado para la simulación opera desconectado de los sistemas controladores de la acción (A), es posible pensar que las operaciones que se llevan a cabo sobre el *output* producto del funcionamiento *online* pueden no tener lugar en el caso del funcionamiento *off-line*. Puede ser el caso que el efecto de los sesgos tenga lugar en los pasos de procesamiento que se siguen en la vía *online* y hasta alimentar los sistemas controladores de la acción. Esto no tiene que ocurrir también en la vía *off-line* que lleva a la predicción. Podría pensarse que los sesgos operan después de que el STD hizo lo propio pero justo antes de que el *output* alimente como *input* a los sistemas controladores de la acción. En este sentido, el funcionamiento desacoplado implica que no hay razones para afirmar que los sesgos aportarán su efecto en la simulación.

Así, el recurso a un mismo sistema no garantiza que el resultado de *mindreading* sea correcto, como se sostiene usualmente. En otras palabras, si la influencia de los sesgos quedara por fuera del funcionamiento del STD, en virtud de que la simulación sólo recluta al sistema cognitivo, podría ser el caso que el sesgo no sea reclutado para simulación. Sin embargo, es necesario que los sesgos sean reclutados para que su aporte se incluya en el resultado de *mindreading*. De lo contrario, no habrá concordancia entre el *output* del STD del blanco afectado por los sesgos y el *output* de *mindreading*, porque éste no se ve afectado por los sesgos que no han sido reclutados. De modo que, de darse este caso, el resultado de *mindreading* no resultaría correcto.

Así, si *mindreading* tiene lugar mediante un proceso de tipo simulacional que recluta al STD sin la intervención de los sesgos, el resultado 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 usualmente este se ve afectado por los sesgos. En este sentido, considero que 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, por ejemplo obviar la existencia de los sesgos cognitivos, y conducir a predicciones erróneas (Stich & Nichols 1992). En este sentido, no basta con el reclutamiento de un mismo sistema cognitivo para que *mindreading* se ejecute con éxito, porque el resultado correcto de

mindreading requiere de la concordancia entre el producto del sistema cognitivo del blanco y el del sistema cognitivo del *mindreader*.

De acuerdo con el análisis realizado, ninguno de los aspectos del criterio de corrección resulta satisfactorio. Por un lado, la afirmación de que la simulación subyace a *mindreading* con resultado exitoso no se sostiene en tanto la simulación no resulta la mejor explicación disponible para todos los aciertos de *mindreading*. La predicción acertada del comportamiento ajeno puede estar basada en una teoría psicológica subyacente. 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 el caso del reclutamiento del STD también deben reclutarse los sesgos cognitivos a los que está sujeta la toma de decisiones, pero puede ser el caso que los mismos no sean reclutados. En este sentido, el criterio de corrección propuesto por Stich & Nichols (2003, Nichols & Stich 2003) no resulta satisfactorio para distinguir entre tipos de procesos subyacentes a *mindreading*.

6. Conclusión

En los enfoques híbridos de *mindreading* se asume que teoría y simulación no están en competencia sino que, más bien, ambos procesos contribuyen a la capacidad. De modo que surge la cuestión de cómo dos tipos distintos de procesos pueden subyacer a una misma capacidad cognitiva. Las propuestas híbridas de *mindreading* disponibles hasta el momento no describen satisfactoriamente la relación entre los procesos subyacentes (Nichols & Stich 2003, Goldman 2006). No obstante, en la propuesta de Nichols & Stich (2003, Stich & Nichols 2003) se sugieren dos formas de colaboración entre los mismos. Por un lado, una relación de cooperación en la que los dos tipos de procesos participan, a la vez, para que tenga lugar *mindreading*, que está ampliamente descripta y desarrollada. Por otro lado, una relación de alternancia entre los dos tipos de procesos, que no está suficientemente descripta, ni desarrollada. Esta última está regida por un criterio de corrección según el cual, a los casos *mindreading* con resultado correcto les subyace un proceso de tipo simulacional (primer aspecto del criterio de corrección), y a los casos de

mindreading con un resultado incorrecto les subyacen procesos ricos en información (segundo aspecto del criterio de corrección).

