PABLO QUINTANILLA

LA COMPRENSIÓN DEL OTRO

Explicación, interpretación y racionalidad

Pablo Quintanilla es profesor principal de filosofía en la Pontificia Universidad Católica del Perú (PUCP). Es PhD en Filosofía por la Universidad de Virginia y magíster en la misma especialidad por la Universidad de Londres (King's College). Es licenciado en Filosofía y bachiller en Humanidades con mención en Filosofía por la PUCP. Se especializa en filosofía del lenguaje y de la mente, epistemología y teoría de la acción. Ha sido decano de Estudios Generales Letras de la PUCP entre los años 2011 y 2017. Es miembro de diversas sociedades académicas internacionales, como el Grupo Interdisciplinario de Investigación Mente y Lenguaje. Ha recibido becas de investigación del Consejo Británico, la Fundación Fulbright, el Instituto Riva-Agüero y la PUCP.

Es coeditor de *Los caminos de la filosofía. Diálogo y método* (2018); *El desarrollo de las competencias genéricas en los Estudios Generales* (2017); *El pensamiento pragmatista en la actualidad: conocimiento, lenguaje, religión, estética y política* (2015); *Pedro Zulen: escritos reunidos* (2015); *Cognición social y lenguaje. La intersubjetividad en la evolución de la especie y en el desarrollo del niño* (2014); *Lógica, lenguaje y mente* (2012); *Desarrollo humano y libertades. Una aproximación interdisciplinaria* (2009). Es editor de *Estudios Generales Letras: lecciones inaugurales 2012-2017* (2017); *Ensayos de metafilosofía* (2009). Es coautor de *Pensamiento y acción. La filosofía peruana a comienzos del siglo XX* (2009).

LA COMPRENSIÓN DEL OTRO

Explicación, interpretación y racionalidad

Pablo Quintanilla

LA COMPRENSIÓN DEL OTRO

Explicación, interpretación y racionalidad

FONDO EDITORIAL
PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

Quintanilla, Pablo, 1964-
La comprensión del otro: explicación, interpretación y racionalidad / Pablo Quintanilla.-- 1a ed.-- 1a reimpr.-- Lima: Pontificia Universidad Católica del Perú, Fondo Editorial, 2019 (Lima: Tarea Asociación Gráfica Educativa).
374 p.; 24 cm.

Contenido: Encontrar algo inteligible -- Interpretar significados y metáforas -- La naturaleza de la comprensión -- La racionalidad y sus límites.
Bibliografía: p. 347-374.
D.L. 2019-14944
ISBN 978-612-317-474-3

1. Comprensión (Teoría del conocimiento) - Ensayos, conferencias, etc. 2. Semántica (Filosofía) 3. Racionalismo 4. Irracionalismo I. Pontificia Universidad Católica del Perú II. Título

La comprensión del otro
Explicación, interpretación y racionalidad

Av. Universitaria 1801, Lima 32, Perú
feditor@pucp.edu.pe
www.fondoeditorial.pucp.edu.pe

Diseño, diagramación, corrección de estilo
y cuidado de la edición: Fondo Editorial PUCP

Imagen de portada: detalle de la pintura *Paraíso*, de Giovanni di Paolo, 1445. Donación del Fondo Rogers al Museo Metropolitano de Arte (The Met), Nueva York, en 1906.

Primera edición: abril de 2019
Primera reimpresión de la primera edición: octubre de 2019
Tiraje: 500 ejemplares

Hecho el Depósito Legal en la Biblioteca Nacional del Perú Nº 2019-14944
ISBN: 978-612-317-474-3
Registro del Proyecto Editorial: 31501361901080

Impreso en Tarea Asociación Gráfica Educativa
Pasaje María Auxiliadora 156, Lima 5, Perú

Índice

Lista de abreviaturas

CP: Peirce, Charles Sanders (1931-1958). *Collected Papers of Charles Sanders Peirce.* Volúmenes 1-6 (1931-1935) editados por Charles Hartshorne y Paul Weiss. Volúmenes 7 y 8 (1958) editados por Arthur W. Burk. Massachusetts: Harvard University Press.

EP: Peirce, Charles Sanders (1992-1998). *The Essential Peirce. Selected Philosophical Writings*. Volúmenes 1-2. Edición de Nathan Houser y otros. Indiana: Indiana University Press.

W: Peirce, Charles Sanders (1982-2000). *Writings of Charles S. Peirce: A Chronological Edition*. Edición de M.H. Fisch y otros. Seis volúmenes. Indiana: Indiana University Press.

Reconocimientos

Algunos capítulos de este libro son inéditos y otros son versiones totalmente reescritas, corregidas y actualizadas de artículos o de fragmentos de artículos publicados previamente por mí, los cuales han sido reestructurados para que mantengan ilación temática. Señalo las referencias bibliográficas de los artículos originales que he tomado como base y que han sido transformados. Capítulo uno: inédito. Capítulo dos: inédito, aunque algunos párrafos han sido tomados de Quintanilla (2017b). Capítulo tres: basado en Quintanilla (2002b). Capítulo cuatro: inédito, aunque algunos párrafos han sido tomados de Quintanilla (1994). Capítulo cinco: basado en Quintanilla (1997). Capítulo seis: basado en Quintanilla (1995, 1999 y 2009b). Capítulo siete: basado en Quintanilla (2004). Capítulo ocho: basado en Quintanilla (2008a). Capítulo nueve: basado en Quintanilla (2014c). Capítulo diez: basado en Quintanilla (2005). Capítulo once: basado en Quintanilla (2001a). Capítulo doce: basado en Quintanilla (2014d).

Prólogo

Los temas tratados en este libro no son solo interesantes por sí mismos sino también porque tienen consecuencias para las diversas ciencias humanas y sociales, así como para la ética y la vida política. Las principales preguntas abordadas son: ¿qué es comprender a un ser humano o a una comunidad humana? ¿Cuándo dos personas o grupos de personas se comprenden y cuándo se malentienden sistemáticamente? ¿Cómo se relacionan la comprensión, la explicación y la interpretación del comportamiento humano? ¿Es posible explicar el comportamiento de alguien sin poder comprenderlo o comprender a una persona sin que podamos explicar su comportamiento? ¿Mediante qué mecanismos neurológicos, psicológicos y culturales atribuimos estados mentales a otros individuos para poder comprenderlos? ¿Cómo comprendemos las expresiones lingüísticas de una persona? ¿Qué es el significado y cómo emerge? ¿Qué es comprender una metáfora? ¿Cuál es el rol que la racionalidad y la irracionalidad cumplen en la comprensión? ¿Qué criterios empleamos para determinar cuándo una interpretación del comportamiento humano es preferible a otra? ¿Cuándo lo que parece comprensión es solo una forma enmascarada de sometimiento? El libro aborda también otros temas, como la diversidad de formas de vida y los problemas del etnocentrismo y el relativismo, en la medida en que su análisis es relevante para las preguntas principales que nos atañen. Como no puede ser de otra manera, estos temas son tratados de forma interdisciplinaria, aunque la estructura argumentativa filosófica es la que hilvana las diversas perspectivas y temáticas.

Este libro es el producto de investigaciones realizadas en el King's College de la Universidad de Londres y en la Universidad de Virginia, gracias a sendas becas de posgrado otorgadas por el Consejo Británico y la Comisión Fulbright, respectivamente. Agradezco a ambas instituciones. Posteriormente, el libro adquirió forma gracias a un semestre de investigación otorgado por la Pontificia Universidad Católica del Perú, a la cual también agradezco especialmente. Son muchas las personas con quienes he conversado sobre estos temas y de quienes he aprendido, pero la lista es tan larga que sería inútil intentar confeccionarla. Quiero formular, sin embargo,

dos agradecimientos principales. Por una parte, a varias generaciones de alumnos —muchos de ellos ahora colegas y amigos— que durante años en diversos cursos de filosofía del lenguaje y de la mente, con sus preguntas, objeciones y comentarios me han permitido pulir las ideas que aquí presento. Por otra parte, a mis amigos y colegas del Grupo Interdisciplinario de Investigación Mente y Lenguaje, con quienes he compartido aciertos y errores por más de once años de constante trabajo. Pero lo más importante es que el libro está dedicado a Lucía, Juan Diego y Álvaro, por su cálida e invalorable presencia en mi vida, sin cuya comprensión esta sería mucho menos plena.

Introducción

Por mucho que andes, y aunque paso a paso recorras todos los caminos,
no hallarás los límites del alma.

Heráclito, fragmento 45, citado por Juan David García Barca, 2009

Los europeos y los occidentales hallan siempre el misterio en la oscuridad,
en la noche, mientras nosotros los griegos lo hallamos en la luz.

Odysseas Elytis, entrevista concedida a Ivar Ivask, 1975

Las palabras «comprensión», «explicación», «interpretación» y «racionalidad» —así como otras asociadas a ellas como «entendimiento» o «intelección»— tienen una pesada carga filosófica, es decir, han sido tantas veces usadas, analizadas y comentadas que es difícil emplearlas sin sugerir alguna posición teórica implícita. Por otra parte, no tenemos por qué suponer que cada una de ellas esconda un significado principal que sea nuestra tarea reconstruir o desocultar. Es posible que estos términos tengan múltiples significados con solo un parecido de familia entre ellos, para aludir a la famosa metáfora de Ludwig Wittgenstein (1988, pp. 483-485). Sin embargo, de una manera bastante general e imprecisa, quizá haya algo que tienen en común: entender, explicar, comprender o dar sentido a algo —sea un fenómeno físico, un proceso histórico o social, el comportamiento de una persona, un texto, un conjunto de signos, un resto arqueológico— es un proceso psicológico mediante el cual uno encuentra el orden que sostiene a un aparente desorden o la estructura subyacente que hace posible un real desorden. Eventualmente también puede ser proyectar un orden en una realidad carente de este. Incluso un relato o una reconstrucción histórica llevan implícitas una búsqueda de inteligibilidad, porque se proponen encontrar u otorgar sentido a una masa de información que podría parecer caótica.

Es una mente la que encuentra o proyecta orden en un supuesto desorden preexistente, porque normalmente el desorden es solo un orden desconocido. Lo anterior puede también ser visto como el proceso por el cual encontramos o construimos la racionalidad que subyace a lo que nos parece enigmático, sea esta una penumbra que necesita ser iluminada o un resplandor que nos obliga a entrecerrar los ojos para poder divisar con más precisión y claridad, porque lo misterioso y desconocido no es solo lo que nos parece oscuro, también puede ser lo que encandila y por su luminosidad nos resulta tan perspicuo que nos sorprende.

Explicar el universo físico, por ejemplo, es descubrir una regularidad subyacente que lo gobierna y de la que somos parte, la cual suele estar oculta, aunque implícita, en lo que observamos de él. Comprender a una persona o a una comunidad de personas —sus acciones, proferencias verbales o estados mentales— es también encontrar una estructura que los moldea —sea individual o colectivamente— a veces de manera consciente, con frecuencia de forma no consciente y en ocasiones inconscientemente. Aquello puede ocurrir en esos tres casos tanto de manera voluntaria como involuntaria, porque hay formas de comportamiento voluntarias que pueden ser inconscientes o no conscientes. Entender un signo, un significado o un texto, asimismo, es hallar la estructura que relaciona sus diversos elementos entre sí, con el contexto y en relación a su autor.

Usaré el verbo «entender» para el proceso psicológico más amplio y general por el que damos sentido o hacemos inteligible a algo. Mientras que seguiré la tradición hermenéutica para hablar de «explicar» la naturaleza física y de «comprender» a las personas o grupos de personas o, en todo caso, a cualquier objeto o criatura dotada de subjetividad. Aunque este libro está dedicado a la comprensión, será necesario decir algo acerca de la explicación, lo que haré en el primer capítulo. Ahora me referiré brevemente a la actividad más general de entender o hacer inteligible algo, y a cómo la explicación y la comprensión son dos aspectos de ella que se encuentran intersecados.

Las metáforas que asocian el entendimiento con el proceso de hallar, construir o proyectar un orden en una realidad previa han sido exploradas por muchos filósofos —paradigmáticamente Kant y sus epígonos—, pero son ancestrales y están presentes en muchas culturas. En el griego antiguo la palabra *káos* designaba a la oquedad original, el vacío primordial o la primigenia ausencia de estructura, de la que surge el *kósmos*. *Káos* no siempre significa desorden, pero sí ausencia de un orden evidente. *Kósmos* significa orden, pero también connota buena organización, así como disposición bella y bien ornamentada. De hecho, la palabra «cosmético» proviene del griego *kosmetikós,* que tiene el sentido de adorno y está etimológicamente emparentada con *kósmos*. De manera análoga, la palabra latina *mundus* —de la que procede mundo—

originalmente significaba limpio, arreglado, organizado y ordenado, por oposición a *inmundus*, o inmundo, que es algo desordenado, caótico y sucio. Pero para que el orden sea inteligible debe poseer un logos, cierta racionalidad que podamos reconocer y compartir. Como es conocido, la filosofía occidental apareció en unas pequeñas poblaciones de las costas orientales del mar Egeo, con la pretensión de encontrar esa racionalidad que se manifestaba a través de aquella lógica.

Probablemente todas las comunidades humanas tengan una inclinación natural por buscar el orden de la realidad para poder predecirla, reaccionar ante ella y, de ser posible, modificarla. Esta realidad incluye tanto el entorno físico como el social. La necesidad de explicar la naturaleza física y el comportamiento de los grupos de individuos —así como el de los individuos en los grupos— se convirtió en una adaptación cognitiva que potenció el crecimiento y la complejidad de los cerebros de nuestros antepasados homínidos en los últimos tres millones de años, lo que nos convirtió progresivamente en científicos y psicólogos natos. Pero asumir o tener la esperanza de que existe un orden que subyace a la fragilidad de nuestra existencia no es solo una postura epistémica sino también una vital y existencial. Se trata de una actitud presente en el pensamiento griego antes de que la filosofía, la ciencia y la religión tomaran caminos diferentes. No es solo un presupuesto metafísico sino también un acto de fe.

Era y sigue siendo necesario para nuestra supervivencia predecir la regularidad de los elementos, pero también imaginar y adelantarnos al complejo comportamiento de los grupos de personas para poder cooperar o competir —según sea el caso— eficientemente. Aunque es posible que versiones rudimentarias de estas habilidades se encuentren en otras especies de animales, es indiscutible que la selección natural nos proporcionó una particular maestría en esos menesteres, y nos convirtió en una especie que tiene una tendencia irrefrenable por entenderlo todo, así como en una que puede cambiar significativamente la realidad y a sí misma dentro de ella. Quizá esta sea nuestra mayor ventaja comparativa respecto de las otras especies, lo que nos ha dado una posición de privilegio sobre ellas, aunque a veces en detrimento de nuestro entorno y de nosotros mismos. Análogamente, los grupos humanos que perfeccionaron su natural tendencia al conocimiento y a la explicación —ya sea como un fin en sí mismo o teniendo la dominación como objetivo— lograron ubicarse en condiciones ventajosas respecto de otras sociedades que no se embarcaron en esa carrera. Es innecesario decir que en muchos casos esto ha tenido consecuencias nefastas, pero ciertamente la solución no podría ser reprimir nuestra curiosidad investigativa sino, por el contrario, potenciarla para que sea nuestra creatividad racional la que elimine los monstruos que ella misma produjo.

El punto es que debemos asumir que nuestros objetos de explicación y comprensión tienen cierto orden o racionalidad, o debemos proyectarlos en ellos, para poder entenderlos. Eso ocurre tanto cuando explicamos objetos naturales como acciones individuales y procesos sociales. En el primer caso, asumimos lo que John Stuart Mill denominó «principio de la uniformidad de la naturaleza» (2002, libro III, capítulo 3, párrafo 1). En el segundo caso presuponemos lo que Donald Davidson (1984c, p. 27), basado en Willard Van Orman Quine (1960, p. 59), llamó «principio de caridad». En el tercer caso, asumimos una combinación de los dos principios anteriores. En los tres casos asumimos que el fenómeno por ser explicado no es aleatorio y que está gobernado por regularidades que contienen relaciones causales. En gran medida explicar algo es conocer esas relaciones causales. Cuando explicamos la naturaleza asumimos que esas regularidades existen de manera independiente del observador, con la discutible excepción de la física cuántica sobre la que aún no hay acuerdo al respecto. Al intentar comprender el comportamiento humano también presuponemos que está gobernado por relaciones causales independientes de él y del observador, aunque además necesitamos asumir que hay cierta dosis de libre albedrío que lo convierte en un agente y no solo en una pieza de la naturaleza. Es decir, suponemos que tiene propiedades como subjetividad y voluntad, que lo hacen capaz de iniciar relaciones causales nuevas de manera autoconsciente. Al explicar procesos sociales intentamos compatibilizar, con cierta dificultad, los presupuestos anteriores y solemos atribuir agencia no solo a los individuos sino también a los grupos de estos.

Lo que deseo subrayar es que «entender algo» o «encontrarlo inteligible», son nociones amplias y genéricas que apuntan a encontrar ese orden que subyace de manera implícita a lo manifiesto o, eventualmente, sugieren construir un orden que puede no estar en el objeto en sí mismo sino en la relación que nosotros tenemos con él. Como ya mencioné, seguiré el vocabulario filosófico habitual para denominar «explicación» a la intelección de la naturaleza y «comprensión» a la de los seres humanos. Así también llamaré «interpretación» a la metodología empleada para intentar comprender a un individuo o una comunidad. Pero explicar y comprender no son conjuntos disjuntos sino intersecados, y muchas disciplinas —sobre todo de origen reciente— se encuentran en esa intersección.

En líneas generales, explicamos la naturaleza buscando relaciones causales gobernadas por regularidades físicas. En el caso de la comprensión de seres humanos, les atribuimos estados mentales que asumimos han causado sus acciones, con el fin de compartir su subjetividad mediante complejos mecanismos interpretativos. Aunque hablaré algo sobre lo primero, este libro está dedicado especialmente a lo segundo.

Una primera tesis que defenderé es que, al comprender a una persona, el orden que construimos y hallamos está en la relación conformada por agente, intérprete y

mundo compartido, y no solamente en la persona interpretada. Una segunda tesis es que la comprensión del comportamiento intencional posee algunos rasgos de la explicación de los eventos físicos —la búsqueda de relaciones causales entre estados mentales y acciones— pero también incluye otros rasgos propios que son de una mayor complejidad y que tienen que ver con que su objeto está dotado de subjetividad. Por ello, una característica importante de la comprensión es que la estructura que se busca es el producto de una red tejida entre quien interpreta, el interpretado y el mundo que ambos comparten o asumen compartir. No es, por tanto, solamente algo que preexista a quien pretende encontrarlo sino es también un objeto construido en el fenómeno mismo de la interpretación. Eso no está presente en la explicación de la naturaleza.

Estudiar la comprensión exige un delicado análisis conceptual y empírico, que tiene como propósito aclarar las siguientes preguntas: ¿qué significa comprender a una persona o a una comunidad humana? ¿Qué acontece cuando dos personas o comunidades se comprenden mutuamente? ¿Qué ocurre y qué deja de ocurrir cuando se malentienden sistemáticamente, lo cual genera la impresión de que se están comprendiendo? ¿Quién determina —y desde qué punto de vista— cuándo dos personas o grupos se comprenden o se malentienden? ¿Qué metodologías interpretativas debemos emplear para lograr descripciones comprehensivas correctas y qué significa que estas lo sean? ¿Cuándo podemos decir que hemos logrado comprender a alguien, cuándo creemos que lo hemos comprendido —aunque solo hayamos proyectado nuestros propios prejuicios en él o ella— y cuándo lo que pasa por comprensión es solo una forma de manipulación o de sometimiento enmascarado? ¿Hay distintas maneras, simultáneamente válidas y complementarias entre sí, de comprender a alguien? ¿Cómo podemos saber que una interpretación permite una mejor comprensión que otra?

Estas interrogantes tienen importantes consecuencias prácticas, sobre todo ahora, cuando muchas personas y comunidades diferentes tenemos que compartir un pequeño y extenuado planeta. Por ello, las posibles respuestas que demos a estas cuestiones tienen consecuencias para la ética, las ciencias sociales, la convivencia entre grupos y culturas diferentes, y para la vida en comunidad.

Uno de los objetivos de este libro es realizar una suerte de radiografía de lo que ocurre en los casos en que nos comprendemos y malentendemos mutuamente —ya sea entre individuos o entre comunidades— y tanto en circunstancias familiares y cotidianas como en aquellos casos especiales en que se produce el encuentro entre sociedades o culturas muy alejadas entre sí, o entre individuos que no comparten ninguna lengua o tradición.

Algunos casos especiales de comprensión son también dignos de análisis, como los que acontecen en el consultorio del psicoterapeuta en que dos personas que inicialmente se conocen muy poco tratan de relacionarse entre sí, o por lo menos intentan saber qué tipo de persona tienen en frente. Hay otros casos en que una persona desea comprender a otra solo para manipularla y utilizarla. También ocurre que una comunidad se propone comprender a otra para imponerse sobre ella, ejerciendo un poder ya existente o esforzándose por obtenerlo. Pero incluso en esas situaciones es necesario preguntarse qué es lo que esa comunidad cree haber comprendido y si la imposición y el sometimiento pueden incluir algún elemento de comprensión o no. Así pues, intentamos comprender al otro incluso cuando no deseamos comprendernos mutuamente. Tratamos de comprender incluso si nuestro objetivo ulterior es otro, o cuando nuestra intención es comprender sin ser comprendidos.

Las habilidades que empleamos para explicar la naturaleza y para comprender a las personas contienen elementos culturales, pero también se enraízan en estrategias que compartimos con la mayor parte de mamíferos sociales y que son el producto de millones de años de evolución de nuestro cerebro, en procesos que son tanto cognitivos como afectivos.

Mi propósito es, pues, analizar lo que de hecho hacemos cuando nos comprendemos, el tipo de estrategias que empleamos y las habilidades que ejercitamos, incluso si no lo sabemos. Interpretarnos e intentar comprendernos mutuamente es algo que hacemos todos los días con muchas personas, tanto conocidas como desconocidas. Es una práctica tan habitual que damos por descontado que podemos hacerlo y que tenemos las herramientas necesarias para ello. Solo nos preocupamos cuando alguien parece no poder hacerlo, ya sea porque no tiene las habilidades sociales requeridas o porque tiene alguna condición especial, como el síndrome de Asperger, o se encuentra dentro del espectro autista[1].

Es claro que estos temas no pueden ser abordados solo de manera conceptual y *a priori*, pues su estudio también requiere de información procedente de la psicología, las ciencias cognitivas, las neurociencias y las ciencias de la evolución, entre otras disciplinas empíricas. Esto implica que con frecuencia nos internaremos en territorios interdisciplinarios y transdisciplinarios. Este libro es, por tanto, un intento por integrar información y reflexiones de distintas disciplinas en una visión filosófica más completa acerca de la comprensión.

Nuestra comprensión de las demás personas es difícilmente separable de nuestra propia comprensión. El autoconocimiento, el conocimiento de la vida psíquica de las

[1] Es materia de debate si el síndrome de Asperger pertenece al espectro autista o es un síndrome diferente de aquel.

otras personas y el conocimiento de la realidad objetiva que compartimos con ellos conforman un inseparable triángulo en el que cada uno de los vértices presupone a los otros dos.

De hecho, en cierto sentido la filosofía griega se inauguró bajo el mandato «conócete a ti mismo» —*gnóthi seautón*— que se encontraba en el pronaos del templo dedicado a Apolo en Delfos, pues los filósofos griegos solían asumir que el conocimiento de cualquier cosa es inseparable del autoconocimiento. En efecto, con frecuencia nuestro poco autoconocimiento hace que nos resulte difícil conocer a los demás o el que nuestra historia individual haya estado aquejada por la dificultad de relacionarnos saludablemente con otras personas puede afectar nuestro autoconocimiento. Comprender a otra persona no es exactamente lo mismo que conocerla y autocomprenderse tampoco es idéntico a autoconocerse, aunque son conceptos entrelazados. La comprensión tiene una connotación de proceso, provisionalidad, subjetividad y afectividad, mientras que el conocimiento alude a la capacidad de desarrollar creencias verdaderas acerca de algo. Sin embargo, bastará con decir que si creemos comprender algo estamos en buen camino para conocerlo y viceversa.

Es tema de investigación si las estrategias que empleamos para comprender a los otros son las mismas que usamos para intentar autocomprendernos. Muchas de las estrategias son semejantes, pues, así como interpretamos a los demás atribuyéndoles estados mentales lo hacemos con nosotros mismos, y en ambos casos cometemos errores. Pero también hay importantes diferencias. En este libro, no obstante, me concentraré en la comprensión de los demás y abordaré brevemente la autocomprensión solo hacia el final del primer capítulo.

Al escribir este libro he realizado un esfuerzo por decir y justificar las ideas con claridad. Cualquier brizna de falsa complejidad, oscuridad o imprecisión es una mácula, un defecto, una carencia indeseada y reconocible solo como una incapacidad. Como señala Elytis en la frase que uso como epígrafe de este prólogo, la claridad, la precisión y la simplicidad son ya suficientemente misteriosas. El enigma está ahí. Mientras más cristalino y diáfano es algo es también más ignoto, pero incorpora la promesa de una mayor profundidad.

Los temas centrales de este libro son fenómenos interconectados y complejos que requieren ser tratados desde diversos ángulos. Por ello, los distintos capítulos abordarán diferentes aspectos de estos fenómenos, con lo cual constituyen una especie de mosaico en que solo se llega a tener una visión de conjunto cuando varias de las piezas ya están en su sitio. He intentado que el libro posea varios niveles de lectura, de manera que sea informativo e interesante para un especialista en filosofía, pero que también sea claro y provechoso para un especialista en otra área, siempre que tenga la necesaria curiosidad como para internarse en estos temas.

Todos los capítulos están atravesados por la influencia del pragmatismo estadounidense, la obra del último Wittgenstein y el pensamiento de Davidson. La primera presencia es implícita, mientras que las otras dos son explícitas. El objetivo, sin embargo, no es reconstruir las posiciones de estos filósofos ni explicarlas más claramente, sino integrarlas a ideas procedentes de otros autores y tradiciones, y tratar de hacer nuevas contribuciones. Pero ninguno de esos dos puntos es un objetivo en sí mismo, son solo medios para lo que sí es un fin: plantear preguntas relevantes para nuestras vidas e intentar aproximarnos a posibles respuestas.

Primera parte
Encontrar algo inteligible

Capítulo uno
Entender, explicar, comprender, interpretar

1.1. Breve historia de los conceptos

Las palabras tienen una larga historia de uso social que reposa sobre sus espaldas. Cuando utilizamos un vocablo estamos evocando a millones de hablantes que, antes de nosotros, se relacionaron entre sí y con su entorno para comunicarse o malentenderse. Una longeva historia de usos lingüísticos nos permite describir de múltiples formas el mundo que compartimos, así como también hace posible que definamos y configuremos nuestra propia vida subjetiva, porque en gran medida constituimos nuestra identidad a partir de las maneras en que nos describimos a nosotros mismos.

Por eso, aunque recurrir a la etimología no nos dice cómo deberíamos usar las expresiones, nos da una pista sobre los procesos sociales que rigen su uso habitual. En tanto el significado de una palabra es inseparable de las creencias compartidas que tienen sus hablantes acerca de los objetos que pretenden describir con ellas —lo que se expresa en las regularidades sociales que gobiernan su uso— puede decirse que un concepto es un resumen miniaturizado de complejas formas de comportamiento social ancestral y que analizar una palabra es examinar ese comportamiento social. Además, como nuestro conocimiento de la realidad tampoco es separable de las prácticas sociales con que la describimos, examinar el significado de una palabra es explorar la realidad en sí misma. Por todas estas razones, será conveniente comenzar este libro reconstruyendo la etimología de algunos conceptos sobre los que vamos a hablar.

La palabra castellana «entender» proviene del latín *intendere. Tendere* es la acción de tender o estirar algo, y el prefijo *in-* que la precede alude a internarse o involucrarse en eso. *Intendere* también se usa como intentar algo o concentrarse en algo. *Intendere*, por tanto, sugiere dirigirse hacia algo para intentar «entrar» en su estructura interna. Esto alude al deseo de incorporarlo a nuestra vida mental, pues implica el estar atento a un objeto. Así, por ejemplo, el castellano «desentenderse» significa

desviar la atención de algo y retirar nuestro compromiso de él. El latín *intendere* ha derivado en el inglés *to intend*, que es intentar hacer algo. Pero tanto «entender» como *to intend* son verbos que implican una acción intencional, es decir, que sugieren el estar dirigidos hacia una realidad diferente de nosotros mismos.

Entender un fenómeno es ser capaz de elaborar una suerte de estructura mental, que puede ser sistemática o narrativa, para encontrarle sentido o hacerlo «inteligible». Esta palabra, a su vez, viene del latín *intelligibilis*, que se descompone en *inter* —entre— y *legere* —leer o escoger—. Inteligir —de *intelligere*— es, por tanto, leer entre líneas o formarse una idea clara de algo a partir de una exploración de su estructura más profunda. Entender algo es encontrar un patrón que lo subyace y el cerebro ha evolucionado para buscar y encontrar patrones.

La palabra «comprensión» por otra parte, viene de *comprehendere*, que alude a capturar, coger, agarrar o atrapar algo, por ejemplo, una idea. Pero el prefijo *com-* implica amplitud o globalidad y *pre* anterioridad temporal. De manera que comprender o comprehender alude a lograr asir algo más bien escurridizo. Se suele traducir el inglés *to understand* por comprender, aunque creo que una mejor opción sería entender y dejar comprender para traducir el inglés *to comprehend*, que además tiene la misma etimología. *Understand* viene del inglés antiguo *under* —debajo— y *stand* —yacer o estar ubicado—, de manera que se relaciona más con entender que con comprender.

Explicar viene del latín *explicatio*, que significa desplegar, exponer, expandir o desarrollar, por ejemplo, en el sentido de extender, desenrollar y desenvolver un mapa o un pergamino. Es interesante que en sus orígenes el verbo también se usara para describir el proceso de estirar las piernas o los brazos, lo que sugiere la idea de ocupar un espacio vacío con el propio cuerpo.

Como señalé en el prólogo, usaré «entender» como un concepto más general que incluye a otros dos más específicos, «explicar» y «comprender», y reservaré «explicar» para el proceso de entender o hacer inteligible la naturaleza, con el fin de desplegar su estructura para reconocer sus elementos más básicos y constituyentes, y «comprender» para el proceso de entender o hacer inteligible a un ser humano o a una comunidad humana, tratando de capturar algo de su subjetividad o vida mental, mientras ampliamos nuestra propia subjetividad y vida mental. Explicar y comprender serían, por tanto, formas de entendimiento o intelección.

Interpretar, por otra parte, viene del latín *interpretatio*, que tiene el sentido de traducir, elucidar o aclarar el sentido de un texto ambiguo o confuso. Vemos, por tanto, que estas diversas palabras se superponen y con frecuencia comparten sentidos específicos. Seguiré la tradición filosófica en usar «interpretar» como el conjunto de estrategias que empleamos para comprender a alguien, de manera que la comprensión sería el objetivo al que apuntamos y la interpretación el medio o instrumento para ello.

Ahora bien, aunque nadie es dueño de los significados de las palabras y estos van evolucionando según criterios muy variados e impredecibles, los filósofos suelen proponer usos técnicos específicos con la esperanza de que resulten más esclarecedores que los coloquiales. De esa manera va avanzando la disciplina. En este capítulo, trataré de iluminar los significados filosóficos y las relaciones entre los conceptos que nos interesan.

El concepto de explicación se fue afinando —desde sentidos previos— a partir del progreso de las ciencias naturales, cuyo desarrollo se aceleró hacia mediados del siglo XVI con autores como Copérnico, Galileo y Bacon. A diferencia de Aristóteles, quien tenía una concepción cualitativa de la explicación científica, ellos proponían una explicación cuantitativa, es decir mensurable empíricamente, pues esto garantizaría mayor precisión y objetividad. Así es como se instaló la idea —ya sea implícita o explícita— de que toda explicación es, en el fondo, predicción y que, en última instancia, deseamos predecir el curso de la naturaleza para poder adaptarnos mejor a esta o para adaptarla a ella a nuestras necesidades. Por eso, de las cuatro causas (*aitíai*) que para Aristóteles eran los cuatro principios explicativos en cualquier ámbito, solo quedó la causa eficiente como criterio de explicación y las otras tres tendieron a desaparecer de la vista. Sin embargo, aunque el concepto de explicación moderno representa una ruptura respecto de Aristóteles, es claro que tiene con él una deuda esencial.

Para Aristóteles explicar un fenómeno o un evento es encontrar las causas que responderían a las preguntas sobre por qué es como es y no de otra forma. Como es conocido, estos cuatro principios explicativos inquieren por: (1) otros eventos que produjeron el fenómeno —causa eficiente—, (2) la base física del fenómeno —causa material—, (3) el concepto instanciado en el fenómeno —causa formal— y (4) la finalidad u objetivo del fenómeno —causa final—. A partir de estos principios explicativos, Aristóteles construye un silogismo para subsumir de manera deductiva, nomológica y causal el fenómeno por ser explicado.

Consideremos como un ejemplo el fenómeno de la caída de los cuerpos. Para Aristóteles los objetos tienden a su lugar natural: aquellos compuestos de tierra y agua se dirigen hacia el centro de la Tierra, y los que tienen mayor parte de aire y fuego hacia la última de las esferas en el mundo supralunar. Explicar la caída de una roca tomaría, entonces, la siguiente forma:

- Premisa mayor: Los objetos principalmente compuestos de tierra tienden naturalmente al centro de la Tierra.
- Premisa menor: Las rocas están principalmente compuestas de tierra.
- Conclusión: Las rocas tienden naturalmente al centro de la Tierra.

Esta explicación tiene tres rasgos importantes: es deductiva, es nomológica y es causal. Es deductiva en tanto la conclusión se infiere lógicamente de las dos premisas. Es nomológica porque el objetivo es subsumir lo particular dentro de lo general, es decir, el evento por ser explicado —la caída de una roca en particular— al interior de una regularidad —la caída de las rocas en general—, lo que a su vez estaría gobernado por una regularidad aún mayor: todos los objetos compuestos mayormente de tierra y agua se dirigen naturalmente al centro de la Tierra, que es el centro del universo. Y la explicación es causal porque hace uso de las cuatro causas ya señaladas, con el objetivo de contestar preguntas que tienen la forma de por qué algo es como es y no de otra forma (Aristóteles, *Física*, II,7).

En esto Aristóteles continúa una tradición griega cuyo representante paradigmático es Platón, según la cual lo que se conoce es lo universal y necesario y lo particular solo puede ser conocido en tanto es una instancia de una Forma universal. Es más, incluso hacia comienzos de la Modernidad, Francis Bacon denominaba a las leyes naturales «las Formas de la naturaleza» (1961, libro II, párrafo 17). Esta intuición sigue presente en nuestra comprensión de la ciencia natural actual, pues solemos considerar que conocer un objeto o evento particular es, en el fondo, saber a qué categoría mayor pertenece. A lo largo de este libro discutiremos hasta qué punto esta idea también está presente y de qué manera, en la explicación y en la comprensión del comportamiento particular de los seres humanos y de sus procesos psíquicos.

Hacia mediados del siglo XIX el filósofo francés Auguste Comte, padre de la sociología clásica, sostuvo que una disciplina científica que no dejara espacio para la metafísica, la religión o la superstición sería aquella que tuviera la capacidad de encontrar las conexiones entre los acontecimientos individuales y las regularidades generales (2002). Aunque claramente esta idea alude a Aristóteles, es el origen de lo que posteriormente se llamaría el modelo de cobertura legal o nomológico deductivo, que apareció en las primeras décadas del siglo XX en el contexto del positivismo lógico del Círculo de Viena. Este modelo sostenía que explicar un evento es ser capaz de encontrar sus conexiones causales y subsumirlas dentro de leyes universales. A mediados del siglo XX, Carl Hempel y Paul Oppenheim (1948; Hempel, 1970, 1973) propusieron de manera formal una versión de este modelo que, aunque fue cuestionado desde sus orígenes, sigue siendo el referente clásico de explicación científica natural.

La objeción principal al modelo nomológico deductivo o de cobertura legal es que este se concentra en la búsqueda de regularidades y deja de lado la razón por la que tales regularidades se producen, es decir, no inquiere sobre la estructura interna de los acontecimientos que genera la aparición de otros acontecimientos (Salmon, 1971, p. 34). En otras palabras, no bastaría con mostrar que, por ejemplo, el calor dilata

los metales y que esa relación causal está gobernada por una regularidad. Sería necesario mostrar qué propiedad del calor hace que ciertas propiedades de los metales produzcan su dilatación.

En una orientación semejante a la de Wesley Salmon, Stephen Grimm (2006, 2016; Grimm, Baumberger & Ammon, 2017) sostiene que el objetivo último de la ciencia —que es un bien en un sentido valorativo— es comprender la realidad, lo que es algo más amplio que simplemente explicarla. Esta es una idea que comparte con Wesley Salmon (1998), Peter Lipton (2004) y Michael Strevens (2006), entre otros. Según esta posición, la finalidad de la ciencia no es solo producir oraciones verdaderas acerca del mundo sino generar un tipo de comprensión sistemática. El objetivo sería desentrañar la estructura de la realidad para mostrar la manera en que sus diferentes partes se relacionan entre sí a la forma de un círculo hermenéutico, es decir, mostrando el lugar que estas tienen en el todo y cómo el todo está conformado por las partes. Una idea que resalta Grimm, no obstante, es que no es la ciencia la que comprende, sino los individuos que la practican.

Peter Manicas también afirma que las ciencias naturales aspiran a la comprensión (*understanding*) y que es un error asociar la explicación a la predicción (1987, 2006). Su argumento principal es que cuando explicamos la naturaleza no nos interesa solamente predecirla y dominarla, sino entender cómo funciona. Esto es, no solo nos interesa saber cuándo una causa produce un efecto sino también por qué lo produce, es decir, cuál es la estructura interna de un evento que origina a otro evento diferente. Piensa Manicas (2006, pp. 16-25) que el concepto de causalidad no alude solamente a una relación de sucesión de eventos, con lo que se aleja diametralmente de David Hume, sino a una relación entre las estructuras que constituyen a ambos eventos y que permite que uno de ellos tenga la capacidad de producir al otro, que es lo que Hume rechazó al suponer que esta es una noción ininteligiblemente metafísica. La idea de Manicas es que en tiempos de Hume esta podría haber sido una noción ininteligible, pero que al día de hoy no lo es, lo que lo conduce a replantear el concepto mismo de causalidad (2006, pp. 26-41). Según Manicas, ahora podemos saber qué características tiene un evento que le permiten causar otro, de manera que una adecuada explicación causal tendría que incluir esa información más allá de la pura constatación de la sucesión regular. Esto le conduce a decir que las ciencias naturales no solo deben aspirar a explicar y predecir sino también deben intentar comprender.

Aunque estos argumentos son razonables, un defensor del modelo nomológico deductivo podría responder sosteniendo que este sí logra examinar la estructura interna de los eventos involucrados en una relación causal. Volvamos al ejemplo de «el calor dilata los metales». El modelo nomológico deductivo no solo registrará

esa relación causal y la subsumirá en una regularidad natural, sino además podrá construir otras regularidades causales que expliciten las propiedades físico-químicas del calor que generan cambios en las propiedades del metal, y originan su dilatación. De esta manera, el avance de la ciencia marchará en dos direcciones opuestas pero complementarias. Por una parte, encontraremos relaciones causales nomológicas cada vez más constitutivas y elementales, es decir, ya no solo en las relaciones entre el calor y el metal, sino en las moléculas o átomos que los conforman, y así hasta donde nos resulte posible. Por otra parte, también buscaremos relaciones causales más comprehensivas y que expliquen más entidades y eventos del universo macroscópico hasta donde sea posible.

De otra parte, Grimm y Manicas usan la palabra comprensión en un sentido mucho más débil que yo. Para ellos se comprende un fenómeno cuando se conoce su estructura interna en relación al todo y a este como conformado por la integración de sus partes. Yo iría más lejos, para mí la comprensión incluye lo señalado por esos autores, pero también la capacidad de imaginar cómo sería ser ese objeto para sí mismo, es decir, desde su experiencia fenoménica de sí y del mundo al que pertenece, mientras uno teje una red interpretativa que conecta sus propios estados mentales, significados y valoraciones con los ajenos, lo que permite que comparta un espacio subjetivo con el otro. Es claro que este sentido de comprensión solo puede aplicarse a criaturas a quienes uno atribuye subjetividad y en quienes se reconoce la capacidad de percibir sentidos y hacer valoraciones. Podría pensarse que la noción de comprensión que empleo es demasiado fuerte, pero tendrá que admitirse que eso es lo que hacemos en la vida cotidiana y que esos procedimientos también son empleados por las ciencias humanas —algunas ramas de la psicología, de las ciencias sociales y de la historia son casos paradigmáticos— y que, en consecuencia, hay que analizarlos teóricamente y diferenciarlos de las actividades y objetivos de las disciplinas que no hacen eso.

En todo caso, hay un sentido importante en el que Grimm y Manicas tienen razón, pues el objetivo de la ciencia no es solo predecir la naturaleza sino hacerla inteligible. Precisamente por eso podría decirse que explicación y comprensión son dos modalidades de un fenómeno más amplio, que es entender algo. Coincido con Grimm en que el objetivo de la ciencia incluye elaborar una visión que muestre la estructura interna que tienen los diferentes elementos que deben estar relacionados para que el universo tenga la forma que tiene y fluya como lo hace. Pero el problema es que Grimm usa la expresión *to understand* tanto para esa actividad como para comprender a los seres humanos. Eso es inconveniente porque necesitamos emplear una palabra diferente para la actividad de compartir la subjetividad ajena, que claramente no podemos hacer con el universo físico, aunque sí con los seres humanos.

Pienso, por tanto, que Grimm yerra al usar el verbo *to understand* tanto para la actividad de dar sentido al universo como para compartir el punto de vista de un agente. Por eso prefiero usar «entender» o «inteligir» para referir a la actividad más general de hacer que algo tenga sentido para uno. Esta se convierte en «explicación» cuando elaboramos descripciones de la naturaleza y «comprensión» cuando lo hacemos de agentes intencionales. En el primer caso —la explicación—, operamos buscando relaciones causales gobernadas por regularidades y esto lo hacemos tanto con fenómenos físicos como con agentes intencionales, pues estos últimos también pueden ser explicables. En el segundo caso —la comprensión— operamos construyendo descripciones que nos permiten capturar algo de la subjetividad ajena. En ambos casos podemos eventualmente predecir, con mayor o menor precisión, el comportamiento del objeto que deseamos entender. En ambos casos la actividad empleada también nos permite hacer inteligible, entender o encontrar sentido al objeto de nuestro interés, y esto puede interpretarse como «convertir lo extraño en familiar», para recurrir a la célebre frase hermenéutica, con el fin de postular un orden más simple que dé cuenta de uno que es más complejo. Hacemos esto para relajar la ansiedad que nos produce lo ignoto y desconocido, en la misma línea en que para Charles Sanders Peirce se fijan las creencias para calmar la ansiedad que produce la duda, como sostiene en «The Fixation of Belief» (*CP*5.358-387, W3: 242-257, *EP* 1.109-123). Esto está relacionado con la tendencia de los mamíferos a familiarizarse con el entorno —representándose una estructura para maximizar sus posibilidades de supervivencia— y a su tendencia natural a desarrollar angustia cuando eso no es posible.

La explicación no puede ser un subconjunto de la comprensión, porque si lo fuera sus objetos tendrían que tener subjetividad. La comprensión tampoco puede ser un subconjunto de la explicación, porque de ser así por lo menos alguna región de la explicación poseería subjetividad. Por ello es mejor ver ambos conceptos como conjuntos intersecados que comparten algunos elementos, pero que también tienen rasgos diferentes, siendo, ambos, parte de una actividad mucho más amplia de intelección.

Lo que sí podría decirse es que la extensión de objetos que pueden ser explicables incluye a la extensión de objetos que pueden ser comprendidos. Esto significa que, en principio, todos los eventos físicos pueden ser objeto de explicación, pero solo algunos de ellos también pueden ser objeto de comprensión, de manera que, en principio, todo lo comprensible es explicable pero no todo lo explicable es comprensible.

Como hemos visto, Manicas (1987, 2006) piensa que el problema central con el modelo de cobertura legal es que está comprometido con el concepto humeano de causalidad, que se limita a registrar la sucesión regular de acontecimientos sin pretender —por considerarlo imposible— examinar la estructura interna del acontecimiento que generó el efecto. Por ello se propone revisar el concepto de explicación

causal para que pueda capturar los mecanismos internos que producen los acontecimientos, más allá de solo ser capaz de predecirlos. Lo curioso es que con ese concepto minimalista de causalidad David Hume pretendió explicar el comportamiento humano.

En efecto, en 1751 Hume publicó su *Investigación sobre los principios de la moral* (2006), en el que acuñó la expresión *moral sciences* para referir a las disciplinas que tienen como objetivo estudiar la conducta de los seres humanos y sus construcciones culturales. Hume confesó que quiso ser el Newton de las ciencias morales, en tanto le interesaba construir una teoría que pudiera explicar los fenómenos humanos de manera tan confiable —y eventualmente predictiva— como Newton pudo hacerlo con el comportamiento de los objetos físicos. Así, el filósofo escocés quiso elaborar una filosofía moral que fuese más empírica y descriptiva que normativa.

A partir del interés de Hume surgieron preguntas sobre si las *moral sciences* —que hoy solemos traducir por ciencias humanas o ciencias sociales[1]— tienen el mismo método que las ciencias naturales y cuál es el rol que la causalidad tiene en ellas. De hecho, en su pequeño texto titulado «Nature», John Stuart Mill (2017) sostenía que es posible explicar los fenómenos sociales y psicológicos a partir de regularidades causales básicas, las cuales se encontrarían gobernadas por regularidades universales. Según Mill nuestra imposibilidad de hacer predicciones estrictas en relación al comportamiento humano provendría de la diversidad y complejidad de las causas, no de que el objeto de estudio sea ontológicamente diferente. Así pues, si Mill hubiera podido conocer el modelo de cobertura legal probablemente habría creído que es aplicable al comportamiento humano y que permite hacer predicciones, inexactas pero aproximadas, a partir del reconocimiento de regularidades más amplias.

Es conocido que Hume sostenía que la causalidad no es parte de la realidad, lo que influyó notablemente en Kant, quien sostenía que la causalidad es una categoría del entendimiento. Pero hay que entender adecuadamente la tesis de Hume. Lo que él quiere decir es que la realidad se compone de eventos naturales observables y que la causalidad no es un evento adicional a ellos. Así, decimos que hay una relación causal entre los eventos A y B si cada vez que ocurre A observamos que también ocurre B y, por tanto, asumimos que esa secuencia está gobernada por una regularidad de la naturaleza que puede ser descrita mediante una oración cuantificada universalmente del tipo siguiente:

$$(x)\ (Ax \rightarrow Bx)$$

[1] Para todos los efectos prácticos asumiré que «ciencias humanas» y «ciencias sociales» refieren al mismo conjunto difuso de disciplinas que estudian el comportamiento humano individual y social, de manera que emplearé ambas expresiones intercambiablemente.