En este trabajo me he ocupado principalmente de esta relación de alternancia y de si el criterio de corrección en que ésta se basa permite distinguir efectivamente entre tipos de procesos en una teoría híbrida de *mindreading*. En contra del primer aspecto, sostuve que los argumentos de simplicidad y de acierto en *mindreading* no son buenos argumentos. En relación con el primer argumento, sostuve que adoptar un enfoque de TT no implica necesariamente postular teorías reduplicadas. De modo que no es el caso que la explicación simulacional que postula un único mecanismo sea más simple que la explicación provista por un enfoque de TT. En relación con el argumento del acierto en *mindreading*, sostuve que no es el caso que el enfoque simulacional sea la mejor explicación disponible para el acierto en *mindreading*. Un enfoque rico en información está en mejores condiciones de brindar una explicación de ciertos casos relevantes de acierto en *mindreading*, por ejemplo, la predicción del comportamiento ajeno.

Contra el segundo aspecto del criterio de corrección, sostuve que puede ser el caso que un enfoque de simulación pueda dar cuenta de *mindreading* incorrecto. A saber, dado el funcionamiento no estándar del sistema cognitivo reclutado en la simulación, hay razones para pensar que si los sesgos operan de manera externa al STD, esto es a nivel del *input* o del *output*, es probable que no se recluten en la simulación, puesto que en la simulación sólo se recluta el sistema cognitivo. Si este es el caso, es probable que el *output* de *mindreading* no concuerde con el *output* del sistema cognitivo del blanco, porque los sesgos no son reclutados en la simulación. En este sentido, en contra de Stich & Nichols (2003, Nichols & Stich 2003), un proceso de tipo simulacional puede generar *mindreading* con un resultado incorrecto.

Referencias

APPERLY, I. "Beyond simulation-theory and theory-theory: why social cognitive neuroscience should use its own concepts to study 'theory of mind'". *Cognition*, 107, pp. 266-283, 2008.

ASTINGTON, J., HARRIS, P., OLSON, D. *Developing theories of Mind.* NY: Cambridge University Press, 1988.

CARRUTHERS, P. "How we know our own minds: The relationship between mindreading and metacognition". *Behavioral and Brain Sciences*, 32, pp. 1-62, 2009.

——— AND SMITH, P. *Theories of theories of mind*. Cambridge: Cambridge University Press, 1996.

DAVIES, M. AND STONE, T. *Folk Psychology: The Theory of Mind Debate*, Oxford: Blackwell, 1995a.

——— *Mental Simulation: Evaluations and Applications*, Oxford: Blackwell, 1995b.

DAVIES, M. AND STONE, T. "Folk Psychology and Mental Simulation". In A. O'Hear (ed.) (1998), pp. 53-82.

FEIGL, H., SCRIVEN, M. *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, 1956.

FLAVELL, J. H. "The development of children's knowledge about the mind: from cognitive connections to mental representations". In J. Astington, P. Harris and D. Olson (eds.) (1988), pp. 244-267.

FODOR, J. *Psychosemantics*. Cambridge: MIT Press, 1987.

GALLESE, V. AND GOLDMAN, A. "Mirror neurons and the simulation theory of mind-reading". *Trends in Cognitive Science*, 3, pp. 493-501, 1998.

GAZZANIGA, M. *The New Cognitive Neurosciences*. Cambridge: Cambridge University Press, 2000.

GOLDMAN, A. "Interpretation Psychologized". In M. Davies, and T. Stone (eds.) (1995a), pp. 60-73.

——— *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. Oxford: Oxford University Press, 2006.

GOPNIK, A. "How we know our own minds: the illusion of first-person knowledge of intentionality". *Behavioral and Brain Sciences*, 16, pp. 1-14, 1993.

——— AND ASTINGTON, J.W. "Children's understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction". *Child Development*, 59(1), pp. 26-37, 1988.

——— AND MELTZOFF, A. "Minds, bodies, and persons: young children's understanding of the self and others as reflected in imitation and theory of mind research". In S. Parker, R. Mitchell and M. Boccia (eds.) (1994), pp. 166-186.

——— AND MELTZOFF, A. *Words, Thoughts and Theories*. Cambridge: Cambridge University Press, 1997.

——— AND WELLMAN, H. "Why the child's theory of mind really is a theory". *Mind and Language*, 7 (1-2), pp. 145-71, 1992.