Para todo objeto x, si este tiene la propiedad A, entonces tiene la propiedad B. Esa afirmación permite oraciones contrafácticas del tipo «Si hubiese ocurrido A, entonces también habría ocurrido B», que es lo que se suele llamar «el principio del carácter nomológico de la causación». Hoy suele asumirse que tanto en las ciencias naturales como humanas este principio no es absoluto, sino más bien probabilístico, de manera que toda relación causal está gobernada por una regularidad que puede tener distintos grados de probabilidad. Esto es tanto una consecuencia de observar que en el mundo natural podría haber regularidades flexibles, como del reconocimiento que la afirmación de que hay leyes inexorables es un presupuesto metafísico que no puede ser justificado de ninguna forma.

Ahora bien, la casi identificación que se hacía en el ámbito natural entre explicación y predicción suscitó en muchos filósofos la pregunta sobre si, en el caso de los estudios humanos, el único objetivo posible es la predicción o si es factible un conocimiento menos cuantitativo y más cualitativo de su objeto de estudio, tomando en consideración que en este caso, a diferencia de lo que ocurre con las ciencias naturales, el objeto de estudio y el sujeto que pretende conocerlo es el mismo y se trata, en consecuencia, al mismo tiempo de un fenómeno de auto y aleoconocimiento.

Hacia comienzos del siglo XX Johann Droysen (1983 [1937]) acuñó la distinción entre explicación (*erklären*) y comprensión (*verstehen*), con el objetivo de diferenciar entre el método de las ciencias físicas y lo que él denominaba las ciencias históricas, pero Wilhelm Dilthey (1989) la amplió y popularizó para distinguir entre el método de las ciencias naturales (*Naturwissenschaften*) y el de las ciencias humanas (*Geisteswissenschaften*). De esta manera, se robusteció la intuición de que el objetivo de las ciencias humanas no puede ser solamente predecir, pues ellas también deben aspirar a comprender, lo que inevitablemente involucra elementos subjetivos. Así también apareció en la tradición hermenéutica la idea de que, si la comprensión es el objetivo, la interpretación es el instrumento.

En otra línea de desarrollo, en la segunda mitad del siglo XX, la filosofía de la mente angloamericana desarrolló la idea de que la interpretación es el proceso de atribuir estados mentales, acciones y significados a los agentes de eventos intencionales. Si bien poca gente duda hoy de que a las ciencias humanas les interesa tanto explicar cómo comprender, se mantiene el debate acerca de las diferencias y semejanzas entre ambos objetivos y sus respectivas metodologías. Más aún, después de varias décadas de predominio de una concepción de la ciencia excesivamente influida por el positivismo lógico, en los últimos años ha renacido la discusión sobre las relaciones entre ciencias humanas y ciencias naturales que no considera a las ciencias naturales como el paradigma de la cientificidad y del tipo de conocimiento al que uno debiera aspirar. Esta polémica, a su vez, obliga a reflexionar sobre el tipo de enfoque

y metodología que brinda la filosofía en medio de la diversidad científica, lo que conduce a interesantes investigaciones de tipo metafilosófico que, sin embargo, están más allá de los objetivos de este libro (véase Williamson, 2008; Quintanilla, 2008c; Rescher, 2014; Lefebvre, 2016; Monteagudo & Quintanilla, 2018). Por eso ahora volveremos nuestra mirada a las relaciones entre explicación y comprensión.

1.2. Explicar y comprender

La explicación de un evento físico, pero también del comportamiento intencional de alguien, requiere de un punto de vista externo que tenga ciertas pretensiones de objetividad. Esa descripción apunta como un ideal regulativo —es decir nunca realizable plenamente, pero al que siempre nos podemos acercar un poco más— el ser válida para cualquier observador capacitado. Así, por ejemplo, las explicaciones de la estructura de un agujero negro, de las funciones de la corteza prefrontal del *Homo sapiens*, de la hiperinflación alemana anterior a la Segunda Guerra Mundial, del enfrentamiento entre Huáscar y Atahualpa, de la etiología de la neurosis obsesivo-compulsiva de un paciente en particular o de las razones por las que Pachacútec conquistó a los chancas son, en principio, independientes de las características personales o culturales del investigador y aspiran a ser lo más objetivas posible, siempre que se den las condiciones cognitivas y sociales que hagan posible la investigación científica. Por lo menos eso se presupone y a eso se apunta. En otras palabras, cuando un historiador quiere explicar por qué Pachacútec conquistó a los chancas no pretende solamente dar su punto de vista entre varios otros, sino que desea mostrar una articulación de causas, razones y evidencias que deberían convencer a quien quisiera escucharlo. Este historiador supone, por tanto, que quien tuviera la evidencia que él tiene y la estructurara de la manera que él considera correcta, tendría que arribar a las mismas consecuencias que él llegó. En ese sentido, su explicación tiene pretensiones de objetividad.

A diferencia de lo anterior, comprender a alguien requiere tener la habilidad para compartir algún aspecto de su perspectiva, aunque sin perder la propia; precisa de participar de su punto de vista subjetivo o de imaginar cómo sería ser él en determinadas circunstancias de su historia personal. También necesita tener la capacidad de elaborar una narrativa comparable con la que el individuo haría de sí mismo, aunque siempre desde la perspectiva de la propia intérprete[2]. Esto supone ver las cosas, por un momento, como estas debieron habérsele presentado al agente, y percibir la valoración

[2] Por razones de claridad y economía verbal, usaré el género femenino para la intérprete y el masculino para el agente. Así, cada vez que diga «ella» se asumirá que me refiero a la intérprete u oyente, y cuando diga «él» deberá entenderse que hablo del agente, hablante o interpretado.

y el significado que él atribuye a ciertos acontecimientos y acciones, lo que permite que la intérprete imagine que ella misma podría haber sido esa persona. Una consecuencia central de esto es que la comprensión está siempre situada en una perspectiva y depende de las características individuales y culturales de quien interpreta. No hay tal cosa como una comprensión no comprometida y toda comprensión implica la construcción de algún tipo de espacio o vínculo entre quien comprende y quien es comprendido[3]. En otras palabras, si la explicación tiene pretensiones de objetividad, la comprensión tiene un componente inevitablemente subjetivo y afectivo.

Se puede comprender a alguien sin estar en condiciones de explicar su comportamiento, así como se puede explicar el comportamiento de una persona sin poder comprenderla. Un ejemplo del primer caso es la psiquiatra que puede explicar lo que ocurre con un psicótico, aunque difícilmente diríamos que ella logra participar de su punto de vista. Otro ejemplo es la historiadora que desea explicar por qué Pachacútec conquistó a los chancas en términos de relaciones económicas y de poder, sabiendo poco o nada sobre la vida mental de Pachacútec. Un ejemplo de lo segundo es la amiga que puede compartir las experiencias de su amigo, aunque no se explica cómo él pudo haber actuado como lo hizo, ni por qué tiene las extrañas creencias y deseos que tiene. La comprensión alude al concepto de empatía, pero esta es condición necesaria, aunque no suficiente. Sobre esto volveré en el capítulo ocho.

Explicar es lo que hacen las ciencias cuando ubican relaciones causales entre eventos, buscando la regularidad que las gobierna y describiéndola con mayor o menor grado de probabilidad. Hacer esto permite predecir el comportamiento del evento que ha sido explicado. Las disciplinas que explican de esta manera son paradigmáticamente las ciencias naturales, pero a las ciencias humanas también les interesa entender causalmente y sobre la base de regularidades —que pueden ser más o menos estrictas— por qué los individuos o las sociedades se comportan como lo hacen. Por eso sería una simplificación afirmar categóricamente que mientras las ciencias naturales explican, las ciencias humanas comprenden. También es simplificador suponer que todas las ciencias naturales emplean un mismo método explicativo y que todas las ciencias humanas comparten un método de interpretación. Hay diferencias importantes entre la manera de explicar de la física, por ejemplo, y la de la biología; también las hay entre las estrategias explicativas de la antropología, la historia y la economía. Más aún, hay disciplinas que están en la intersección entre las ciencias naturales y las humanidades, como la lingüística o la psicología, que se proponen explicar y también comprender.

[3] Esta es una intuición que está presente en autores de distintas tradiciones como, por ejemplo, Wittgenstein, Gadamer y Davidson. En este libro trataré de explicitar y enriquecer esa idea.

También hay ciencias que necesitan hacer uso de la explicación para que la comprensión sea posible y de la comprensión para que pueda darse la explicación. Un ejemplo de lo primero es la psicóloga que necesita conocer ciertas regularidades del comportamiento del paciente para poder capturar algo de su subjetividad. Un ejemplo de lo segundo es la socióloga que desea conocer algunas experiencias subjetivas colectivas para poder explicar determinadas regularidades sociales.

Otros ejemplos interesantes son la antropología o la historia, para no hablar de disciplinas más recientes como las neurociencias, el neuropsicoanálisis, la arqueología cognitiva o la psicología evolucionista. La mayor parte de estas disciplinas no existían o no tenían las características que tienen ahora, cuando en el siglo XIX la hermenéutica alemana hizo la distinción entre ciencias naturales y ciencias humanas. Por eso es conveniente suponer que no hay una frontera, sino una intersección rica en diversidad y contenido.

En lo que respecta a la filosofía, aunque esta actividad suele ser ubicada por razones administrativas en los departamentos de humanidades, muchos de los filósofos y de sus producciones, tanto del presente como del pasado, podrían con justicia pertenecer a otros departamentos, sean estos de física, biología, psicología, historia, literatura o ciencia política, para mencionar solo algunos ejemplos, porque estas disciplinas emergieron de la filosofía como si de una Gran Explosión se tratara. Pero, más allá de la atención que los filósofos deben dar a la evidencia empírica y al desarrollo de las diversas ciencias, su objetivo principal es analizar los conceptos con que nos acercamos a la realidad para entenderla.

Mi tesis, entonces, es que las ciencias naturales y las ciencias humanas son dos círculos que se intersecan, pues comparten muchos rasgos, pero también tienen diferencias. Las ciencias naturales solo explican; las ciencias humanas explican y comprenden. Asimismo, explicar y comprender son dos puntos de vista o perspectivas de distintos aspectos de los mismos objetos. Esta idea es consistente con la tesis que defenderé, según la cual, en el caso del ser humano, la descripción mental —o psicológica— y la descripción corporal —o física— iluminan dos aspectos diferentes de una misma realidad.

Visto de una manera general, el universo es un conjunto de eventos físicos en el que un subconjunto sumamente pequeño puede también describirse como acciones intencionales. Así, aunque toda acción intencional es también un evento físico, no todo evento físico es una acción intencional. Por ello, aunque todo evento que puede ser comprendido —una acción— podría en principio ser explicado en términos físicos, no todo evento físico que puede ser explicado también podría ser comprendido como una acción intencional. Solo puede ser comprendido un agente dotado de intencionalidad y, a pesar de que todo agente intencional es un objeto físico, es claro

que no todo objeto físico es también un objeto intencional. Del conjunto de objetos físicos que conforman la naturaleza solo un muy pequeño subconjunto puede ser descrito de manera intencional como para poder ser sujeto de comprensión.

Un evento físico es también una acción cuando puede ser descrito de manera intencional. Esta es una idea recurrente en Gertrude Anscombe (2000) y en Donald Davidson (1980b), así como en muchos otros autores. Tal como lo veo, inspirado en Davidson (1980c [1971], p. 195), la intencionalidad es una propiedad relacional triádica entre un agente, la descripción que hace de él una intérprete y un conjunto de eventos objetivos del mundo que comparten. La descripción hace que el evento sea intencional seleccionando una propiedad de este. Compárese, por ejemplo, las siguientes dos oraciones:

(1) Edipo mató a su padre en el lugar L y en el tiempo T.

(2) El segundo esposo de Yocasta mató al primer esposo de Yocasta en el lugar L y en el tiempo T.

Es claro que mientras la oración (1) reconoce la intencionalidad de Edipo de matar a su propio padre, la (2) no lo hace, aunque el evento físico es exactamente el mismo. La intencionalidad, por tanto, no es una propiedad inmanente del evento físico sino una propiedad relacional que emerge en la situación interpretativa.

Solo son pasibles de comprensión las criaturas cuyo comportamiento puede ser descrito como un conjunto de acciones causadas —por lo menos parcialmente— por sus propios estados mentales, de manera que estas criaturas son reconocidas como agentes dotados de subjetividad y agencia. Tendemos a asumir que los agentes intencionales que pueden ser comprendidos son solo los *Homo sapiens*, pero la frontera es imprecisa porque podría haber especies no humanas que tengan algún grado —por pequeño que sea— de subjetividad y agencia, y ciertamente no todos los seres humanos están dotados de estas. Más aún, los que lo están no lo están en todos los momentos de su vida ni en el mismo grado o en la misma forma. No lo están, por ejemplo, los bebes recién nacidos ni las personas en estado de coma. Tampoco lo están plenamente los psicóticos, quienes se encuentran bajo el efecto de ciertas drogas, los que son víctimas de manipulación, sometimiento político o de algún tipo de ideología.

La subjetividad es una propiedad fenoménica que tiene un individuo que se experimenta a sí mismo. Esta experiencia fenoménica es privada, aunque sin duda se desarrolla siempre en una comunidad. Hay razones para sostener que una persona mantenida en absoluta soledad no podría vivir físicamente y que, de hacerlo, carecería de subjetividad, aunque podría tener conciencia nuclear, como veremos luego. Sin duda, la subjetividad se enriquece ontogenéticamente con el desarrollo neurológico y la vida en comunidad. Probablemente lo mismo ocurre en el nivel filogenético.

En el caso de la agencia, es discutible si somos agentes cuando soñamos estando dormidos, es decir, si podemos realizar acciones en sueños. También es tema de debate si las comunidades humanas pueden tener agencia, esto es, si hay agencia colectiva. En cualquier caso, la agencia tampoco es una propiedad que se tiene o no se tiene, sino una que puede estar presente en mayor o menor medida en distintos momentos de la vida de las personas o los grupos. La agencia es una propiedad relacional emergente que se constituye intersubjetivamente —de manera que es tanto descubierta como atribuida—, como sostendré que son los contenidos proposicionales de los estados mentales que constituyen la subjetividad, aunque no la experiencia fenoménica de la subjetividad misma, que es privada. Sobre esto volveremos más adelante.

Subjetividad y agencia están interrelacionadas. Por ejemplo, algunas personas o comunidades tienen más agencia que otras en tanto poseen más capacidades, como afirmaría el modelo de desarrollo como ampliación de las capacidades propuesto por Amartya Sen (2000). Pero sostengo que, en líneas generales, autoconocimiento y agencia son directamente proporcionales (Quintanilla, 2009c). La agencia es una propiedad con la que los infantes humanos no nacen, pero que desarrollan progresivamente a lo largo de muchos años hasta alcanzar cierto grado hacia la segunda década de su vida. Cuando eso ocurre las personas también adquieren otras propiedades como una rica vida subjetiva, el libre albedrío y la responsabilidad moral. Estas son propiedades emergentes y relacionales de individuos que reconocemos y atribuimos cuando los interpretamos. También es una propiedad emergente, pero en este caso no relacional sino monádica, aunque igualmente constituida en comunidad: la experiencia fenoménica de la subjetividad.

La distinción entre explicación y comprensión es algo gruesa y simplificadora, pero, en líneas generales, útil. La explicación busca subsumir lo particular en lo general. La comprensión también lo hace, puede ser causal y estar concernida en buscar regularidades, pero con un interés adicional en algo que la explicación no hace: capturar un punto de vista, una vivencia, una experiencia fenoménica o —lo que técnicamente se suele llamar en filosofía de la mente— los *qualia* de la subjetividad. Como ha sugerido Daniel Dennett (2017, p. 94), la comprensión no es una experiencia abrupta. No es, por tanto, como si uno se diera cuenta de algo en un golpe de lucidez. Uno puede comprender algo mejor o peor y puede mejorar o empeorar su comprensión. Ciertamente, hay momentos en que sentimos que tenemos una suerte de epifanía que nos permite comprender finalmente algo, pero posteriormente uno puede comprender que su supuesta experiencia de comprensión no era tal, o que faltaba aún mucho por comprender.

Como hemos visto, explicar e interpretar son distintas estrategias que tienen como finalidad última hacer algo inteligible para alguien y, en el caso de la interpretación,

la comprensión es su objetivo. Ahora podemos pasar a describir y analizar algunas de las características y presupuestos de cada una de estas estrategias.

De manera general, explicar un fenómeno exige contestar dos tipos de preguntas: ¿cómo es algo? y ¿por qué es cómo es y no de otra manera? Es decir, requiere, en primer lugar, de una descripción de sus propiedades más notables y, en segundo lugar, de un tipo de razonamiento —una articulación de razones— que muestre y haga inteligible por qué ese fenómeno tiene las propiedades que tiene y no otras que también podría haber tenido. Aunque no hay una sola forma de explicación y las distintas ciencias explican de maneras diferentes, dos rasgos compartidos son la subsunción de lo particular en lo general y la reconstrucción de su historia causal.

Como vimos, la palabra «explicación» proviene del latín *explicatio*, que significa desplegar o desenvolver. Para el filósofo alemán del siglo XV, Nicolás de Cusa (1957 [1440]), el universo es un despliegue —*explicatio*— de la divinidad que, aunque de manera imperfecta y limitada, desarrolla de muchas maneras aquellos atributos que se encuentran en Dios, aunque condensados de manera unitaria —*complicatio*—. Pero lo central en el concepto de *explicatio* es que uno desenvuelve algo para poder observar sus características implícitas, aquellas que se encontraban ocultas antes del despliegue y que permiten aquellas características explícitas que hacen que el objeto sea lo que es y no otra cosa.

Casi siempre explicar algo es reducirlo a algo más básico y, por eso, toda explicación es una simplificación. Por ejemplo, explicar el movimiento gravitatorio de los astros requiere traducir esa gran complejidad a un simple conjunto de ecuaciones. Explicar la complejidad del comportamiento de una persona o de una comunidad implica traducir esa enorme diversidad a un pequeño conjunto de objetivos, creencias o intereses. Explicar la evolución de los homínidos necesita traducir un sinnúmero de eventos que ocurrieron a lo largo de muchos millones de años a un conjunto de acontecimientos paradigmáticos. En esas traducciones hay inevitables simplificaciones, pero ese costo está justificado gracias a la ventaja de poder hacer inteligibles, para alguien, procesos que de otra manera serían totalmente incomprensibles. Por otra parte, el que la explicación incorpore una traducción de lo diverso y complejo a lo básico y simple no implica que esos elementos básicos sean un conjunto de principios fundamentales en sí mismos. Son solo un conjunto de claves que tienen sentido para una persona o una comunidad epistémica, según el propósito explicativo que estas tengan en mente.

En todo caso, en medio de ciertos parecidos de familia hay una gran variedad de formas de explicación y esta diversidad constituye la pluralidad del mundo académico actual. Así, aunque el objetivo de toda ciencia, tanto natural como humana, es explicar las características principales de su objeto de estudio y esto incorpora

algún tipo de subsunción y la búsqueda de las relaciones causales entre los eventos que constituyen ese fenómeno, las diversas ciencias explican de modos diferentes.

Se denomina *explanandum* al fenómeno por ser explicado y *explanans* a la articulación de criterios, razones, datos, encadenamientos históricos, etcétera, que nos permitirán realizar la explicación. Aunque no hay un solo modelo de explicación científica sino varios, la mayor parte de ellos ve una conexión entre explicación y relación causal. Hay diversas condiciones que se supone deben estar presentes en toda relación causal, pero la más general y menos comprometedora es aquella que sostiene que A es causa de B, si A es condición necesaria, aunque no suficiente, de B o, en otras palabras, si A es *conditio sine qua non* B o, formulado de manera más coloquial, si cada vez que ocurre A debe ocurrir B.

En un libro clásico, Ernst Nagel (1961) caracterizó cuatro modelos de explicación (véanse también Salmon, 1988, 1990 y Gonzáles, 2002): el deductivo, el probabilístico, el funcional o teleológico y el genético. Según el modelo deductivo, el fenómeno por ser explicado —*explanandum*— se sigue como una consecuencia lógica necesaria de un conjunto de premisas —*explanans*—. El modelo hipotético deductivo de Karl Popper (1982, 1983) sería una versión de ello y también lo sería la versión clásica del modelo de cobertura legal. La forma tradicional de plantear esta concepción es bajo el nombre de modelo nomológico deductivo, como lo denominaron Hempel y Oppenheim (1948). Como vimos, existen objeciones centrales a esta posición en la línea de que es solamente predictiva y no puede aclarar la estructura interna de los eventos que se aspira a explicar. Por ello, creo que es seguro afirmar que, al día de hoy, es solo un ideal regulativo general asumido por los científicos que cultivan algunos tipos de ciencia, sobre todo las ciencias naturales, antes que una práctica extendida en todas las actividades científicas.

Para el modelo probabilístico, el *explanandum* no se sigue necesariamente del *explanans,* sino con cierto grado de probabilidad. Este modelo es de particular utilidad en las ciencias sociales, pero cada vez más también en las ciencias naturales, sobre todo desde los desarrollos en biología y física cuántica.

De acuerdo con el modelo funcional o teleológico, el *explanans* es un objetivo, finalidad, propósito o función que hace inteligible al *explanandum*, es decir, que muestra por qué algo ocurrió de la manera como ocurrió, incluso si hubiera podido ocurrir de otra manera. Este modelo es de particular utilidad en la biología y en las ciencias humanas, especialmente en la psicología. Así, por ejemplo, podemos decir que una persona realizó cierta acción porque tenía determinado objetivo. En este caso, el objetivo sería la causa final de la acción, aunque, como es obvio, no en un sentido eficiente de causalidad. Sin embargo, como veremos en su momento, los estados mentales podrían tener un rol causal —tanto final como eficiente— en la explicación

de acciones como, por ejemplo, si decimos que alguien realizó una acción porque tenía cierta creencia o determinado deseo. En ese caso, la creencia y el deseo causaron eficientemente la acción y también nos permiten entenderla en función al objetivo de la acción. Más aún, como sostuvo Davidson en su famoso artículo «Actions, Reasons and Causes» (1980a [1963]), los estados mentales no solo causan acciones sino también son razones que las justifican.

Siguiendo al modelo genético, cierto conjunto de acontecimientos anteriores en el tiempo, que sería el *explanans*, hace inteligible a otro conjunto de acontecimientos posteriores en el tiempo, el *explanandum*, que no se habría producido de no ser por los acontecimientos anteriores que los explican genealógicamente. Este modelo explicativo comienza con una reconstrucción de acontecimientos y termina mostrando una articulación causal en ellos, lo que permite mostrar por qué ocurrió lo que ocurrió asumiendo que, en principio, pudo haber ocurrido algo distinto. Naturalmente, la ciencia que emplea de manera paradigmática este modelo es la historia, pero también podría serlo el psicoanálisis y la biología evolutiva. Es claro, sin embargo, que una misma ciencia puede hacer uso de varios de estos modelos explicativos.

Aunque hay debate respecto de qué tienen en común estos modelos, en los cuatro está presente la causalidad, a pesar de que podrían ser conceptos ligeramente diferentes de causalidad. En la discusión contemporánea hay un sentido amplio de causalidad, uno teleológico y uno específico o clásico, característico de la física. Según el sentido amplio, A es causa de B si B ocurrió porque A ocurrió. Según el sentido teleológico, A es causa de B si B es una acción que cierto agente realizó teniendo como propósito u objetivo que se produjera A. El sentido amplio no merece mayor debate porque es tan general que tiene poco contenido. Ha habido mucha polémica respecto de si el sentido teleológico responde efectivamente a un concepto de causalidad. Finalmente, el sentido clásico es el que más se ha discutido en la filosofía de la ciencia. Con frecuencia se ha asumido que es el único concepto de causalidad, lo que ha tenido como consecuencia que, al objetar su uso en algunos tipos de ciencias, por ejemplo, las sociales, se crea equivocadamente que se está eliminando toda causalidad en ellas. En algunos casos esto ha conducido a que se crea que esas disciplinas no son científicas, lo que muestra una concepción excesivamente estrecha de cientificidad y de causalidad.

Ahora concentrémonos en la causalidad clásica, cuya representante paradigmática es la física. Al pretender explicar la realidad, las ciencias naturales se proponen describir el mundo como un entramado de cadenas causales en el que cada evento está causalmente relacionado con otros, en un sentido clásico de causalidad que debe cumplir por lo menos con tres condiciones. A es causa de B si:

(1) Cada vez que ocurre A también ocurre B.

(2) Hay una regularidad —ya sea nomológica o no— que gobierna esta relación causal.

(3) A y B son conceptualmente independientes, es decir, pueden ser «concebidos» de manera independiente como eventos diferentes y no conectados entre sí.

En el caso de la interpretación de agentes intencionales, al desear hacerlos inteligibles, les atribuimos estados mentales porque intuitivamente asumimos que estos causaron los eventos que queremos explicar y comprender, a saber, sus acciones y otros estados mentales. En otras palabras, imaginamos qué estados mentales pudieron haber causado los eventos que queremos explicar y los atribuimos al agente. Aquí surge un problema con la tercera cláusula de la causalidad, porque es discutible si al interior de un escenario interpretativo los estados mentales atribuidos son conceptualmente independientes de los eventos que supuestamente causaron. En este terreno hay posiciones diferentes.

Una posición, influida por el segundo Wittgenstein (1958, 1988), considera inviable explicar causalmente el comportamiento humano a partir de estados mentales porque no hay independencia conceptual entre la causa y el efecto, es decir, entre los estados mentales y las acciones, ya que solo se puede determinar el contenido de las acciones a partir de los contenidos de los estados mentales que supuestamente los causaron y solo se puede determinar el contenido de los estados mentales a partir del contenido de las acciones que supuestamente causaron.

Un ejemplo es el siguiente: Lara observa a Diego caminando apresurado hacia un cine, a pocos minutos de que comience la función. Lo que en principio ella observa es solo un comportamiento físico, es decir, a una persona desplazándose de un punto a otro en el espacio. Ese evento físico será reconocido como una acción solo si ella le atribuye a él un conjunto de estados mentales que ella considera causaron el evento. Así, por ejemplo, ella le adscribe la creencia de que en ese cine proyectarán una película y el deseo de verla. Es solo gracias a que ella le atribuyó esos estados mentales que podrá reconocer y describir al evento físico como la acción de dirigirse al cine. Pero aquí hay una circularidad, porque ella reconocerá el comportamiento físico como una acción solo si previamente le atribuyó al agente estados mentales, los cuales serán adscritos en la medida en que ella pueda reconocer ciertos eventos físicos como acciones. En este ejemplo da la impresión de que la interpretación comenzara con el reconocimiento de que un evento físico puede también ser visto como una acción, pero no es así. Si ella cree que Diego detesta el cine, no hará esa atribución. Tampoco la hará si sabe que él trabaja en una oficina que queda al lado del cine. Eso muestra que para que ella reconozca al evento físico como una acción en particular necesita

haber atribuido ciertos estados mentales y no otros. ¿Dónde se inicia, entonces, la interpretación: en la acción o en los estados mentales que supuestamente la causaron? Se trata de un círculo hermenéutico en el que el todo da sentido a las partes y las partes son inteligibles en relación con el todo. No hay un único punto de partida sino varios posibles. Se reconoce la acción a la luz de los estados mentales atribuidos y los estados mentales a la luz de la acción atribuida. Y todo ello es parte de un proceso interpretativo más amplio y abarcador.

El punto es que, al parecer, no hay independencia conceptual entre los estados mentales y las acciones que las justifican al interior de una interpretación intencional, esto es, no se pueden «concebir» independientemente o, mejor aún, no se puede describir uno sin describir el otro. Según esa línea de argumentación, dentro de una interpretación intencional el contenido de la acción solo es concebible, definible y comprensible a partir de los estados mentales que supuestamente lo causaron. De igual manera, el contenido de los estados mentales atribuidos solo es concebible a partir de las acciones que supuestamente causaron, pues atribuimos estados mentales al agente a la luz de las acciones que queremos entender. Esto se sigue de la concepción holista de la interpretación. Es decir, estados mentales y acciones se interdefinen y explican mutuamente, por lo que no son conceptualmente independientes. Si eso es correcto no podríamos hablar de relaciones causales, en un sentido clásico, cuando describimos el comportamiento de los agentes intencionales en términos de estados mentales y acciones. Esta es la posición de, por ejemplo, Robin George Collingwood (1946) y Peter Winch (1958). Este es el problema de la conexión lógica.

Pero el abandonar la causalidad en este terreno tiene dos consecuencias: por una parte, convierte en misteriosa la relación entre estados mentales y acciones, como si no hubiera conexión alguna entre ellos o como si esa conexión fuese esencialmente diferente de la que hay entre cualesquiera otros eventos de la naturaleza. De otro lado, no toma en consideración el que las distintas disciplinas humanas —y también el habla coloquial— se expresan en términos causales cuando desean explicar por qué alguien actuó como actuó, con lo que le atribuyen a los estados mentales un significativo rol causal.

Aquí surge una aparente pero falsa dicotomía: ¿es o no posible explicar causalmente el comportamiento intencional humano? Si la respuesta es afirmativa y se usa un concepto clásico de causalidad, habrá que afirmar que hay independencia conceptual entre los estados mentales y las acciones al interior de una interpretación intencional. Si la respuesta es negativa habrá que desterrar el concepto de causalidad del ámbito de las ciencias humanas, lo que es un precio demasiado alto de pagar, además de que en la práctica nadie estaría dispuesto a pagarlo.

Para superar esta disyuntiva, Davidson (1980a [1963]) defiende tres tesis que a primera vista parecen incompatibles entre sí, a menos que, como desea ese autor, se adopte el monismo anómalo que él desea sostener. Las tesis son las siguientes:

(1) Las razones para actuar pueden ser causas de las acciones.

(2) Hay una regularidad nomológica que gobierna toda relación causal.

(3) No hay leyes psicofísicas ni psicológicas, aunque sí físicas.

La idea de Davidson es que los estados mentales causan y al mismo tiempo justifican otros estados mentales y acciones, de manera que al explicar por qué alguien hizo lo que hizo recurriendo a su red de creencias y deseos, es factible decir que esta explicación es causal y que al mismo tiempo expresa las razones que el agente tuvo para actuar como actuó, es decir, las justifica a ojos del agente, desde el punto de vista de una intérprete. Esto sería posible porque fue un evento físico —por ejemplo, cierta configuración neuronal— lo que causó el comportamiento físico del agente. Pero ese mismo evento físico —la configuración neuronal— puede ser descrito en términos intencionales como un estado mental —por ejemplo, como una creencia o un deseo—, de manera que el mismo evento es causa y también razón del comportamiento del agente.

El punto es que los estados mentales que causan y justifican las acciones y las acciones causadas sí pueden ocurrir separadamente —y por tanto son concebibles independientemente—, pero se hacen mutuamente inteligibles cuando son descritos asociados entre sí en una interpretación. La conexión lógica se da entre los contenidos proposicionales de las razones del agente y las acciones descritas de una u otra manera bajo cierta interpretación. Por el contrario, las relaciones causales se dan entre eventos, independientemente de cómo sean estos descritos. En otras palabras, las relaciones causales no dependen de nuestra interpretación de ellas, a diferencia de la explicación racional que se da en una narración interpretativa.

Una acción es un evento espacio-temporal que puede satisfacer distintos predicados y, por tanto, puede ser descrito de diferentes maneras como poseyendo distintas propiedades. Por ejemplo, la oración «El autor de *Trilce* escribe un poema» describe la misma acción que la oración «El autor de *Los Heraldos Negros* escribe un poema» o, incluso, que «El más grande poeta peruano nacido en Santiago de Chuco escribe un poema». En los tres casos estoy hablando de una acción realizada por César Vallejo, el escribir un poema, pero en cada caso estoy describiendo una propiedad diferente de ese poeta. Se trata del mismo evento descrito de diferentes maneras. Análogamente, un evento físico puede ser descrito de manera intencional haciendo explícita la acción sin que la acción sea diferente del evento físico.

Pueden ser muchas las razones del agente que dieron lugar a una acción y no hay manera de precisar la razón que tuvo mayor efecto causal, de todas las otras razones que el agente también pudo haber tenido para actuar pero que no fueron causalmente eficientes. En estos casos uno debe hacer una inferencia a la mejor explicación y buscar las razones que resulten más explicativas.

Ahora bien, las razones para actuar son causas de la acción tanto en un sentido eficiente como en un sentido final, pues hay un elemento normativo en la justificación que va más allá de la causalidad eficiente. Así entonces, mientras que la relación causal eficiente es independiente de una interpretación, el rol causal teleológico de una razón es parte de una interpretación.

Davidson se alinea con Aristóteles para sostener que encontrar las razones por las cuales alguien actuó es explicar causalmente su comportamiento, de manera que las razones que explican el comportamiento del agente son los estados mentales que lo causaron. Aristóteles (1985) explica el comportamiento intencional mediante un silogismo práctico que incluye como premisas creencias y deseos, y cuya conclusión es una acción. Esto también está cerca de la posición humeana de asumir que la explicación del comportamiento requiere de un elemento conativo o volitivo y uno cognitivo. En la primera sección de este capítulo vimos cómo el modelo de explicación causal de las ciencias procede en gran medida de Aristóteles. Vemos ahora que esta deuda también está presente en la explicación del comportamiento humano. El silogismo práctico aristotélico puede tener la siguiente forma:

- (Premisa 1) Tengo el deseo de realizar la acción A.
- (Premisa 2) Creo que puedo realizar la acción A.
- (Conclusión) Realizo la acción A.

El silogismo práctico es una reconstrucción racional del comportamiento del agente, pues es obvio que nadie actúa a partir de la consideración de premisas. Sin embargo, Aristóteles nota que las premisas no tienen que ser conscientes. Esto lo discutiremos en el último capítulo cuando abordemos el problema de la irracionalidad, pero lo que Aristóteles sostiene es que cuando uno actúa de manera irracional ha «olvidado» transitoriamente alguna de las premisas, de manera que estas siguen teniendo un rol causal y continúan justificando la acción, incluso si el agente no lo sabe, o por lo menos no sabe que lo sabe (*Ética Nicomaquea*, VII). La intuición aristotélica es muy fina porque está cerca de sostener un «olvido» freudiano que permita explicar cómo una persona puede actuar en contra de su mejor juicio o puede aceptar creencias que sean contradictorias con sus otras creencias. Así, las premisas del silogismo práctico no tienen que ser conscientes. Pueden ser no conscientes o incluso

inconscientes en un sentido psicodinámico, aunque evidentemente Aristóteles carecía de este último concepto.

Pero regresemos ahora a la relación entre acciones y eventos. Lo central en el argumento de Davidson es que los eventos del universo son de una sola categoría ontológica, de ahí el término «monismo», pero pueden ser descritos mediante el lenguaje[4] físico o el lenguaje intencional, según las necesidades explicativas que uno tenga. Al utilizar conceptos como «sinapsis», «hormona», «neurotransmisor» o «desplazamiento» estamos empleando el lenguaje físico. Si usamos conceptos como «creencia», «deseo», «intención» o «acción», estamos utilizando el lenguaje intencional. Ambos lenguajes versan sobre los mismos eventos, aunque los describen de diferente manera resaltando distintos aspectos de ellos, por lo que esta posición también ha sido llamada «dualismo de aspectos» o «monismo de aspecto dual». Se trata de un monismo ontológico asociado a un dualismo metodológico o epistemológico. Podría parecer sorprendente que un mismo evento tenga propiedades o aspectos diferentes, pero no lo es. Considérese, por ejemplo, un evento climático que puede ser al mismo tiempo físico, hermoso o destructivo, según el tipo de descripción que uno utilice para los fines que se proponga realizar. El evento es el mismo, aunque puede dar lugar a diferentes descripciones según los aspectos que uno desee resaltar.

De acuerdo con el modelo davidsoniano (1980a [1963]), las relaciones causales entre eventos son independientes del lenguaje que uno use para describirlos, pues la realidad se compone de eventos y estos existen de manera previa e independiente de nuestra descripción de ellos. No obstante, podemos construir leyes que gobiernan las relaciones causales si usamos la descripción física. Por el contrario, no podremos hacerlo si empleamos la descripción psicológica o intencional. Por otra parte, dado que solo hay un conjunto de eventos que pueden ser descritos mediante estos dos vocabularios, no tiene sentido decir que haya leyes que gobiernen las relaciones entre eventos físicos y psicológicos, porque no son dos conjuntos diferentes de eventos. La posición de Davidson es un naturalismo no reductivista de corte spinozista. Su propuesta central es que los estados mentales son estados físicos —esto es, lo mental y lo físico son distintas descripciones de los mismos eventos—, pero los conceptos mentales no son reducibles a conceptos físicos ni viceversa y las explicaciones mentales tampoco son reducibles a explicaciones físicas ni viceversa.

La posición suena razonable excepto en que no queda claro de qué manera existen los fenómenos psicofísicos como, paradigmáticamente, los psicosomáticos, cuya existencia nadie —y menos un psicólogo o un psiquiatra— estaría dispuesto a negar.

[4] Usaré «lengua» cuando desee referirme específicamente a las lenguas naturales, y «lenguaje» para referirme al sistema lingüístico y también a las lenguas naturales y artificiales.

Por otra parte, como algunos filósofos han notado, parece que solo tienen rol causal los eventos cuando son descritos físicamente, de manera que la descripción mental sería irrelevante desde el punto de vista causal y, por tanto, explicativo, si uno asocia explicación a causalidad. Esta es la célebre posición de Jaegwon Kim (1989, 2000, 2007) y de otros autores, quienes consideran que el naturalismo no reductivista contenido en el monismo anómalo de Davidson es inviable, precisamente porque, a juicio de ellos, toda la relevancia causal recaería en lo físico y lo mental resultaría prescindible, lo que convertiría a la posición de Davidson en una forma de epifenomenalismo. Sobre este punto regresaré más adelante.

Creo que la posición de Davidson se puede sostener, pero haciendo algunas cualificaciones. Es razonable afirmar que las relaciones causales se dan entre eventos, independientemente de cómo son descritos. También que la descripción física da lugar a regularidades con mayor o menor grado probabilístico, aunque la descripción intencional también puede dar lugar a regularidades, a pesar de que estas no son nomológicas. Sobre si la posición de Davidson es un tipo de epifenomenalismo en el que lo mental es causalmente irrelevante —como cree Kim—, pienso que puede sostenerse que el modelo davidsoniano es compatible con alguna forma de emergentismo, en la que los eventos que no solo pueden ser descritos como físicos sino también como mentales tienen relaciones causales más complejas, precisamente porque tienen propiedades emergentes que permiten la descripción mental además de la física. En el apartado 9.3 discutiré las concepciones más importantes acerca de las relaciones entre la mente y el cuerpo, y explicaré cómo se puede mantener la psicosomática en un modelo monista de aspecto dual. También intentaré aclarar de qué manera se relacionan causalmente los estados mentales y las acciones.

Volviendo a los rasgos de la causalidad, algunos filósofos incluirían una condición para utilizar este concepto en el sentido clásico. La relación causal entre dos eventos A y B es nomológica, es decir, está gobernada por regularidades que tienen forma de ley y que no tienen excepciones. Según esa condición, si hubiera una excepción a esa regularidad esta no constituiría una ley, y si no hubiese una ley que gobernara la relación entre A y B, A no sería causa de B. Aunque esta condición fue asumida como evidente desde Hume, según algunos autores hay un contraejemplo obvio: la singularidad que dio origen al universo en una Gran Explosión. Por definición, esta relación causal no podría estar gobernada por ley alguna, con lo cual tendríamos una relación causal no nomológica e irrepetible. Me inclino, sin embargo, por no considerar este aparente contraejemplo como concluyente, dado que poco o nada sabemos acerca de las relaciones causales que originaron el universo. Lo más probable es que ni siquiera podamos usar el concepto de causalidad para hablar del origen del universo, precisamente porque el concepto de causalidad nos permite entender fenómenos

que se dan al interior de este y constituiría por lo menos un exceso, sino un error categorial, aplicarlo para explicar el origen mismo del universo.

Como ya vimos, la noción tradicional y más conocida de explicación es la que recibe los nombres de modelo nomológico deductivo o modelo de cobertura legal. Lo característico de este modelo es sostener que explicar un evento es encontrar las relaciones causales que lo han generado, así como las regularidades nomológicas que gobiernan esas relaciones causales para, posteriormente, describir esas relaciones causales mediante una ley. Pero hay que notar que no es lo mismo una regularidad y una ley. La primera es una propiedad de la naturaleza según la cual el universo tiene un curso uniforme y no errático, mientras que la segunda es una propiedad de las teorías científicas —es decir, de nuestras construcciones conceptuales y lingüísticas— con que describimos la naturaleza, presuponiendo que estas descripciones carecen de excepciones. Así, por ejemplo, si quiero explicar por qué los metales se dilatan con el calor, deberé encontrar las relaciones causales que conectan esos eventos —la dilatación de los metales— con otros eventos naturales —por ejemplo, el someter objetos metálicos a cierta temperatura bajo determinadas condiciones del entorno—. Al encontrar esas conexiones habré descubierto las regularidades naturales que gobiernan tales eventos, con lo cual podré describir aquellas regularidades en forma nomológica, es decir, mediante una ley. A diferencia de lo que ocurre en las ciencias humanas, se asume que las leyes naturales no tienen excepciones. Si encontráramos un caso que no cumpliera con la ley tendríamos que admitir que hemos encontrado un contraejemplo y, por tanto, la ley habrá sido falsada, es decir, se habrá descubierto que no es una ley, o que es un caso anómalo y rebelde que requiere de una explicación más refinada.

El modelo de cobertura legal presupone el llamado «principio de la uniformidad de la naturaleza»[5], según el cual esta no es aleatoria ni errática sino que está determinada con una alta probabilidad, de tal manera que la misma causa produce, en casos semejantes y ante las mismas condiciones del entorno —*ceteris paribus*—, los mismos efectos. Recientemente la física cuántica nos ha llenado de dudas en relación con este presupuesto, pero aún no se sabe hasta qué punto sus evidencias modificarán nuestras concepciones clásicas de causalidad, de explicación natural o acerca de la uniformidad de la naturaleza. Además, aunque en un nivel cuántico no parece haber regularidades plenamente determinadas, los grados de azar y probabilidad —de comprobarse— serían demasiado pequeños como para obligarnos a cuestionar la tesis general de que la naturaleza está gobernada por regularidades. Por ello, a pesar del aparente indeterminismo cuántico, las ciencias naturales siguen asumiendo

[5] Así fue denominado por John Maynard Keynes (1943). Nicholas Rescher, por su parte, lo llama el «principio de la sistematicidad de la naturaleza» (1958).

como un presupuesto el hecho de que, en líneas generales, el futuro será semejante al pasado, que ningún acontecimiento es producto del azar y que todo evento tiene causas anteriores, de manera que el universo es una estructura de eventos físicos causalmente conectados entre sí de forma inexorable donde, como decía Hume, la causalidad es el cemento del universo.

Muchos filósofos no aceptan hoy el determinismo natural sino alguna forma de probabilismo. Esta posición, propuesta originalmente por Peirce (*CP*4 y 5) con el nombre de *tychismo*, sostiene que las regularidades que gobiernan el universo no están determinadas sino son solo altamente probables. La idea es que en un mundo aleatorio hay más probabilidades de que ocurran ciertas regularidades que otras. Pero incluso en ese caso, cuando uno desea explicar un fenómeno natural pretende que su descripción está en condiciones de predecir su curso futuro, con lo cual asume cierta regularidad cuyo conocimiento sería uno de los objetivos de la explicación científica. Por eso la explicación natural presupone o bien el determinismo de la naturaleza o, por lo menos, un alto grado de probabilidad. Asimismo, la explicación pretende describir hechos externos, públicos y objetivos, que pueden ser intersubjetivamente constatados y medidos todas las veces que sea necesario.

La filosofía de la ciencia ha cambiado mucho en las últimas décadas, en gran medida como reacción a los presupuestos del positivismo lógico a partir de la obra de Popper y, especialmente, de Thomas Kuhn, como veremos en el capítulo tres. Una concepción demasiado estrecha de la ciencia conduce a una visión equivocada de la práctica real de las ciencias naturales y a una desvalorización de las ciencias humanas. En la órbita del positivismo lógico y los filósofos influidos por este, con frecuencia se cometía el error de suponer que la física es el modelo del conocimiento ideal y que en tanto las otras ciencias se acerquen a ella serán más científicas. A veces se veía este ideal como el de la mayor matematización posible, es decir, se creía que mientras más matematizable fuese una disciplina sería más científica. Es discutible que ese ideal sea aplicable a todas las ciencias naturales, pero sin duda no es aplicable a las ciencias humanas, pues estas también tienen una metodología cualitativa. Pero es importante no recaer aquí en la dicotomía del siglo XIX. El que las ciencias humanas necesiten una metodología cualitativa no impide que también puedan tener una cuantitativa para algunas regiones de sus objetos de estudio. La filosofía de la ciencia pos-kuhniana relativiza muchas viejas dicotomías que ya no se aplican al mundo académico actual, como veremos más adelante.

Hasta aquí he sostenido que, aunque las expresiones «ciencias naturales» y «ciencias humanas» refieren a dos conjuntos difusos de disciplinas diferentes, tienen una gran intersección y muchas de las ciencias más interesantes, nacidas o desarrolladas desde fines del siglo XX, se encuentran en ese espacio. Sostengo que las ciencias naturales

y las ciencias humanas son explicativas, aunque las ciencias humanas también son comprensivas. En este apartado hemos discutido algunas relaciones generales entre explicación, comprensión e interpretación; ahora nos detendremos en la naturaleza de la interpretación de los agentes intencionales, sean estos individuales o colectivos.

1.3. Interpretar

Como hemos visto en el apartado anterior para el caso de las ciencias naturales, cuando se desea explicar un evento uno aspira a encontrar las causas físicas que lo produjeron, para luego hallar las regularidades naturales que gobiernan esas relaciones causales y posteriormente describir tales regularidades mediante leyes o, por lo menos, mediante descripciones con un alto grado de probabilidad.

Se suele llamar «método hipotético-deductivo» al proceso como esto se realiza. A partir de la observación de la naturaleza se elabora un conjunto de hipótesis acerca del tipo de relaciones causales y regularidades que podrían estar presentes. Luego se deducen las consecuencias observacionales que deberían ocurrir si las hipótesis fuesen correctas. Mediante sucesivas correcciones de las hipótesis y las deducciones observacionales, progresivamente se describen las relaciones causales y las regularidades que las gobiernan.