GORDON, R. "Folk psychology as simulation". *Mind and Language*, 7 (1-2), pp. 11-34, 1992.

HARRIS, P. "From simulation to folk psychology: the case for development". *Mind and Language*, 7 (1-2), pp. 120-144, 1992.

HAUGELAND, J. *Artificial Intelligence: The Very Idea*. Cambridge, Mass.: The MIT Press, 1985.

HEAL, J. "Replication and Functionalism". In M. Davies and T. Stone (eds.) (1995a), pp. 45-59.

KAHNEMAN, D., SLOVIC, P. AND TVERSKY, A. *Judgment under uncertainty: heuristics and biases*. NY: Cambridge University Press, 1982.

LESLIE, A. "Pretense and representation: the origins of 'theory of mind'".*Psychological Review*, 94 (4), pp. 412-426, 1987.

——— ""Theory of mind' as a mechanism of selective attention". In M. Gazzaniga (ed.) (2000), pp. 1235-1248.

LEWIS, D. "How to define theoretical terms". *The Journal of Philosophy*, 67 (13), pp. 427-446, 1970.

——— *Papers in Metaphysics and Epistemology*. Cambridge: Cambridge University Press, 1999.

——— "Psychophysical and Theoretical Identifications". The Australasian Journal of Philosophy, 50, 1972. Repr. in D. Lewis (1999), pp.248-261.

MACHERY, E. *Doing without concepts*. Oxford: University Press, 2009.

NICHOLS, S. AND STICH, S. *Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding of Other Minds*. Oxford: Oxford University Press, 2003.

———, LESLIE, A. AND KLEIN, D. "Varieties of off-line simulation". In P. Carruthers and P. Smith (eds.) (1996), pp. 39-74.

NICKERSON, R. S. "How we know -and sometimes misjudge- what others know: imputing one's own knowledge to others". *Psychological Bulletin*, 125, pp. 737-759, 1999.

NISBETT, R. AND ROSS, L. *Human inference: strategies and shortcomings of social judgment*. New Jersey: Prentice-Hall, 1980.

O´HEAR, A. Current Issues in Philosophy of Mind. Cambridge: Cambridge University Press, 1998.

PARKER, S., MITCHELL, R., BOCCIA, M. *Self- Awareness in animals and Humans*. NY: Cambridge University Press, 1994.

PERNER, J. "Simulation as explicitation of predication-implicit knowledge about the mind: arguments for a simulation-theory mix". In P., Carruthers and P., Smith (eds.) (1996), pp. 90-104.

SCHOLL, B.J. AND LESLIE, A. "Mind, modules and meta-analysis". *Child Development*, 72(3), pp. 696-701, 2001.

SELLARS, W. "Empiricism and the Philosophy of Mind". In H. Feigl and M. Scriven (eds.) (1956), pp. 253-329.

STICH, S. *Deconstructing the Mind*. Oxford: Oxford University Press, 1996.

——— AND NICHOLS, S. "Folk psychology: simulation or tacit theory?". *Mind and Language*, 7(1-2), pp. 35-71, 1992.

——— "Second thoughts on simulation". In M. Davies and T. Stone (eds.) (1995b), pp. 87-108.

——— "How do minds understand minds? Mental Simulation versus Tacit Theory". In S. Stich (ed.) (1996), pp. 136-167.

——— "Cognitive penetrability, rationality and restricted simulation". *Mind and Language*, 12, pp. 297-326, 1997.

——— "Folk Psychology". In S. Stich and T. Warfield (eds.) (2003), pp. 253-255.

——— AND WARFIELD, T. *The Blackwell Guide to Philosophy of Mind*. Malden, Oxford: Blackwell Publishing Ltd, 2003.

WELLMAN, H. AND WOOLEY, J. "From simple desires to ordinary beliefs: the early development of everyday psychology". *Cognition*, 35, pp. 245-273, 1990.

WELLMAN, H., CROSS, D. AND WATSON, J. "Meta-analysis of theory-of-mind development: the truth about false belief". *Child Development*, 72(3), pp. 655-684, 2001.

WIMMER, H. AND PERNER, J. "Beliefs about Beliefs". *Cognition*, 13, pp. 103-128, 1983.