Normalmente se trata de muchas relaciones causales entrecruzadas y nunca un evento tiene una sola causa, aunque con frecuencia seleccionemos la que consideramos más relevante. Tampoco solemos explicar eventos individuales, sino, usualmente, muchos eventos integrados entre sí. Al buscar regularidades presuponemos que la naturaleza no es errática, sino que está determinada o, por lo menos, que tiene regularidades con altos grados de probabilidad de ocurrencia. El principio de la uniformidad de la naturaleza es, por supuesto, un postulado metafísico imposible de demostrar, pero necesario para toda explicación física. Todo esto nos permite hacer predicciones exitosas acerca del futuro y, por tanto, modificar la realidad según nuestras necesidades o adaptarnos a ella. En todos los casos asumimos que los eventos, las relaciones causales y las regularidades que las gobiernan son independientes de nuestra conciencia y voluntad. Tanto los eventos, las relaciones causales como las regularidades son contingentes, es decir, pudieron haber sido diferentes de cómo son si el universo hubiese tenido características distintas, lo que perfectamente hubiera podido ocurrir.

Ese es un resumen apretado de la manera en que explicamos en las ciencias naturales, pero en el caso de las ciencias humanas la situación es distinta. La diferencia principal es que la explicación natural presupone un punto de vista externo y con pretensiones de objetividad acerca de algo que carece de subjetividad. La comprensión, por su parte, solo es posible de criaturas intencionales dotadas de subjetividad y

se produce cuando logramos compartir algo de ese espacio personal, es decir, cuando logramos capturar algo de su punto de vista. Para hacer eso necesitamos interpretar al agente intencional. A fines del siglo XX se acuñó la expresión «giro interpretativo» para aludir a una tendencia en la filosofía y en las ciencias humanas, aunque en algunos casos también en las ciencias naturales, a abandonar la idea de que el conocimiento aspira a una mirada descontextualizada y «desde ningún sitio», para subrayar el lugar de la intérprete (véase Hiley, 1991 y Pedace, 2018, p. 208).

Así pues, la interpretación tiene como objetivo la comprensión y es la estrategia propia de las ciencias humanas. Como el objetivo de la interpretación es la comprensión, solo interpretamos a individuos o comunidades de individuos que —asumimos— están dotados de voluntad, esto es, que actúan con algún grado de libre albedrío. También asumimos que tienen mentes, es decir, sistemas cognitivos complejos que les permiten procesar información de su entorno físico, de los otros individuos con quienes comparten relaciones sociales y de ellos mismos, es decir, del funcionamiento de su cuerpo y de su propia mente. Igualmente, asumimos que tienen la capacidad de representarse la realidad que existe fuera de sus propias mentes, las representaciones que otros individuos tienen y también sus propias representaciones. Suponemos que se trata de criaturas dotadas de subjetividad, es decir, de un punto de vista propio que les permite tener conciencia de sí mismos en varios niveles de complejidad. En el capítulo nueve discutiré con mayor detalle qué es la mente y cómo se relaciona con el cuerpo; ahora me basta con afirmar que las ciencias humanas tienen como objetivo interpretar el comportamiento y la vida mental de individuos a quienes consideramos están dotados de agencia, mente, subjetividad y autoconciencia.

Así como en el caso de las ciencias naturales, al interpretar a los agentes intencionales buscamos las causas que produjeron su comportamiento y sus estados mentales, pero con una diferencia importante. Aunque muchas causas son externas al individuo, también hay causas internas, estas son sus estados mentales. La expresión «estado mental» alude a procesos psíquicos como creencias, afectos, deseos, voliciones, fantasías, sueños, temores, etcétera. Muchos de estos estados mentales están dotados de contenido proposicional, esto es, son actitudes acerca de una proposición, como, por ejemplo, mi creencia, deseo, anhelo, sospecha, temor, etcétera, de que «mañana lloverá», que «el candidato x ganará las elecciones» o que «tú me acompañarás al cine esta noche». Pero existen también estados mentales que carecen de contenido proposicional y son únicamente sensaciones o estados afectivos puramente fenoménicos, como un dolor o un estado generalizado de placidez.

Por otra parte, describimos algunos de los eventos naturales con los que están comprometidos los individuos a quienes atribuimos estados mentales como «acciones», es decir, como eventos físicos que están dotados de intencionalidad porque se dirigen

hacia objetos diferentes del estado mental mismo y han sido causados por sus estados mentales, con lo cual estos individuos pasarán a ser considerados agentes.

De esta manera, al interpretar a un agente solemos buscar una relación causal entre sus estados mentales y sus acciones y, por tanto, la interpretación incluye también un elemento de explicación causal, pero bastante más refinada. En algunos casos, asumimos que estas relaciones causales están gobernadas por regularidades —el agente suele comportarse de esta u otra manera, o las personas normalmente reaccionan de esta forma ante tales o cuales estímulos—, pero no creemos que estas sean regularidades nomológicas, es decir, que puedan ser descritas mediante leyes, precisamente porque asumimos que se trata de agentes que tienen algún grado de libre albedrío y subjetividad.

La interpretación es la actividad de atribuir sistemas interconectados y, en líneas generales, consistentes de estados mentales, acciones, significados y valoraciones al comportamiento de un agente intencional. El modelo interpretativo clásico en la filosofía del lenguaje y de la mente, del cual Davidson es un representante paradigmático, se basa en la atribución de creencias y deseos a los agentes, y de significados a sus proferencias verbales. Sin embargo, es claro que también les atribuimos valores y afectos —emociones, sentimientos y pasiones— y que no solo adscribimos significados a sus proferencias verbales —que son un subconjunto de las acciones—, sino a todas las acciones o por lo menos a aquellas que tienen un mínimo de regularidad. Cuando hacemos eso podemos decir que hemos comenzado a interpretarlo. Pero esta definición tan gruesa genera muchos problemas de detalle, algunos de los cuales comentaré ahora. La concepción de interpretación que desarrollaré está inspirada en el modelo davidsoniano, pero añade muchos otros elementos que resulta incierto si el propio Davidson hubiera aceptado, así como otros que ciertamente no lo hubiera hecho.

Un primer punto por señalar es que el fenómeno de la interpretación no siempre es consciente, esto es, con frecuencia reconocemos y atribuimos estados mentales a un agente sin percatarnos de que lo estamos haciendo. Eso ocurre, por ejemplo, cuando conocemos a una persona y nos cae particularmente bien o mal, nos genera desconfianza o nos inspira una inmediata sensación de paz y sosiego, aunque no sepamos por qué. Más aún, podría generarnos paz y sosiego, o malestar e incomodidad, sin que seamos conscientes de que es eso lo que estamos sintiendo. Podríamos tener una experiencia afectiva difusa que no llegamos a reconocer y que solo lograremos categorizar mucho tiempo después.

En esos casos, hay estados mentales o rasgos que estamos reconociendo en él aunque no tengamos metacognición de ello, es decir, aunque no tengamos otros estados mentales acerca de esos mentales que estamos teniendo y reconociendo.

Con el paso del tiempo y gracias a cierta experiencia tratando con personas y lidiando con nosotros mismos, seguramente descubriremos que él nos inspiraba confianza porque de manera no consciente le atribuíamos ciertos estados mentales y no otros. O quizá descubramos que lo que nos producía no era tanto confianza sino seguridad, o tal vez no seguridad sino cierta sensación de predictibilidad. Con mayor práctica interpretativa acaso lleguemos a tener estados mentales sobre otros estados mentales en varios niveles de intencionalidad, como, por ejemplo: «Su comportamiento de confianza excesiva en sí mismo me genera desconfianza, lo que, a su vez, me produce cierta incertidumbre respecto de mí mismo. Creo que eso es porque me hace recordar a un amigo de la infancia que parecía muy seguro cuando en realidad era una persona que sufría mucho. Eso me conmueve y me hace dudar de mi propia seguridad».

Un estado mental de solo un grado de intencionalidad es aquel que representa directamente el mundo como, por ejemplo, «Creo que Antonio está en Egipto»[6]. Un estado de dos grados de intencionalidad sería aquel en el que nos representamos la manera en que otro —o nosotros mismos— nos representa(mos) el mundo. Por ejemplo: «Creo que Claudia cree que Antonio está en Egipto» o «Creo que deseo que Antonio esté en Egipto». Un estado de tres niveles de intencionalidad sería, por ejemplo, «Temo que Claudia piense que Paola no desea que Antonio esté en Egipto». Uno de cuatro podría ser «Me atormenta pensar que me avergüence saber que mis emociones son tan evidentes». Un ejemplo de seis niveles sería: «Me atormenta pensar que me avergüence saber que creo que mis emociones son evidentes acerca de que Paola no desea que Antonio esté en Egipto».

Los niños entre tres y cuatro años pueden alcanzar hasta dos o tres grados de intencionalidad, y los seres humanos llegamos hasta cuatro o cinco niveles de intencionalidad sin confundirnos. A partir de seis niveles nos enredamos o simplemente nos perdemos en la cadena de niveles. El número de niveles de intencionalidad que alcanza la especie depende del crecimiento que ha alcanzado la corteza prefrontal, en los últimos tres millones de años de evolución de nuestro cerebro, que ha pasado de aproximadamente 450cc de volumen encefálico a 1450cc en promedio en la actualidad. A su vez, el número de niveles de intencionalidad que alcanza un niño depende de la madurez física, cognitiva y emocional de su cerebro, lo que depende de un desarrollo neurológico sano y de un entorno social apropiado.

El punto es que atribuimos de manera consciente o no consciente estados mentales —ya sea en dos o más niveles de intencionalidad— sobre la base de un fondo no consciente de estados mentales que tenemos, reconocemos en el otro y atribuimos,

[6] O simplemente «Antonio está en Egipto», donde el «creo» es tácito. Según la teoría de la redundancia, «Creo que p», «Creo que p es verdad» o «P» dicen exactamente lo mismo.

a partir de mecanismos automáticos afectivos y cognitivos que nos toma la vida llegar a conocer. Ese fondo de estados mentales no conscientes —propios y reconocidos en el otro— hace posible que tanto bebes como adultos interactúen con los demás de manera muy intuitiva. Es conocido que los infantes tienen reacciones inmediatas de confianza o desconfianza en otras personas, sin que puedan dar cuenta de las razones que los han motivado. Y en gran medida parte de lo que hacemos a lo largo de nuestras vidas es aprender a reconocer nuestros propios estados mentales y los ajenos, para conocernos mejor y para conocer más a quienes nos rodean. Si vemos este tema desde una perspectiva filogenética, observaremos que los homínidos reconocían y atribuían estados mentales de manera no consciente y que progresivamente fueron haciéndose conscientes de ello, primero en solo dos grados de intencionalidad y luego en varios. Posteriormente la cultura, la literatura, el arte y la experiencia vital misma hacen que muchas personas logren una mayor finura al reconocer la sutileza y los matices de sus propias experiencias cognitivas y afectivas, así como las que encuentran en los otros. Ello requiere de toda una educación cognitiva y afectiva. Sin embargo, la estructura fundamental de atribución psicológica consciente se mantiene cuando reconocemos estados mentales de manera no consciente, así que podemos reconstruir lo que ocurre en el primer caso para explicitar lo que también acontece en el segundo.

En este punto voy a hacer algunas distinciones fundamentales que reaparecerán en otros momentos con mayor complejidad. Por estados mentales «conscientes» entenderemos aquellos sobre los cuales tenemos experiencia fenoménica. Por ejemplo, la sed o el dolor. En principio cualquier especie con un sistema nervioso central mínimamente desarrollado tendría este tipo de conciencia, que también suele denominarse «conciencia nuclear»[7].

Por estados mentales «autoconscientes» entenderemos aquellos que implican por lo menos dos grados de metacognición, es decir, estados mentales sobre otros estados mentales. Cuando esto ocurre se suele hablar de conciencia reflexiva, conciencia autobiográfica o autopercepción. Por ejemplo, si no solo tengo un dolor sino además me preocupa tenerlo.

Los estados mentales «preconscientes» son aquellos que se encuentran almacenados en nuestro aparato psíquico, aunque no seamos siempre conscientes de ellos.

[7] Múltiples experimentos sugieren que los mamíferos, las aves, los crustáceos y los peces —no así los insectos— tienen experiencia fenoménica del dolor. Véase Godfrey-Smith, 2017, capítulo 4, aunque ese autor llama «experiencia subjetiva» a lo que yo prefiero llamar «experiencia fenoménica» o «conciencia nuclear», y reservo la expresión «subjetividad» para la autoconciencia, pues esta última involucra por lo menos dos niveles de intencionalidad, así como una perspectiva del mundo y de sí mismo. Hasta donde se cree, los sistemas nerviosos habrían comenzado su evolución hace aproximadamente 700 millones de años y la experiencia fenoménica del dolor se habría originado en el período cámbrico de la era paleozoica, hace unos 541 millones de años.

Por ejemplo, si le pregunto a usted cuál es la capital de Turquía probablemente me responda que Ankara. Ahora le resulta un estado mental consciente, pero antes que se lo preguntara se encontraba de manera preconsciente.

Podemos tener estados mentales «inconscientes» en un sentido psicodinámico, que es lo que ocurre cuando se trata de aquellos que han sido reprimidos en un sentido psicoanalítico. Si, por ejemplo, me resulta intolerablemente doloroso tener cierta creencia, mi aparato psíquico podría reprimirla como un mecanismo de autoprotección, con lo cual se vuelve inconsciente. Pero esto no significa que haya sido eliminada de mi mente, pues casi son certeza reaparecerá generando efectos causales en mi comportamiento —o en otros estados mentales míos— de manera involuntaria y sin que yo tenga conciencia de ello. Como se sabe, Freud (2002) realizó el análisis más elaborado de la noción de inconsciente dinámico —es decir, el que es producto de la represión— a comienzos del siglo XX.

También hay estados mentales «no conscientes», siempre que puedan llegar a ser conscientes o puedan representar algo diferente de sí mismos. En este caso, estamos hablando de procesos cognitivos que gobiernan nuestro comportamiento generando representaciones del mundo. Sin embargo, un estado que no sea consciente ni pueda nunca por principio llegar a serlo y que tampoco represente nada diferente de sí mismo, no sería un estado mental sino simplemente un estado físico. Ejemplo de estos estados son la circulación de la sangre o los procesos electroquímicos del cerebro.

Daniel Wagner postuló en 2002 la existencia de un «inconsciente adaptativo» (Wilson, 2004) que realiza procesos cognitivos y afectivos —y en esto se diferencia del «inconsciente cognitivo» de la psicología cognitiva conductual— cuando la conciencia de ellos es innecesaria o contraadaptativa. Este nos permite procesar y lidiar de manera más rápida, con menor gasto de energía y de forma eficiente y automática con el mundo físico y social, lo que permitiría concentrar nuestras conciencias solo en aquellos procesos en los que la experiencia fenoménica sería indispensable. Ejemplos de esto último son el placer, el dolor y las emociones. Pero con la evolución del cerebro surgieron también aquellos procesos que requerirían de deliberación explícita, aunque es debatible si la conciencia sería un beneficio para ello. Esto ocurriría en un nivel filogenético y ontogenético. En el primer caso, se trata de procesos elegidos por la selección natural para favorecer la supervivencia; en el segundo caso se trata de procesos que el individuo ha automatizado a lo largo de su desarrollo, como montar bicicleta, practicar un deporte, «recitar» un argumento clásico o reaccionar emocionalmente ante determinadas circunstancias que tienden a repetirse.

Pero también necesitamos volver a preguntarnos a qué o a quién deberemos considerar un agente intencional y qué significa estar dotado de «intencionalidad».

Desde 1874 cuando Franz Brentano publicó *La psicología desde un punto de vista empírico* (1995), el concepto de intencionalidad ha sido empleado de varias maneras, pero siempre asociado a la característica principal —casi definitoria— de la mente humana. Brentano retomó este concepto de la escolástica para referir a la propiedad que tiene la mente de estar dirigida hacia algo diferente de sí misma: su objeto intencional. En la discusión actual hay dos propiedades que se suelen reconocer en la intencionalidad: propósito y capacidad representacional.

Con frecuencia se ha asumido que para que un evento físico pueda ser descrito como intencional es necesario que su agente sea consciente, lo que deja fuera los estados inconscientes que parecen también ser intencionales en un sentido psicoanalítico, así como los procesos no conscientes que son parte de nuestro comportamiento y quizá constituyen la mayor parte de él. A veces se entiende lo intencional como voluntario, lo que no solo abre el debate sobre la naturaleza de la voluntad sino también plantea la pregunta de si podría haber comportamiento involuntario que, sin embargo, pudiera ser considerado causado por estados mentales y, por tanto, ser reconocido como un conjunto de acciones. También se ha entendido la acción intencional como aquella dotada de sentido, es decir, como parte de un lenguaje o de un sistema de comunicación regulado por normas (Salmon, 2002; Sperber, 1985, 1996). Más frecuentemente se entiende «intencional» en el sentido de representacional. De esta manera, un estado es intencional si está dirigido a algo diferente de sí mismo, es decir, si refiere a algo, es acerca de algo o posee lo que se ha denominado *aboutness*. John Searle (1983, 1992), por ejemplo, define la intencionalidad como la capacidad de la mente de representar —o referir a— algo distinto de sí mismo. También usa el concepto de «intencionalidad colectiva» (1995) para aludir a las creencias, deseos o intenciones compartidas por varios agentes individuales.

Mi posición, siguiendo a Anscombre (2000) y Davidson (1980b), es que una acción es un evento físico si también puede ser descrito intencionalmente, es decir, como causado por estados mentales y dotado de intencionalidad. Más allá de esos autores, sostendré que un estado mental es un estado físico que posee por lo menos una de dos propiedades: conciencia o intencionalidad. Esto supone que la intencionalidad no es necesariamente consciente. Un agente, por otra parte, es una entidad —normalmente una criatura y probablemente solo un ser humano— que es capaz de causar acciones de manera autónoma a partir de sus propios estados mentales. Pero para que estas definiciones no sean circulares será conveniente desarrollarlas.

Los estados mentales son procesos psíquicos, ya sea intencionales o conscientes, o ambas cosas, como creencias, deseos, afectos, voliciones, sensaciones, fantasías, dolores, etcétera. Como ya vimos, las acciones son eventos físicos que pueden ser descritos como causados por un agente intencional, es decir, por una criatura que ha

realizado un evento de manera representacional, motivado por sus estados mentales y con un propósito u objetivo ulterior. El primer autor en describir los estados mentales fue Aristóteles (*Acerca del alma*) y fue también él quien utilizó una expresión completa para referir a ellos. En su tratado *Sobre la interpretación* (16a 4-9) este filósofo discute las relaciones entre el lenguaje escrito, el lenguaje hablado, los estados mentales y los hechos del mundo representados por la mente y el lenguaje (Para un análisis de estas relaciones en Aristóteles véase Quintanilla, 1990). Aristóteles acuña la expresión *ta pathémata tes psychés*, cuyo análisis es iluminador. *Pathémata* es una forma plural de *pathos*, que fue traducida por los latinos como *affectio* y pasó al castellano como «afecto», con el significado de aquello que nos afecta. *Tes psychés* es una forma en genitivo de *psyché,* que los griegos entendían como el lugar del pensamiento y la causa del movimiento y las sensaciones. De hecho, Aristóteles definía la *psyché* (*Acerca del alma*) como *arché tes kinéseos kai aísthesis,* es decir, principio del movimiento y de la sensación. Los latinos tradujeron *psyché* por *anima* y así fue como pasó al castellano «alma». El griego antiguo *psyché* es un sustantivo del verbo *psychein,* que significa respirar, por lo que hay una asociación entre *psyché* y «lo que respira». A su vez, las palabras latinas *anima* y *animus* que son cognados del griego *ánemos* —viento— también significan hálito y principio vital. Todos estos términos proceden del protoindoeuropeo. Así, *ta pathémata tes psychés* sería traducido al castellano como «las afecciones del alma», en el sentido de todo aquello que afecta a la vida psíquica, que es aproximadamente lo que hoy llamamos «estados mentales».

Como señalé anteriormente, los estados mentales pueden o no incorporar un contenido proposicional. Si lo hacen, son actitudes respecto de una proposición que describe el mundo, si no lo hacen son solo experiencias afectivas sin contenido concreto. Las creencias, por ejemplo, son estados mentales con contenido proposicional, en las que ese contenido es la representación del mundo que incorporan. Así, si yo creo que «Antonio está en Egipto», el contenido de mi creencia es la representación de que Antonio está en Egipto. Como las creencias describen un estado del mundo, sus contenidos son portadores de verdad, esto es, pueden ser verdaderos o falsos, según si el mundo es como es descrito o no. Cuando las creencias son sobre el pasado de uno mismo constituyen recuerdos y cuando son sobre nuestro propio futuro son expectaciones. Si un estado mental no incorpora representaciones suele ser denominado «no proposicional». Un estado mental no proposicional podría ser, por ejemplo, un dolor o una sensación muy básica. Así es como los bebes muy pequeños y los animales tienen estados mentales no proposicionales.

Aunque existen muchos tipos de estados mentales —la frontera entre ellos tiene una base neurológica, pero en gran medida es cultural y establecida por el lenguaje—,

para los efectos de interpretar a los demás seres humanos podemos establecer cuatro grupos de estados mentales principales. Las creencias, que son representaciones de cómo es el mundo para nosotros; los deseos, que son representaciones de cómo quisiéramos que fuese el mundo; los valores, que son representaciones de cómo creemos que debería ser el mundo; y los afectos, que nos informan sobre la manera como los acontecimientos del mundo impactan en nosotros mismos. Los afectos, a su vez, se pueden clasificar en tres grandes grupos según su intensidad y duración temporal. Las emociones —por ejemplo, el amor, el odio o la gratitud— son intensas y duraderas; las pasiones —como la ira, los celos o la fascinación ante algo— son intensas y poco duraderas, y los sentimientos —como la ternura, la placidez o la ansiedad— son poco intensos y duraderos. En todos estos casos, por supuesto, hay gradualidades y las fronteras no son nítidas.

Es importante insistir en que cuando uno interpreta a un agente y le asigna creencias, deseos, valores y afectos, los estados atribuidos se interdefinen mutuamente, de manera que sus contenidos —para los efectos de la interpretación— dependen de los contenidos de los otros estados atribuidos. Este es el holismo de lo mental, tema sobre el que regresaremos.

Pero cuando interpretamos a alguien no solo le atribuimos estados mentales sino también asignamos significados a sus proferencias verbales y otras acciones intencionales. El concepto de significado tiene por lo menos dos sentidos importantes. Por un lado, solo se puede atribuir significado a una acción intencional —una expresión lingüística, un gesto, un movimiento corporal, etcétera — que sea parte de una práctica social regular que suele tener una función comunicativa en una comunidad específica. El significado de esta acción intencional sería precisamente la práctica social que gobierna su uso al interior de una comunidad de agentes intencionales. Sobre este punto volveremos en el capítulo cuatro. Por otro lado, el significado —también llamado «significación»— es la valoración o importancia que un agente o grupo de agentes confieren a una acción o cosa, en relación con los propósitos y objetivos de ese agente o esa comunidad. Ambas acepciones de significado deben ser tomadas en cuenta cuando se interpreta a un agente, pero, por razones de claridad, emplearé «significado» solo en el primer sentido y hablaré de «valor» para referir al segundo sentido.

Ahora bien, como ya mencioné, el objetivo de la interpretación es la comprensión del agente, y esto incluye la capacidad de compartir algo de su punto de vista subjetivo: percibir, valorar, evaluar o experimentar las cosas, los acontecimientos y a las personas como él lo hace, es decir, participar en algunos de sus estados mentales sin por ello perder los propios. Más aún, y a diferencia de lo que ocurre con la explicación, cuando uno interpreta las acciones intencionales de un agente asume que este

es libre de actuar como actúa, pues podría haber obrado de una manera diferente de haberlo deseado; de otra manera no estaríamos hablando de acciones intencionales sino solo de eventos naturales. Así, la intérprete describe a los agentes como iniciando una cadena causal. En ese sentido, los agentes son delineados como, por así decirlo, primeros motores o causas incausadas que mueven *ex nihilo*, no porque esas sean propiedades reales de los agentes, sino porque eso es lo que una intérprete debe suponer si quiere interpretar intencionalmente al agente y no solamente explicar físicamente su comportamiento. Con frecuencia, sin embargo, mezclamos ambos niveles y atribuimos libre albedrío a los agentes, aunque también asumimos que hubo causas que condicionaron su comportamiento restringiendo su ámbito de elección.

En este punto surge una paradoja en la que no nos internaremos porque daría para una investigación en sí misma: cuando interpretamos a un agente explicamos su comportamiento causalmente, es decir, establecemos los estados mentales que consideramos causaron sus acciones, pero al mismo tiempo le asignamos libre albedrío, es decir, asumimos que actuó como lo hizo, aunque pudo haberlo hecho de una manera diferente. La paradoja radica en que asumimos que actuó causado por sus estados mentales, pero al mismo tiempo creemos que lo hizo libremente. Podríamos explicar por qué tuvo los estados mentales que tuvo y así consideraremos causas externas al sujeto —como acontecimientos del mundo u otras personas— y también causas internas a él —como otros estados mentales que causaron los estados mentales que causaron sus acciones— pero en ningún lugar de esta estructura causal encontraremos nada parecido a lo que llamamos «voluntad» ni tampoco diríamos que hay estados mentales o acciones totalmente incausadas que surgieron de la nada. Aun así, al interpretar a un agente, lo asumimos libre.

Solemos asumir que el ámbito de libertad se reduce en tanto el agente tiene menos control de su comportamiento —esto es, en tanto tiene menos capacidad para obrar de una manera diferente de como lo hizo—, con lo cual también se reduce su responsabilidad moral, ya que esta es directamente proporcional al grado de libre albedrío atribuido. No voy a discutir ahora hasta qué punto ese presupuesto es válido y si es posible conciliar la atribución de libre albedrío con el determinismo natural, solo me limitaré a señalar que eso es lo que solemos asumir cuando interpretamos a un agente, es decir, cuando describimos su comportamiento en términos de estados mentales y acciones.

Si la atribución de libre albedrío se reduce hasta desaparecer abandonaremos progresivamente la interpretación, esto es, la descripción intencional, para pasar únicamente a la explicación al interior de la descripción física. Eso pasa con una persona cuando por enfermedad, sea física o mental, dejó de tener control sobre su comportamiento, de manera que ya no hablamos de acciones sino solo de eventos físicos.

Pero también podría ocurrir lo contrario, que ampliemos la atribución de libre albedrío y responsabilidad moral, asumiendo que el comportamiento del agente no solo puede ser explicado físicamente, sino que también puede ser interpretado intencionalmente. Este es el caso, por ejemplo, cuando consideramos que un infante ya no es solo un conglomerado de fuerzas biológicas, pues progresivamente se convierte en un agente dotado de intenciones y estados mentales, lo que en su momento generará conciencia, libre albedrío y responsabilidad moral. También podría pasar si llegamos a la conclusión que una especie animal no es únicamente un conjunto de fuerzas de la naturaleza, sino que tiene rudimentos de lo que llamamos «intencionalidad». Si así ocurriera, progresivamente atribuiríamos a los miembros de esas especies derechos o quizá hasta responsabilidades, por lo menos en algún grado.

Lo que tenemos aquí es un continuo y no hay una frontera clara que separe a criaturas que solo pueden ser explicadas como objetos físicos de agentes que también pueden ser interpretados intencionalmente. No hay una línea precisa que señale cuándo un niño ya es un agente moral y tampoco tenemos razones incontrovertibles que indiquen que solo los seres humanos pueden ser descritos intencionalmente. No es imposible que mientras más conozcamos a los primates superiores y a otros mamíferos, sobre todo sociales, estemos más dispuestos a atribuirles algún grado de intencionalidad, con todo lo que esto implica.

Pero mientras la descripción física soporta un sentido clásico de causalidad en que la causa y el efecto son conceptualmente independientes, parece que la descripción mental no soporta ese sentido porque, como ya señalé, el contenido del estado mental y el de la acción que fue supuestamente causada por el estado mental se interdefinen mutuamente. Eso no significa que los estados mentales no sean causa de las acciones, lo son, pero la relación causal se da en el nivel de los eventos en sí mismos, aunque nosotros los describamos con un lenguaje físico o con un lenguaje intencional. Al usar la descripción física hablamos de causalidad clásica, pero en el caso de la descripción intencional también nos referiremos a la causalidad funcional o teleológica. Todo eso en el ámbito de las descripciones. Pero ello no nos impide afirmar que los estados mentales causan acciones, porque los estados mentales son idénticos a los eventos físicos. Sobre esto volveré en el capítulo 9, pero ahora deseo redondear un poco más la idea.

Para formularlo con más claridad, es posible sostener que un estado físico, por ejemplo, la presencia o ausencia de serotonina en el cuerpo de un agente, causó en un sentido clásico otro evento igualmente físico. Pero si decimos que su creencia de que «*Las troyanas* de Eurípides refleja la condición efímera de la suerte humana» causó la experiencia fenoménica de su melancolía, estamos haciendo un uso de causalidad clásica siempre que admitamos que tanto la creencia como la experiencia

de la melancolía son idénticas a ciertos fenómenos físicos que acontecen en su cuerpo. Lo importante es tener claro que las relaciones causales se dan entre eventos, independientemente de las descripciones que hagamos de ellos.

Si aceptamos que un estado mental es idéntico a un estado físico solo que bajo otra descripción, es posible decir que bajo la descripción física hay una relación causal clásica entre la causa y el efecto. Bajo la descripción intencional puede haber una relación causal clásica y también una teleológica, porque la descripción intencional sugiere una motivación ulterior para realizar la acción.

Volvamos un momento a la idea de que las razones por las que alguien actuó son también las causas de su acción. Cuando una intérprete adscribe a un agente ciertas razones para actuar se propone resaltar algo que el agente consideró o tuvo en cuenta para actuar como actuó. Al hacer esa atribución, el comportamiento del agente se hace inteligible para la intérprete, pues ella puede reconocer en las razones del agente una apropiada motivación para su comportamiento. A Davidson (1980a [1963]) le parece importante sostener que las razones del agente también fueron las causas de su acción, de manera que así se podrá explicar su comportamiento tanto de manera causal como racional.

Pero, como ya hemos visto, aquí surge un problema, porque para que haya una relación causal entre dos eventos estos tienen que ser conceptualmente independientes, es decir, no deben tener conexión lógica, porque si no la explicación causal sería circular y es obvio que si decimos que un estado mental del agente fue su razón para realizar su acción, estamos estableciendo una conexión conceptual entre los dos eventos, la razón y la acción. Esto ha conducido a que muchos autores sostengan que las razones por las que uno realizó una acción no pueden explicar causalmente las acciones.

La respuesta de Davidson a esta objeción es sofisticada. En primer lugar, distingue entre «razones para actuar» y «razones por las que se actuó» (1980a [1963], p. 9). La razón «para» actuar es la que tuvo el agente en mente, mientras que la razón «por» la que actuó es la que tuvo la eficiencia causal, independientemente de si el agente fue consciente de ello. En el capítulo doce discutiremos el rol de los estados mentales inconscientes en nuestro comportamiento, pero ahora nos concentraremos solo en la distinción que acabo de mostrar. Supongamos, por ejemplo, que el agente tuviera una razón para asistir a una reunión de trabajo —debía presentar un informe—, pero finalmente asistió porque su jefe lo llamó y le dijo que lo estaba esperando, y no quiso quedar mal con él. Hubo dos razones que justificaron su acción —cumplir con su obligación y no quedar mal con el jefe—, pero solo una de ellas tuvo el rol causal. En este caso la razón que tuvo el rol causal es la que explica la acción, aunque en algunos casos ambas razones podrían también coincidir y ambas podrían tener rol causal.

Ahora bien, hay conexión conceptual entre la causa —la razón para actuar— y la acción, según cómo se describan ambas. Desde un punto de vista físico, es claro que no la hay, aunque sí desde un punto de vista intencional o psicológico. Pero la relación causal es anterior e independiente de la descripción que hagamos de ella. Este es un tema complicado sobre el que Davidson ha dejado algunos puntos oscuros, pero pienso que es factible decir que mientras la descripción física ilumina el rol causal en un sentido clásico de causalidad, la razón para actuar iluminar el rol causal en un sentido de causalidad que puede también ser el clásico, pero que adicionalmente es un sentido teleológico.

De otro lado, no es el caso que cada vez que ocurra la causa —el sistema integrado de estados mentales— deba ocurrir necesariamente el efecto —las acciones—, dado que los estados mentales y las acciones son atribuidos por una intérprete según sus propios criterios de interpretación. Asimismo, como los estados mentales no son conceptualmente independientes de las acciones que pretenden explicar, las acciones del agente solo se entienden a la luz de ciertos estados mentales atribuidos, los cuales a su vez se entienden en relación a las acciones que estos pretenden explicar. En otras palabras, reconocemos un evento como una acción de un tipo y no de otro en función a las creencias, deseos, afectos y valores atribuidos al agente. De igual manera, atribuimos una creencia a un agente a la luz de los deseos, afectos, valores y acciones que creemos reconocer, y adscribimos un deseo a la luz de las creencias, valores y acciones reconocidas. Y así en adelante. Aquí hay un círculo holista en el que cualquier modificación en la atribución de estos elementos a un agente generará inevitables modificaciones en los otros elementos atribuidos, dado que el contenido de cada uno de ellos se puede adscribir, reconocer y entender solo a la luz del contenido de los otros, y siempre desde el punto de vista de la intérprete y en relación con el mundo que ambos comparten.

Volvamos ahora a la discusión sobre la naturaleza de la descripción intencional que es, propiamente, lo que he llamado «interpretación». Sostendré que los rasgos más notables de este tipo de descripción son el carácter holista, racional, teleológico y normativo de la atribución de estados mentales, significados y acciones a los agentes intencionales.

Un primer rasgo de la descripción intencional es el holismo, en tanto esta descripción asume que solo podemos atribuir conjuntos coherentes e interconectados de estados mentales, significados, valores y acciones. Naturalmente, decir que podemos atribuir una acción a un agente equivale simplemente a decir que podemos describir su comportamiento como realizando una acción intencional. No sería posible atribuir a alguien un estado mental aislado ni un significado o una acción que se encontraran desconectados de un sistema de ellos. Es más, el contenido de un estado

mental, de un significado o de una acción procede de las interconexiones que este tiene con otros estados mentales, significados y acciones[8]. Así, por ejemplo, el contenido de mi creencia que «Platón escribió el *Timeo*» depende de mis creencias acerca de los objetos referidos por «Platón» y «*Timeo*», lo que a su vez depende de un corpus más amplio de creencias que tenga acerca de la filosofía griega. Entonces, si yo creo que «Platón» es el nombre de un filósofo griego nacido en Atenas, mientras que mi vecino cree que es el nombre de un compositor napolitano del siglo XIX; y si yo creo que «*Timeo*» es el nombre de un diálogo filosófico mientras que mi vecino cree que es el nombre de una ópera, tanto mi vecino como yo creemos que «Platón escribió el *Timeo*», pero nuestras creencias serán muy diferentes y eso dependerá de las otras creencias con las que cada creencia esté asociada. La noción de holismo se encuentra cercana a la idea de círculo hermenéutico, en tanto sostiene que el contenido de un estado mental depende de sus interconexiones con un sistema de ellos y que el sistema está, a su vez, constituido por los contenidos de los estados mentales específicos. La estructura de la red de estados mentales depende de los contenidos de sus elementos y los contenidos de los elementos dependen de su lugar en la red. Una modificación en el contenido de un estado mental tendrá consecuencias, más o menos severas según el caso, para todo el sistema en su conjunto.

Un segundo rasgo de la interpretación es la racionalidad. Mientras la descripción física tiende a explicar mostrando que un fenómeno está en conformidad con una regularidad, la interpretación del comportamiento intencional funciona de una manera diferente. Una instancia de comportamiento resulta inteligible para una intérprete si ella lo encuentra razonable. El criterio que guiará las atribuciones de la intérprete será sus propias concepciones acerca de lo que es verdad y lo que resulta razonable creer y hacer dada la evidencia disponible. La intérprete usará como criterio para la atribución de estados mentales al agente, los estados que ella imagina que tendría si fuera él y estuviera en las circunstancias en que ella cree que él se encuentra. Para interpretar el comportamiento intencional de un agente es condición necesaria hallarlo básicamente racional y eso es encontrar suficiente consistencia entre sus estados mentales. Usaré «consistente» en un sentido lógico cuando me refiera a creencias con contenido proposicional. En todos los otros casos usaré la expresión «comportamiento consistente» en un sentido más amplio para aludir a un tipo de comportamiento que, a los ojos de la intérprete, no muestra incompatibilidades entre estados mentales y acciones.

[8] A lo largo de todo el libro y por razones de brevedad, cada vez que hable de estados mentales del agente o de la intérprete, incluiré también sus acciones y los significados asignados a las acciones, sean estas comportamientos físicos intencionales o proferencias verbales. También incluiré la significación o valoración dada a los estados mentales y acciones. Incluiré también estados mentales en más de un nivel de intencionalidad, es decir, estados mentales sobre estados mentales, ya sea del agente o de la intérprete.

Los estados mentales y las acciones son incompatibles entre sí si se excluyen mutuamente. Esto ocurre, por ejemplo, si dos acciones parecen dirigirse a finalidades incompatibles, si un subconjunto de estados mentales excluye a otro subconjunto de ellos o si un subconjunto de estados mentales excluye a un subconjunto de acciones. Sin embargo, en todos estos casos la inconsistencia debe encontrarse al interior de una interpretación más amplia. En otras palabras, un estado mental aislado no puede excluir una acción; debe tratarse de un conjunto de estados mentales interconectados y básicamente consistentes, que excluyen un subconjunto de acciones igualmente interconectadas y básicamente coherentes entre sí. Pero la inconsistencia que nos concierne está en el comportamiento del agente según la interpretación de la intérprete, dado que los eventos físicos son reconocidos como acciones de acuerdo a los criterios que ella tenga en circunstancias interpretativas específicas.

Es indiscutible que con frecuencia debemos atribuir a los agentes inconsistencias entre sus estados mentales, es decir irracionalidad, pero lo podemos hacer solo al interior de un sistema más amplio de estados mentales que consideramos consistente y con el objetivo de encontrar al agente básicamente inteligible. En otras palabras, cuando nos vemos obligados a atribuir irracionalidad lo hacemos con el objetivo de poder encontrar al agente básicamente inteligible, es decir racional. Sobre este tema volveré en el capítulo doce.

Un tercer rasgo de la interpretación es que esta tiene un componente normativo, porque cuando atribuimos a un agente acciones o estados mentales no los vemos gobernados por leyes físicas generales, sino estando en conformidad con ciertas normas o prácticas sociales. Al adscribir estados mentales y acciones atribuimos también al agente compromisos normativos respecto del futuro[9]. Al atribuir al agente estados mentales lo comprometemos con otros estados mentales que él debería tener para que lo podamos seguir considerando inteligible, es decir, racional. No solo tenemos creencias acerca de cómo él debería comportarse, creer, valorar y sentir, sino lo comprometemos con tal comportamiento y tales estados mentales, y esos compromisos normativos serán parte de nuestra interpretación de él. Por ejemplo, si le atribuimos la creencia de que p y también la creencia de que p implica q, creemos que en condiciones ideales él debería creer que q. Asimismo, si le atribuimos la creencia de que x es la acción correcta por seguir, el deseo de

[9] Esta idea es discutida ampliamente por Carlos Moya, quien cree que lo que distingue la acción del mero acontecimiento es la intencionalidad, que debería ser entendida como cierto tipo de compromiso respecto del futuro. Dice: «nuestra habilidad para comprometernos para hacer cosas en el futuro es una parte esencial de la agencia y de nuestra conciencia de ser agentes» (1990, p. 48). Esta y todas las traducciones son mías, a menos que diga explícitamente lo contrario. Pedace también insiste en el carácter normativo de toda interpretación (2017).

hacer lo correcto y la creencia de que no hay ningún obstáculo para hacerlo, le atribuiremos también el compromiso normativo de realizar la acción x. Atribuir un compromiso normativo a un agente equivale a decir que él debería creer, desear, valorar, sentir o hacer algo[10]. Esto vale también para los afectos. Si atribuyo a una persona amor por otra y temor a que algún daño le suceda, deberé también atribuirle un comportamiento que maximice la seguridad de la persona amada, así como la emoción de la desdicha si la amada sufriera algún percance. Si ante la calamidad de la amada el amante no sufriera desdicha alguna, habría que reevaluar las atribuciones afectivas que le estamos haciendo o, en todo caso, habrá que decir que él no se está comportando racionalmente, es decir, que no podemos entender los estados mentales —y en concreto las razones— que gobiernan su comportamiento. Supongamos que la amada sufre una injusticia de parte de una tercera persona y que este hecho enfurece al amado. Tendríamos que admitir que su ira es perfectamente razonable, es decir, que es comprensible en términos de razones o, lo que es lo mismo, que —ante nuestros ojos— sus sentimientos de amor causan y justifican la emoción de la ira. Lo incomprensible e irracional sería que el amado observara impávido el sufrimiento injusto de su amada. En ese caso tendríamos que optar entre varias opciones: ¿realmente la ama como dice? ¿Se ha dado cuenta de lo que está ocurriendo? ¿Sufre algún desorden afectivo? ¿Ha quedado paralizado por la sorpresa? ¿La ama, pero también tiene fuertes sentimientos negativos por ella que motivan su extraño comportamiento? Como se verá, optaremos por la hipótesis que sea más coherente con el resto de atribuciones que le hagamos y que, por tanto, lo muestre más racional frente a nuestros ojos. En tanto la historia siga fluyendo, haremos —y deberemos hacerlo si somos buenos intérpretes— pequeños y progresivos ajustes en nuestras atribuciones para maximizar, en lo posible, la inteligibilidad, coherencia y racionalidad del agente. Volveremos con más detalle sobre el concepto de racionalidad en la cuarta parte; hasta aquí será importante notar que ser un agente interpretable es ser una criatura racional —a ojos de la intérprete— y que ser racional es tener ciertas obligaciones normativas. Al mismo tiempo, solo puede ser intérprete una criatura racional que está en capacidad de atribuir exigencias normativas y que ella misma tiene exigencias normativas para poder interpretar al agente. Sí, por ejemplo, ella piensa que el agente cree que p y que también cree que p implica q, ella deberá atribuirle a él la exigencia normativa de creer que q,

[10] Esta idea es resumida por John McDowell de la siguiente manera: «Las actitudes proposicionales tienen su lugar propio en explicaciones en las cuales las cosas son inteligibles al ser reveladas como siendo, o aproximándose a ser, como racionalmente deberían ser. Debe contrastarse esto con un estilo de explicación en el que las cosas son inteligibles al representarlas como una instancia particular de cómo las cosas tienden a ocurrir» (1985, p. 389).

aunque no necesariamente la creencia de q. La exigencia normativa es doble: es para el agente, pues él debería creer que q, y es para la intérprete, pues ella debería atribuirle a él la exigencia normativa de creer que q.

Un quinto rasgo de la interpretación es que esta es una descripción teleológica, porque damos sentido al comportamiento intencional de un agente en relación con lo que consideramos son sus propósitos y objetivos. Esto no implica que cada estado mental deba ser teleológico, pero la intérprete podrá atribuir un estado mental solo como parte de un sistema que tenga componentes teleológicos. En otras palabras, para que la intérprete pueda comprender los estados mentales y acciones del agente debe asumir que su comportamiento general está dirigido por propósitos. Con frecuencia esto requerirá adscribirle también atribuciones respecto de lo que él considera valioso y relevante. El primer rasgo —holismo— está presente tanto en la descripción intencional como en la física. Los otros tres rasgos —racionalidad, normatividad y teleología— están ausentes en la descripción física, pero son esenciales para la intencional.

Algunos autores como Charles Taylor (1985), piensan que la distinción entre explicación e interpretación refleja una distinción ontológica más profunda entre objetos físicos independientes de la mente —como una galaxia o una neurona— y objetos culturales dependientes de la mente —como las clases sociales, los adverbios o los tabúes—. Otros autores, como Thomas Kuhn (1991), sostienen que la interpretación puede ser un método apropiado incluso para el caso de las ciencias naturales. Hay también quienes piensan, como los eliminativistas y fisicalistas reduccionistas, que no existe una distinción metodológica u ontológica relevante entre las ciencias naturales y las humanas, pues todo lo que puede ser descrito en términos psicológicos o culturales también podría ser descrito y explicado apropiadamente en términos físicos (Churchland, 1979a, 1979b). Finalmente se encuentran aquellos que piensan que la distinción misma entre explicar e interpretar está comprometida con una distinción ontológica que ellos desean evitar (Rorty, 1980, 1991a). De lo que he estado diciendo se sigue que el punto de vista que defiendo es que la distinción entre explicación e interpretación es puramente metodológica y no está asociada con ningún tipo de dualismo ontológico, lo que es consistente con el monismo de aspecto dual que asumo y defenderé más adelante.

Un importante rasgo de la interpretación del comportamiento intencional de un agente es que requiere de la aplicación del principio davidsoniano de caridad, mientras que la explicación al interior de una descripción física no lo permite. La idea central de este principio —que discutiré detalladamente en la tercera parte— es que una condición necesaria y *a priori* para la interpretación de un agente intencional es asumir que él es básicamente racional, un creyente de verdades y una persona

que actúa de acuerdo con lo que él considera que es su mejor opción disponible, todo según los criterios que la intérprete le atribuye. En este sentido, la interpretación es un tipo de descripción del comportamiento y de los estados mentales del agente en términos de los estados mentales de la intérprete, tomando como marco de referencia el mundo objetivo que ambos comparten. Este es el proceso de triangulación descrito por Davidson[11], que está formulado de manera técnica mediante la producción de oraciones-T, como se verá en el capítulo cinco.

Así pues, podríamos ver el proceso de interpretación de una persona de esta manera. Interpretamos al agente para comprenderlo, es decir, para encontrarlo inteligible ante nuestros propios ojos y frente a nuestra propia vida mental. ¿Pero qué es exactamente comprenderlo? Es encontrar conexiones —lógicas y causales— entre sus diferentes estados mentales y los nuestros, lo que es propiamente compartir parte de su subjetividad, es decir, ver por un momento las cosas y verse a uno mismo como creemos que él las está viendo y se está viendo. También es ser capaz de imaginar algo del significado y de la valoración que el agente confiere a sus propias acciones y a los acontecimientos del mundo, incluidos nosotros en este. En algunos casos podemos incluso llegar a experimentar algo de lo que creemos son sus afectos. De esa manera, comenzamos el proceso de interpretación, al proyectar al agente parte de nuestro propio sistema de estados mentales interconectados. También proyectamos a las acciones y proferencias del agente los significados y valoraciones que conferimos a nuestras propias acciones y proferencias.

Comenzamos la interpretación asumiendo que el agente tiene las mismas creencias y deseos que nosotros creemos que tendríamos si fuéramos él y estuviéramos en las circunstancias en que creemos que él está; que el agente actuaría básicamente de la manera como nosotros lo haríamos si fuéramos él; que el agente atribuiría, en líneas generales, los mismos significados que nosotros atribuiríamos a las palabras ante los mismos contextos, es decir, que él y nosotros usaríamos las expresiones de una manera parecida ante semejantes circunstancias; que él asignaría el mismo valor a las cosas y situaciones que nosotros daríamos, estando en la misma posición, y que él experimentaría los mismos afectos, o análogos, que nosotros experimentaríamos si estuviéramos pasando por lo que creemos él está pasando. Sin embargo, después de hacer estas atribuciones, seguramente reconoceremos que parte de su comportamiento no es coherente con nuestras expectativas originales o, lo que es lo mismo, que nuestras atribuciones no nos permiten predecir o hacer inteligible su comportamiento. Por ello deberemos modificar nuestras atribuciones para dar

[11] Esta idea aparece en muchos de sus textos, pero puede verse especialmente en Davidson, 2001b. Para discusiones recientes sobre este concepto véase Myers y Verheggen, 2016; y Verheggen, 2017.

sentido a su nueva e inesperada conducta. Para poder producir mejores atribuciones, intentaremos identificar la situación en que se encuentra y trataremos de imaginar los estados mentales que tendríamos si fuéramos él y estuviéramos en su situación. Como resultará claro, la capacidad de imaginar escenarios alternativos y estados mentales ajenos, en condiciones contrafácticas, es central en todo este proceso, tema sobre el que volveremos en el próximo capítulo.

Lo que deseo resaltar, entonces, es que encontrar al agente inteligible es trazar conexiones entre los estados mentales que le atribuimos y los nuestros. Cuando interpretamos sus acciones, percibimos por qué él actuó como lo hizo, dados los estados mentales que asumimos él tiene y considerando lo que creemos que es racional hacer, creer, desear, valorar o sentir para él y en las circunstancias en que creemos él está. Al involucrarnos en el proceso interpretativo descubriremos que muchos estados mentales atribuidos al agente no nos permiten hacer inteligible su comportamiento y que, por tanto, deberemos reorganizar las atribuciones para hacerlas consistentes con la nueva evidencia. Como la interacción es un proceso dinámico que va en ambos sentidos, la comunicación entre interlocutores requerirá de la producción de mejores atribuciones —de cada uno respecto del otro— para dar sentido a nuevas formas de comportamiento. La calidad de las atribuciones se medirá por su capacidad para dar más coherencia a mayor evidencia procedente de su comportamiento, lo que nos permitirá compartir algo más de la subjetividad de nuestro interlocutor.

Las ideas que sugiero en este capítulo, y que espero desarrollar más ampliamente en los que sigue, están inspiradas en las investigaciones recientes de la filosofía de la mente y del lenguaje, así como de la teoría de la acción, especialmente en la tradición anglosajona en la que el trabajo de Davidson ha sido fundamental. Sin embargo, más allá de hacer exégesis davidsoniana o de defender a este autor de los cuestionamientos que se le han formulado, mi interés es explorar con detalle el fenómeno intersubjetivo que se constituye entre la intérprete y el agente durante la situación comunicativa en que ambos se atribuyen mutuamente estados mentales para comprenderse o para creer que se están comprendiendo frente a la realidad que ambos comparten. Es claro, sin embargo, que algunas de estas ideas han aparecido también en formulaciones diferentes en otras tradiciones, especialmente en la hermenéutica alemana y en sus desarrollos posteriores. Un autor clásico que es un referente obligado en las ciencias sociales y cuyas intuiciones tienen interesantes semejanzas con los temas que estamos abordando es Max Weber. En su libro *Economía y sociedad. Esbozo de sociología comprensiva* (1964), publicado originalmente en 1922, Weber define los conceptos que desea usar para la fundamentación de la sociología como ciencia de una manera parecida a la que hemos planteado aquí aunque, evidentemente, sin entrar en los detalles, porque su interés es construir una teoría de las ciencias sociales

y no una filosofía de la comprensión (véase también Weber, 2013). Por ejemplo, Weber afirma que la acción social es:

> [U]na conducta humana (bien consista en un hacer externo o interno, ya en un omitir o permitir) siempre que el sujeto o los sujetos de la acción *enlacen* a ella un *sentido* subjetivo. La «acción social», por tanto, es una acción en donde el sentido mentado por su sujeto o sujetos está referido a la conducta de *otros*, orientándose por esta en su desarrollo (1964, p. 5)[12].

En otras palabras, piensa que la acción es un tipo de comportamiento mediante el que alguien tiene un propósito y en el que este propósito está dirigido a otros individuos. Considera también que, de manera metodológica, comprender a alguien supone encontrar su racionalidad (1964, p. 7). También que al comprender una acción nos interesa capturar tanto su sentido —propósito, objetivo— como su motivación (1964, p. 8). En este último caso, las motivaciones son lo que nosotros llamaríamos los estados mentales que causaron la acción.

Un tema sobre el que aún he dicho poco y cuya importancia y complejidad nos obligará a retornar a él es la pregunta sobre qué califica como una interpretación correcta y con qué criterios podemos saber cuándo estamos frente a una. Sobre este punto dice Weber que «Una *interpretación* causal *correcta* de una acción concreta significa: que el desarrollo externo y el motivo han sido conocidos de un modo certero y al mismo tiempo *comprendidos* con sentido en su conexión» (1964, p. 11).

Al intentar aclarar cómo puede determinarse semejante conocimiento «certero», Weber alude a la evidencia empírica, las estadísticas y la probabilidad, sin dar mayor precisión, a pesar de que esta es requerida.

Pero es fundamental saber qué criterios valorativos empleamos y debemos emplear para determinar cuándo una interpretación es mejor que otra. Sostengo que no existe tal cosa como una interpretación correcta y definitiva, pero que sí existen criterios para determinar cuándo una interpretación es preferible a otra porque permite dar una mayor inteligibilidad. Pienso que los criterios formales que nos permiten determinar que una interpretación es preferible son los mismos que emplearíamos para determinar que una teoría científica —tanto en las ciencias naturales como humanas— es más explicativa que otra. Estos criterios son llamados «valores» o «virtudes» epistémicas y podemos agruparlas en cuatro clases, teniendo siempre en cuenta que se trata de criterios que nos permiten valorar las interpretaciones para elegir entre ellas, si se encuentran en disputa.

[12] En esta y todas las citas de Weber las cursivas son del autor y las traducciones corresponden a las ediciones en castellano citadas en la bibliografía.

En primer lugar, están las que tienen que ver con la predicción exitosa y sistemática del comportamiento del agente, que preservan o mejoran el éxito observacional de otras interpretaciones, y producen nuevas y originales hipótesis predictivas. Considérese, por ejemplo, la posibilidad de predecir —en líneas generales—, el comportamiento de un paciente psicológico, sus capacidades de reacción y recuperación, etcétera. O la posibilidad de predecir los nuevos estados mentales y acciones de una persona a quien tratamos de entender. Otro ejemplo sería afirmar que, si nuestras hipótesis sociológicas son correctas, ciertos acontecimientos sociales tendrían que ocurrir. Finalmente, también podríamos sostener que, si ciertas hipótesis históricas son acertadas, tendríamos que encontrar ciertos textos o restos arqueológicos en determinada zona y con cierta antigüedad.

En segundo lugar, tenemos el apoyo interteórico, es decir, la posibilidad de integrar una interpretación con otras interpretaciones previamente aceptadas y que ya han demostrado ser exitosas para comprender a un agente en particular. Por ejemplo, si de dos interpretaciones en disputa una es más consistente que la otra con interpretaciones previas que han resultado exitosas, hay buenas razones para preferir la primera interpretación a la segunda.

En tercer lugar, tenemos a la compatibilidad con creencias justificadas del entorno. Por ejemplo, si una interpretación es consistente con evidencia observacional que asumimos compartir con el agente o con información que nosotros creemos que él debería tener, sería preferible a otra interpretación que no tenga esas características. Imaginemos a Cristóbal Colón llegando a América e intentando interpretar a un grupo de nativos. Él preferirá una interpretación que no asuma información que sería muy extraño que ellos tengan —como, por ejemplo, que él es un enviado de los Reyes Católicos— y que maximice la evidencia que ellos sí podrían tener, como, por ejemplo, que los recién llegados pueden resultar un peligro.

En cuarto lugar, figuran la consistencia interna, la simplicidad ontológica y la economía explicativa. En otras palabras, siempre es preferible una interpretación coherente a una plagada de contradicciones, una que no tenga que postular toda suerte de entidades misteriosas e innecesarias y, finalmente, una que explique más evidencia o comportamiento con menos atribuciones innecesarias.

En resumen, siempre nos manejamos con una diversidad de interpretaciones frente a la misma evidencia observacional y nos enfrentamos a la necesidad de elegir entre ellas. En líneas generales, actuamos de manera semejante a como lo haría un científico cuando tiene que elegir entre varias teorías en disputa y, por tanto, realiza lo que se llama una «inferencia a la mejor explicación», que se define de esta manera: dado un conjunto de eventos sorprendentes, estamos justificados en creer la teoría que los explique de la mejor manera posible, incluso si es solo de manera provisional.

Pero obviamente eso nos conduce a preguntarnos qué criterios debemos emplear para determinar cuál es la teoría que los explica mejor.

Aunque varias teorías podrían ser simultáneamente correctas al iluminar distintos aspectos de los mismos acontecimientos, con frecuencia son incompatibles entre sí, por lo que necesitamos criterios para determinar cuándo una interpretación es preferible a otra. Esos criterios son las virtudes o valores epistémicos que acabo de describir y que no son muy distintos de los que emplearíamos en otros tipos de ciencias, tanto naturales como humanas (Quintanilla, 2003a).

Ahora bien, estas virtudes epistémicas permiten evaluar mejores y peores explicaciones científicas e interpretaciones del comportamiento de agentes intencionales, pero a veces podrían no ser de mucha ayuda para la autocomprensión y, aunque este libro está dedicado a la comprensión del otro, es decir a la aleocomprensión, será necesario terminar este primer capítulo abordando algunas ideas generales acerca de la comprensión de uno mismo.

Es a partir del modelo triangular inspirado en Davidson (2001e [1991]) que me propongo desarrollar que el conocimiento del mundo exterior, el de uno mismo y el de las otras mentes se requieren mutuamente. Esto significa que solo es posible conocer la realidad objetiva si comparamos nuestras creencias y percepciones con las de otras personas. Análogamente, no es posible atribuir estados mentales a otras personas sin poder reconocer también los de uno mismo, con lo que se relacionan ambos tipos de estados mentales —propios y ajenos— con el mundo compartido. Tampoco es posible la introspección si no tenemos la capacidad de atribuir estados mentales a los demás, en relación con el mismo mundo que cohabitamos. Esto es central en la idea de que la subjetividad se constituye de manera intersubjetiva en relación con el mundo objetivo, lo que no supone, sin embargo, que la autocomprensión tenga exactamente las mismas características que la aleocomprensión, aunque veremos que ambas formas de conocimiento no son muy diferentes. Deberemos preguntarnos ahora, por tanto, cómo opera la autointerpretación y qué es el autoconocimiento.

La motivación detrás del mandato délfico «conócete a ti mismo» es doble. Por una parte, está la intuición de que nada puede ser conocido si no se conoce uno a sí mismo también. De otro lado, está la sospecha de que nuestra motivación última hacia cualquier forma de conocimiento proviene de nuestra necesidad de autoconocernos. Lo curioso, sin embargo, es que desde las primeras reflexiones de los griegos hacia el siglo VI a.C. hasta fines del siglo XX, es muy poco lo que la filosofía ha dicho acerca de qué es autoconocerse. Sin duda se ha escrito abundantemente sobre la naturaleza del entendimiento humano —sobre todo en la modernidad— y desde mediados del siglo XVIII en adelante sobre la aleocomprensión, pero es muy poco lo elaborado

sobre la naturaleza misma de la autocomprensión y su relación con otras formas de comprensión. La razón de ello es que hasta mediados del siglo XIX solía asumirse que nuestras creencias introspectivas sobre nosotros mismos y sobre nuestros propios estados mentales son básicamente incorregibles, de manera que lo problemático es entender cómo es posible el conocimiento del mundo exterior y de las otras personas, pero no cuáles son nuestros propios estados mentales. Ciertamente, nadie asumía que tuviéramos un conocimiento pleno de nosotros mismos, pero sí que ya sabíamos cuál era el mecanismo del autoconocimiento —la introspección autorreflexiva— y que, en general, el conocimiento de nuestros propios estados mentales era correcto. Se pensaba que lo que no tenemos claro era el mecanismo del conocimiento del mundo exterior y el de las otras mentes. El texto paradigmático de esa posición es la primera de las *Meditaciones metafísicas* de Descartes (1968), que comienza con la afirmación de que nada es tan fácil de conocer como el espíritu para sí mismo. Precisamente por ello Descartes toma como certeza indubitable el autoconocimiento para luego intentar demostrar la existencia del mundo externo, a través de una previa demostración de la existencia de Dios como garante epistemológico de toda certeza. Es recién hacia mediados o fines del siglo XIX que el autoconocimiento se vuelve problemático cuando se considera que alguna, o gran parte, de nuestra vida mental es no consciente o inconsciente. Esto comienza a ocurrir con Arthur Schopenhauer, Friedrich Nietzsche, William James y, por supuesto, Sigmund Freud. Solo recientemente la filosofía de la mente se ha planteado la pregunta por el autoconocimiento de manera radical, incorporando información procedente de otras ciencias, sobre todo psicología, neurociencias y ciencias de la evolución (Bilgrami, 2006; O'Brian, 2007; Carruthers, 2011; Gertler, 2011; Cassam, 2014; Green, 2017; Renz, 2017).

Aunque estas tres formas de conocimiento se requieren entre sí, tienen características y metodologías algo diferentes, por lo que será necesario compararlas para entender un poco mejor cómo procede el autoconocimiento. La definición tradicional de conocimiento objetivo es esta:

Un sujeto A conoce la proposición p si y solo si se cumplen por lo menos tres condiciones:

- A cree que p.
- A está justificado en creer que p.
- P es una proposición verdadera[13].

[13] Como consecuencia de un célebre artículo de Gettier, 1963, la mayor parte de autores considera que esas tres condiciones no son suficientes y que se necesita una cláusula adicional, pero este es un tema que no discutiré aquí.

Como es claro, esta definición es válida solo para el conocimiento proposicional y gran parte de nuestro conocimiento del mundo externo es tácito o práctico (Polanyi, 1966). Asimismo, nuestro conocimiento de las otras mentes es solo parcialmente proposicional, pues también existe un conocimiento tácito o práctico de la vida subjetiva ajena que es no consciente, porque es consecuencia de mecanismos que no generan en nosotros experiencias fenoménicas ni metacognición. Sin embargo, incluso en muchos casos de las formas de aleocomprensión tácita no consciente es posible reconstruir racionalmente los mecanismos de atribución psicológica para que adopten una forma proposicional.

En el caso del autoconocimiento, con frecuencia nos autoatribuimos conscientemente estados mentales con contenido proposicional, como, por ejemplo, si trato de entender por qué actué de manera tan peculiar ante determinada situación o por qué tengo las creencias, deseos y afectos que tengo. Así es como nos autointerpretamos. Pero con mucha más frecuencia nuestro autoconocimiento es solo tácito y no está acompañado de autoatribuciones con contenido proposicional, aunque en muchos de esos casos podamos hacer una reconstrucción racional proposicional.

Tanto en el caso del aleoconocimiento como en el del autoconocimiento —y solo en esos dos casos— tratamos de capturar el elemento subjetivo contenido en un punto de vista. En el primer caso, el objeto conocido suele cambiar si hay interacción, pero en el segundo caso el proceso y el fenómeno del autoconocimiento modifica inevitablemente el objeto conocido, que es uno mismo. Así en el caso del autoconocimiento hay una dialéctica entre el punto de vista de la primera persona —la introspección— y el de la tercera persona —un punto de vista externo—, de suerte que para conocernos no solo tenemos que hacer un esfuerzo de autoconciencia reflexiva sino, además, debemos tratar de vernos a nosotros mismos como si fuéramos otra persona, con el propósito de ganar cierta objetividad. Pero no solo debo verme a mí mismo como podría ver a otra persona sino también como parte de un contexto mayor. Por ejemplo, ciertamente ganaré en autoconocimiento si sé que soy el producto de la evolución de varias especies de primates y que desciendo de una de ellas, que salió de África hace aproximadamente cien mil años.

Hay una asimetría básica entre el autoconocimiento y el aleoconocimiento. Uno conoce la experiencia fenoménica de sus propios estados mentales de manera incorregible, aunque no la ajena. Por ejemplo, nadie sabe mejor que yo cómo me duele una muela y nunca sabré cómo le duele la muela a otra persona, aunque me describa meticulosamente su sensación. Por otra parte, hay una diferencia fundamental entre el autoengaño y el engaño a otros. En el primer caso, aunque hay intencionalidad, suele tratarse de un fenómeno de división de la mente, como veremos en el capítulo doce. En el segundo caso, aunque también puede uno engañar a otros autoengañándose,

en general el engaño suele tratarse de un caso deliberado de deshonestidad. Esto tiene como consecuencia que a veces uno puede conocer a otra persona mejor de lo que se conoce a sí mismo y que otras personas podrían conocernos mejor de lo que nosotros mismos nos conocemos, sobre todo en aspectos puntuales de nuestra vida mental. Por eso, como el objeto del autoconocimiento cambia en el desarrollo mismo de conocerse, el autoconocimiento es siempre un proceso de autodescubrimiento y transformación personal que no tiene fin.

Conocemos la experiencia fenoménica de nuestros estados mentales de manera inmediata e incorregible, como en el caso de un dolor o de una emoción. Nadie podría convencerme de que no me duele la muele o de que no tengo una experiencia afectiva intensa respecto de alguien y nadie puede conocer mejor que yo la experiencia fenoménica de mis estados mentales. Lo que no conocemos inmediatamente, es corregible y alguien podría interpretarlo mejor que uno mismo es el contenido proposicional de nuestros estados mentales. Esto conduce a que con frecuencia deba confiar en la opinión ajena para conocer los contenidos de mis propios estados mentales. También conduce a que, a veces, la mejor manera de conocerse uno mismo es analizándose como si uno fuera otro, adoptando una perspectiva de tercera persona. Por eso es que no hay una línea clara entre la introspección y la extrospección cuando uno reflexiona sobre los contenidos proposicionales de sus estados mentales.

Por ejemplo, puedo reconocer que tengo una experiencia fenoménica peculiar, aunque no sepa si es alegría, tranquilidad, alivio o entusiasmo. Y, en cualquier caso, tampoco sé si lo es porque he encontrado un libro que durante mucho tiempo estuve buscando o por las recientes declaraciones del presidente de la república. También podría ser que se trate de una experiencia que combine varias categorías de emociones porque, finalmente, aunque las emociones básicas tienen una realidad universal que en muchos casos compartimos con otros mamíferos, las categorizaciones de ellas son lingüísticas, culturales y contingentes. Por eso, en muchos casos, la mejor manera de conocer los contenidos de nuestros propios estados mentales no es mediante introspección sino observándonos a nosotros mismos, ya sea en nuestro comportamiento o, mejor aún, en nuestros vínculos afectivos.

En el caso de las creencias, no existe experiencia fenoménica. Es decir, no «sentimos» lo que creemos, como sí sentimos deseo, alegría o dolor. Por ello reconocemos lo que creemos preguntándonos a nosotros mismos qué nos resultaría bien justificado creer. En la mayor parte de situaciones esto es inmediato, como, por ejemplo, si alguien me pregunta si creo que el primer alcalde de Lima fue don Nicolás de Ribera y Laredo, el viejo. Pero si alguien me pregunta si ese mismo conquistador fue andaluz o castellano, quizá deba detenerme a meditar al respecto o, lo que es lo mismo, tendré que preguntarme qué creencia estaría mejor justificada, dadas las otras creencias que tengo.

En algunos casos simplemente no tendré creencias respecto de algo porque no conseguiré encontrar justificación alguna, como, por ejemplo, si alguien me pregunta si creo que el primer alcalde de Sidney fue el señor Charles Windeyer.

La estrategia de averiguar lo que uno cree preguntándose qué resultaría razonable para uno mismo, es decir, qué creencias nos parecerían bien justificadas, ha sido denominado por Cassam «método de la transparencia» (2014, p. 2), para referir a la técnica empleada y discutida por autores como Edgley (1969), Evans (1982), Byrne (2011), Moran (2012), Fernández (2013) y otros. Cassam objetó este método por considerarlo poco realista y porque, según él, genera una brecha —lo que él llama una «disparidad»— entre el *Homo philosophicus* —una criatura idealmente racional inventada por los filósofos— y el *Homo sapiens* real. De hecho, asumir que siempre creemos lo que es racional es poco realista y no toma en consideración que con frecuencia somos criaturas irracionales y creemos lo que «sabemos» que es falso, así como también actuamos en contra de nuestro mejor juicio, como discutiremos en el capítulo doce.

Por eso, la idea del método de la transparencia es que debemos asumir que, en líneas generales, creemos lo que nos resulta bien justificado a la luz de nuestras otras creencias previas, de manera que yo puedo detectar qué es lo que creo respecto de algo preguntándome qué estaría bien justificado creer, según mis propios criterios de justificación y sobre la base de mis creencias previamente aceptadas. Este es un punto de partida desde el cual, con mucha frecuencia, después deberé hacer modificaciones y correcciones para detectar creencias más complejas y situaciones menos predecibles, en una suerte de «equilibrio reflexivo», para usar el célebre concepto desarrollado por Nelson Goodman (1955) y John Rawls, según el cual, a través de un proceso deliberativo de ajustes mutuos, entre principios generales y juicios particulares, se logra un estado de balance o coherencia en un conjunto de creencias. Pienso que puede verse el método de la transparencia como la versión aplicada a la autointerpretación de lo que en la aleointerpretación sería el principio de caridad, que discutiremos ampliamente en este libro.

No es que seamos siempre racionales, sino que nos tratamos como si lo fuéramos (Dennett, 1998b [1987], p. 52) y nos interpretamos a partir de ese supuesto, para luego hacer modificaciones y correcciones. En el caso del autoconocimiento, para averiguar qué cree, siente o desea uno, nos preguntamos qué se seguiría naturalmente de los estados mentales que nos parece que tenemos y partimos de la base de que esos podrían ser nuestros estados mentales, para luego preguntarnos si no podríamos tener otros estados mentales no conscientes o inconscientes que causen o sean causados por otros estados mentales que tampoco conocemos. Al hacer eso actuamos desde la perspectiva de primera persona, mediante la introspección,

y desde la de tercera persona, tratando de explicar objetivamente lo que nos ocurre como si fuéramos un mecanismo externo. Pero también podemos emplear una perspectiva de segunda persona, tratándonos a nosotros mismos como si fuéramos un tú al que intentamos dar sentido tanto cognitiva como afectivamente.

Pero para abordar, aunque sea de manera somera como pienso hacer aquí, la naturaleza del autoconocimiento, es necesario preguntarse qué es lo que este tipo de conocimiento pretende conocer. En un sentido genérico queremos conocer todo lo que tiene que ver con nosotros y con nuestras vidas, pero, de manera específica, lo que nos interesa es saber cuáles son los estados mentales que tenemos, es decir, aquellos que causan y justifican nuestras acciones y otros estados mentales. También queremos saber cómo y por qué adquirimos esos estados mentales, si están bien justificados y de qué manera participan en nuestra relación con el mundo y con las otras personas pues, en un importante sentido, como señalaron James (1983) y Winnicott (1971), uno es también sus vínculos con las otras personas. En efecto, si el contenido de nuestros estados mentales es relacional, es decir, si está constituido por sus relaciones con otros de nuestros estados mentales y también por las otras personas con las que nos vinculamos, el autoconocimiento no puede ser solo introspectivo sino también exige explorar nuestros vínculos con los demás.

Ahora bien, queremos conocer nuestros estados mentales pero el método de la transparencia solo nos permite conocer creencias. ¿Es que las creencias tienen alguna primacía respecto de los otros estados mentales? En un sentido epistémico lo es de dos maneras. En primer lugar, aunque no todo conocimiento es proposicional, el conocimiento proposicional sí se formula en términos de creencias, que pueden verse como representaciones del mundo y disposiciones para actuar en él. En segundo lugar, casi todos los estados mentales involucran creencias o están parcialmente moldeados por creencias. Los deseos, por ejemplo, suelen incorporar la creencia de que algo es deseable porque es realizable. Y las emociones suelen estar estructuradas cognitivamente sobre la base de creencias: por ejemplo, el miedo puede suponer la creencia de que un daño puede sobrevenirnos y la ira, la creencia de que somos víctimas de una situación injusta. De hecho, el primer autor en explicitar la relación entre emociones y creencias fue Aristóteles, en el segundo libro de la *Retórica* (2002; véase Quintanilla, 2007).

Pero hay otro sentido en que las creencias pueden modificar significativamente nuestros estados mentales, incluso nuestra experiencia fenoménica de ellos. Imaginemos la siguiente situación. Fui invitado a cenar a casa de un amigo, quien me sirve un delicioso plato de carne. Al finalizar la cena le digo que el lomo estuvo maravilloso. Mi amigo, sorprendido, me dice que no fue lomo sino lengua de res almendrada y me pregunta si deseo más. Como nunca había comido lengua y la sola

idea me resulta desagradable, rechazo el pedido. Mi amigo me pregunta: «¿Pero no te ha gustado?». Le respondo: «Ya no». En efecto, me gustó mucho el plato hasta que me enteré de que era lengua. En ese momento el sabor del guiso en mis papilas gustativas cambió. Mi experiencia fenoménica fue otra. Ya no fue una experiencia de placer sino más bien algo parecido al asco. ¿Qué ocasionó que mi experiencia fenoménica cambiara sin que nada más cambie con ella? Solo una creencia, nada más que una creencia.

Aunque los conceptos de autoconocimiento y autocomprensión se asemejan, no son exactamente lo mismo. El primero connota certeza y convicción; el segundo sugiere proceso, falibilidad y transformación. Sin embargo, la diferencia es de matiz, de suerte que para la mayor parte de efectos prácticos autocomprensión y autoconocimiento son intercambiables. Por otra parte, no es lo mismo autocomprensión y autoconciencia. La comprensión alude al proceso de descubrimiento de mis propios estados mentales, vida psíquica e identidad, tanto mediante la introspección como gracias a la observación de mi comportamiento y de mis propios vínculos, y también adoptando una perspectiva de tercera persona respecto de mi propia historia, contexto y entorno. La autoconciencia alude a la habilidad de identificarse a uno mismo y de diferenciarse de otras entidades, es decir, a la capacidad para reconocer que ciertos estados mentales me pertenecen a mí y no a otra persona, y que están integrados conformando un sistema. Puede haber autoconciencia sin autoconocimiento, pero no autoconocimiento sin autoconciencia. Asimismo, la autoconciencia presupone la conciencia —en tanto pura experiencia fenoménica— de manera que podríamos ver tres niveles de complejidad en los que el último presupone los anteriores, pero no viceversa: conciencia, autoconciencia y autocomprensión.

De una manera bastante general, la autocomprensión pasa por el reconocimiento de nuestros propios estados mentales, pero estos pueden ser conscientes, no conscientes e inconscientes. Como vimos los primeros requieren de experiencia fenoménica y, en el caso de la autoconciencia, de por lo menos dos niveles de metacognición. Los segundos son procesos cognitivos que tienen un rol causal en nuestras vidas, aunque nosotros no los conozcamos, pero que pueden llegar a ser conscientes. Los últimos son estados mentales que han sido reprimidos en un sentido psicodinámico.

El autoconocimiento nos permite distinguir entre los estados mentales que tenemos y aquellos que tenemos razones para tener o que deberíamos tener, lo que involucra un elemento normativo. Por ejemplo, si creo que he realizado una acción que yo mismo considero incorrecta, creeré que debo sentirme culpable. No somos observadores pasivos de nuestra vida psíquica, sino agentes de su formación mediante el proceso de su conocimiento. Eso es precisamente lo que nos convierte en agentes

y por eso el autoconocimiento se concentra más en lo que hacemos que en lo que nos ocurre. Parte de lo que hacemos es transformar nuestra propia vida mental en tanto la conocemos.

Lo anterior conduce a un punto importante. Algunos de nuestros estados mentales son propiedades monádicas, como, por ejemplo, un dolor de muelas, que es pura experiencia fenoménica, pero otros son propiedades relacionales, como aquellos que tienen contenido proposicional atribuido por una intérprete en relación con un objeto que ella y nosotros compartimos. En general, los estados mentales dotados de intencionalidad, es decir aquellos que representan hechos o situaciones diferentes de ellos mismos, son relacionales. Ahora bien, en el caso de los estados mentales que son relacionales no es posible conocerlos solo por pura introspección, porque parte de lo que ellos son, esto es, parte de su contenido, depende del tipo de relación que tienen con otros objetos del mundo, ya sea otras personas o hechos de la realidad. Esto prueba que la pura introspección no basta para el autoconocimiento. Más adelante volveré con detalle sobre la ontología de las propiedades.

Pero es claro que conocer los estados mentales propios no es suficiente para el autoconocimiento. Es también necesario conocer las causas de nuestros estados mentales, como nuestra historia individual y grupal, así como el entorno físico y social. Las causas de nuestros estados mentales pueden ser otros estados mentales —propios o ajenos—, así como eventos físicos y sociales del entorno. Averiguar, no solo de manera introspectiva sino empírica, los procesos por los cuales llegamos a determinar nuestros propios estados mentales es también una forma de autoconocimiento, pero en un grado de abstracción mayor. Asimismo, saber cómo se constituyeron estos procesos, elaborar una genealogía del reconocimiento de nuestros propios estados mentales y poder establecer las conexiones que tienen con otras formas de conocimiento, es, sin duda, una forma de autoconocimiento. El tipo de conocimiento que tenemos a partir de la descripción de los procesos psíquicos que están presentes cuando tenemos procesos cognitivos acerca de nuestros propios procesos cognitivos es una forma de autoconocimiento de un nivel adicional, aquel según el cual describimos con cierta pretensión de objetividad lo que ocurre cuando tenemos momentos de introspección subjetiva. Este proceso, a su vez, está conectado y es difícilmente separable de aquel por el cual conocemos los estados mentales de otras personas, así como conocemos los hechos del entorno físico y social. Así pues, el autoconocimiento no es solo un acto de introspección en el que uno se observa a sí mismo sino también un acto de observación de algo que es, al mismo tiempo, exterior e interior a uno: sus vínculos, sus relaciones con los demás, sus reacciones ante los acontecimientos del mundo. Eso implica observarse a uno mismo en relación con los otros, es decir, observar las emociones que los otros nos generan y las que nosotros

generamos en ellos. Dado que la mente se constituye en relación con los demás, conocerse implica comprender la manera en que uno se relaciona con los otros, pues, como señaló James (1983), nuestros vínculos con los demás, así como ellos mismos, son parte de lo que somos.

Es importante distinguir entre los aspectos cognitivos y afectivos de nuestros estados mentales. Los primeros involucran procesamiento o almacenamiento de información, por ejemplo, de información social, en el caso de la cognición social. Los segundos involucran experiencia fenoménica que «afecta» al ser humano en su totalidad, y hacen que valore y «coloree» la realidad, de una manera u otra.

Esto es importante porque la vida afectiva constituye una forma privilegiada de autoconocimiento y autocomprensión. Uno se reconoce en sus vínculos y en sus emociones, así como en las emociones que le generan sus vínculos. De igual manera, las emociones son valoraciones de la realidad o de nuestras acciones que tienen como función conectarnos con nuestros objetivos y propósitos, para evaluarlos y motivarnos a proseguirlos o, por el contrario, para modificarlos. Así como las percepciones sensoriales tienen como objetivo permitirnos sobrevivir exitosamente en un entorno agreste, al proporcionarnos conocimiento del mundo exterior, las emociones nos proporcionan conocimiento de nuestro mundo interior: qué queremos, tememos, deseamos, logramos o valoramos. Las emociones nos permiten dar significado y valor a los acontecimientos de nuestras vidas, y generan en nosotros sutiles reacciones y experiencias internas que nos permiten tomar decisiones apropiadas en situaciones complejas. Las emociones son, por tanto, un tipo de autopercepción que permite al individuo tener una gama de datos sobre él mismo, análoga a la manera en que los sentidos le proveen de información acerca del mundo externo. Aunque ambos tipos de información son fundamentales para guiar la acción, las emociones tienen la función de comunicarnos con nosotros mismos para dirigir nuestra atención a eventos relacionados con nuestros propósitos y objetivos, y a nuestra percepción de si estos son exitosos o frustrados. Pero las emociones no son moralmente neutras, las sociedades suelen valorarlas positiva o negativamente, con lo cual las prácticas sociales las moldean y constituyen e incluso les dan connotaciones normativas. Esto también puede tener una función adaptativa: las emociones que facilitan las relaciones humanas y la cooperación social tienen una connotación moral positiva —y suelen llamarse emociones prosociales—, mientras que las que la dificultan la tienen negativa. Sin embargo, aunque haya un componente adaptativo previo —probablemente universal— para la conformación de las emociones, estas solo se van a cristalizar en una forma de vida que incluso podrá llegar a moldear el componente adaptativo. Así pues, las emociones son una forma de autoconocimiento y el cultivo de nuestras emociones puede ser una forma más profunda de autoconocimiento.

Hay un último punto que abordaré en este capítulo someramente, porque volveremos sobre él. ¿Cuál es el estatuto ontológico de los estados mentales? ¿En qué sentido y de qué manera existen? Hay dos posiciones opuestas y extremas, y entrambas una diversidad de posibilidades. De un lado está el interpretacionismo antirrealista. De otro lado figura un realismo de estados mentales internos. El interpretacionismo, en general (Child, 1994, p. 8), sostiene que podemos conocer la naturaleza de lo mental mediante el análisis del proceso de interpretar a alguien y que al adscribir estados mentales a un agente debemos intentar maximizar el acuerdo entre los estados mentales de ese agente y los que sería razonable que tuviese, desde nuestro punto de vista, en las circunstancias en las que está. Esta posición sería abrazada por Daniel Dennett (1998b [1987]) y por Donald Davidson. Pero el interpretacionismo antirrealista, que no aceptarían Dennett (1998b [1987], p. 49) ni Davidson, afirmaría que a partir de lo anterior se sigue que no existen los estados mentales *per se*, sino solo en tanto son atribuidos a alguien. El realismo de estados internos, por otra parte, sostendría que los estados mentales son entidades internas a la mente o al cerebro, ya sea como representaciones encriptadas en procesos neuronales o de cualquier otra forma. Lo característico de esta posición es sostener que los estados mentales son propiedades monádicas de alguien.

En este libro sostendré que la experiencia fenoménica de un estado mental —lo que uno siente cuando tiene ese estado mental, es decir, la conciencia de ello— es una propiedad monádica subjetiva de una criatura, la cual ha emergido del sistema complejo que es su cerebro. Pero el contenido de un estado mental —aquello acerca de lo que ese estado mental es, por ejemplo, una representación del mundo— es una propiedad relacional triádica que se constituye en una relación interpretativa que incluye tres elementos: el agente, la intérprete y el mundo objetivo que ambos comparten. Esta es una posición interpretacionista y también realista, porque sostiene que esas propiedades relacionales son objetivas y no dependen únicamente del agente ni de la intérprete.

Como es claro, el contenido variará según las características de los tres elementos de la situación comunicativa. Así, por ejemplo, mi cerebro está actualmente en determinados estados conformados por configuraciones neuronales. Si una intérprete desea dar sentido a mi comportamiento en relación con sus propios estados mentales y al mundo, me atribuirá ciertas creencias, deseos y afectos. Esas atribuciones son maneras de capturar una realidad que no es únicamente cerebral sino también social. Al intentar hacerlo, la intérprete asignará contenidos a mis estados mentales. Eventualmente otras intérpretes podrán intentar hacer lo mismo, con lo cual me atribuirán diferentes estados mentales que darán lugar a distintos contenidos. Aquí se generará cierta indeterminación de la interpretación, pero habrá

criterios epistémicos que permitirán determinar con cierto grado de precisión cuál interpretación es preferible por ser más explicativa. El proceso neuronal es real y, en principio, puede observarse desde una perspectiva de tercera persona. La experiencia fenoménica también es real, pero solo puede ser experimentada desde la perspectiva de primera persona. El contenido del estado mental es igualmente real y se constituye en la relación triangular entre intérprete, agente y realidad, y requiere una perspectiva de segunda persona.

Se podría objetar que de aquí se sigue que no hay un contenido sino muchos, dado que podría haber muchas intérpretes, lo que trivializaría el concepto mismo de contenido. Pienso que este fenómeno debe verse en analogía con el significado. Aunque distintos hablantes usen una palabra de diferentes maneras, atribuyéndole, por tanto, distintos significados, el significado de la palabra será el producto de la intersección sistemática de muchas atribuciones semánticas exitosas. En ese sentido, los significados son reales y socialmente constituidos. De la misma manera, distintas intérpretes pueden atribuir a un agente diferentes estados mentales, aunque podría haber cierto acuerdo en que él cree que p y desea que d, incluso si hay cierto margen de indeterminación acerca del contenido de p y d, ya que estos se conforman en relación a los otros estados mentales del agente y en relación con la intérprete. Negar existencia a los contenidos porque se constituyen en situaciones intersubjetivas sociales sería tan absurdo como decir que no existen los adverbios, las clases sociales o las relaciones afectivas de parejas, solo porque estas siempre pueden verse de muchas y distintas maneras.

Los colores también son propiedades relacionales triádicas que requieren de la luz, un objeto que refracte la luz y un sistema retiniano y cerebral en buen estado; pero sería absurdo decir que no existe el color azul, aunque la experiencia fenoménica que tú tienes del azul del mar en cierta circunstancia y hora del día sea una propiedad subjetiva tuya.

Yo nunca sabré si tú ves cierto tono del azul del mar de la misma manera como yo lo veo, porque nuestros diferentes sistemas de percepción visual podrían procesar la luz de maneras ligeramente diferentes. Por eso la experiencia fenoménica del azul es subjetiva. Pero lo importante es que yo sé que tú ves el azul más o menos como yo lo veo y que cuando digo que «El mar es azul», tanto tú como yo sabemos que estamos hablando del mismo objeto con las mismas propiedades y sabemos que cada uno de nosotros sabe que estamos hablando del mismo objeto con las mismas propiedades. Ambos sabemos también que el azul del mar es una realidad objetiva, en tanto no depende de la subjetividad individual de ninguno de nosotros, y que cualquier ser humano que tenga órganos sensoriales normales y sistemas neuronales aptos para procesar la luz tendrá, en este mar y a esta hora del día, algún tono de azul

y no uno de rojo ni de amarillo. También sabemos que si lo viera rojo o amarillo tendríamos que llevarlo a un oftalmólogo. Más aún, sabemos que es lógicamente posible, aunque sumamente improbable, que él lo vea rojo, aunque lo llame azul y que, por tanto, tenga una experiencia subjetiva muy diferente de la nuestra. Pero incluso, en ese caso, seguiríamos hablando de «azul» y seguiríamos diciendo que el mar es azul.

Análogamente podemos suponer que distintas intérpretes que pertenezcan a parecidas comunidades epistémicas deberían atribuir estados mentales semejantes a un agente en circunstancias comunicativas afines. Sabemos que, si veo a alguien dirigirse a una refrigeradora y sacar una jarra de limonada, en condiciones normales le atribuiré el deseo de beber un líquido para calmar la sed y la creencia de que la limonada puede cumplir esa función. También describiré ese evento físico como una acción intencional. Solo en condiciones muy especiales, y si hay mucha más evidencia que lo amerite, le atribuiré el deseo de envenenar la limonada y la creencia de que de esa manera podrá deshacerse de un vecino incómodo. En general, la mayor parte de intérpretes racionales hará atribuciones parecidas en circunstancias semejantes. Esto es análogo a la atribución de significados a las expresiones lingüísticas, como veremos en la segunda parte.

En el sentido expuesto, el azul del mar y los estados mentales de las personas son propiedades objetivas de la realidad, aunque el azul no pueda existir sin la luz ni criaturas con un determinado tipo de capacidad sensorial, y los estados mentales no puedan existir sin situaciones comunicativas y criaturas racionales que tengan ciertas habilidades de atribución psicológica. De esta manera, también, los contenidos de los estados mentales son propiedades relacionales triádicas objetivas de la realidad. Pero en tanto propiedades que son —tanto el azul como un determinado estado mental— existen a partir de la existencia física de otros objetos naturales de los cuales son propiedades, como también lo son los significados de nuestras expresiones. En otras palabras, aunque el azul y los estados mentales son propiedades reales, existen a partir de fenómenos naturales de los cuales precisamente son propiedades. En el caso del azul, la relación entre la luz, el mar y un organismo sano. En el caso de un estado mental, la relación entre una criatura interpretable, una intérprete y un objeto del mundo compartido por ambos.

Me parece improbable, aunque lógicamente posible, que los contenidos de los estados mentales —vistos, por ejemplo, como representaciones cognitivas— estén codificados o encriptados en procesos cerebrales específicos, de manera que la creencia «La limonada es bebible» dé lugar al mismo cableado neuronal en cualquier persona que la tenga a manera de una identidad tipo-tipo. Si así fuera, esas encriptaciones serían la base neurológica de una creencia específica, un deseo, una emoción, etcétera.

Pero desde el punto de vista de la interpretación, que es lo que ahora nos interesa porque nuestro objetivo es aclarar qué es comprender a alguien y cómo ocurre ese fenómeno, el contenido de un estado mental es aquella propiedad relacional que emerge en el proceso comunicativo.

La vida mental es un flujo continuo de experiencias subjetivas que nosotros, para lograr comprenderla mejor, clasificamos, ordenamos y nombramos con conceptos culturales como «creencia», «deseo», «pasión», etcétera. La vida mental no es una construcción discreta conformada por piezas independientes, como si fueran los electrones, protones y neutrones que constituyen la materia. Toda descripción de la subjetividad es, por tanto, en algún grado arbitraria, y es factible que haya lenguas y culturas que la ordenen de distintas maneras. Por ejemplo, las lenguas distinguen de diferentes formas las emociones y podría haber alguna que carezca del concepto de creencia. Sin embargo, es de suponer que todas las sociedades humanas tengan mecanismos para que sus miembros se comprendan unos a otros y que estos mecanismos incluyan estrategias de atribución comunes, ya sea si lo hacen con conceptos semejantes o no. Sería improbable, por ejemplo, que una sociedad carezca de una manera de describir lo que nosotros llamamos «creencias», en tanto actitudes que uno puede tener acerca de la manera en que es el mundo y que pueden ser verdaderas o falsas. Es altamente adaptativo que los seres humanos tengamos creencias y conceptos de creencias. También lo es que podamos atribuirnos creencias mutuamente y que contemos con mecanismos para distinguir entre creencias que describen el mundo como pensamos que es y aquellas que no lo hacen. Es de suponer que todas las lenguas lo hacen, incluso si no tienen una palabra para «creencia» o «verdad» y en vez de ellas usan evidenciales o alguna otra estrategia gramatical.

En esa línea, la recientemente desarrollada «filosofía experimental» se propone explorar evidencia empírica interdisciplinaria que pueda corroborar o falsar tesis tradicionalmente filosóficas (por ejemplo, Devitt, 2004; Goldman, 2007; Kauppinen, 2007; Machery, 2017). Para el caso que acabo de mencionar, esta filosofía podría investigar, por ejemplo, si existen intuiciones o presupuestos universales subyacentes a las diferentes culturas acerca de temas filosóficos fundamentales como el conocimiento, la verdad, la comprensión del otro, la sabiduría, la identidad de uno mismo, la agencia y el libre albedrío, etcétera[14].

Más aún, de la misma manera como cada lengua tiene formas propias de clasificar la realidad según sus necesidades funcionales, las lenguas clasifican de distinta forma la vida mental de sus usuarios. Ser hablante de una lengua, por tanto, favorece

[14] Véase, por ejemplo, el proyecto *The Geography of Philosophy*, que explora transculturalmente y de manera interdisciplinaria la universalidad y diversidad en algunos conceptos filosóficos fundamentales: https://www.geographyofphilosophy.com/

—aunque ciertamente no determina— que clasifiquemos el mundo y nuestra propia vida mental de cierta manera. Esto implica que la comprensión requiere que podamos compartir o imaginar la manera en que los otros clasifican la realidad objetiva y la vida mental subjetiva.

En este primer capítulo he trazado una suerte de mapa general de los temas que están asociados a la pregunta central que nos interesa: ¿qué es comprender al otro? En el próximo capítulo nos concentraremos en los mecanismos psicológicos que permiten que una intérprete dé sentido al comportamiento de un agente mediante la atribución de estados mentales, esto es, el fenómeno de la atribución psicológica.

Capítulo dos
La atribución psicológica

2.1. Psicología folk y la perspectiva de tercera persona

La discusión filosófica sobre la naturaleza de la comprensión y las dificultades de comprender a quienes son diferentes a nosotros se puede rastrear hasta la Antigüedad. Hay frases iluminadoras en aquellos diálogos platónicos que versan sobre la dialéctica, así como en los tratados aristotélicos *De Interpretatione* (2015), *Poética* (1968) y *Retórica* (2002), aunque no haya en el pensamiento griego un análisis propiamente comprehensivo sobre el tema. Con el advenimiento de la modernidad, el *Ensayo sobre el entendimiento humano*, de John Locke (1982), y el *Tratado sobre la naturaleza humana*, de David Hume (1975), plantean la discusión en nuevos términos. Para Hume la comprensión está asociada a la *sympathy*, que es la capacidad de compadecerse de otra persona y requiere de la habilidad para escapar al propio interés para aprehender la perspectiva de otra persona. Por ello, Hume piensa que la *sympathy* es condición de posibilidad y fundamento de la vida moral.

En el caso de la filosofía más reciente que se puede rastrear hasta Friedrich Schleiermacher, Wilhelm Dilthey, Martin Heidegger y Hans-Georg Gadamer —en su vertiente alemana— y el pragmatismo estadounidense, Ludwig Wittgenstein, Willard Van Orman Quine y Donald Davidson —en la tradición angloamericana— hay una fascinación en torno del problema del significado y la interpretación. Es posible presenciar en ambas tradiciones un alejamiento implícito de la hermenéutica intencionalista —aquella según la cual entender al otro es conocer sus reales intenciones— y un acercamiento a la concepción según la cual la comprensión es un fenómeno que se produce en la interacción entre horizontes o formas de vida, como la generación de un territorio común.

Deseo explicitar esa intuición, que me parece se encuentra implícita en esos autores y radicalizarla, para así sostener que comprender al otro no es reconstruir

su vida mental sino crear un espacio compartido, que es una invitación al cambio y a la adaptación. También propongo que tanto el significado como la comprensión emergen en la interacción entre una intérprete y un hablante, y que no son posibles sin el concurso cooperativo de ambos, como lo veremos detalladamente en el capítulo cuatro.

Esos son los temas sobre los que discutiremos largamente en este libro. Sin embargo, en este capítulo me concentraré en un tema más bien técnico y reciente en el que confluye la filosofía con diversas disciplinas empíricas, que es el problema de la atribución psicológica, esto es, el fenómeno por el cual empleamos diversos mecanismos cognitivos y afectivos para adscribir estados mentales a otras personas con la finalidad de comprender, explicar y predecir su comportamiento. La atribución psicológica no equivale a la comprensión, pero es condición de posibilidad de ella.

Como las investigaciones sobre atribución psicológica son relativamente recientes, la terminología es variada, repetitiva y con frecuencia se superpone, es decir, distintos vocablos designan fenómenos parecidos o superpuestos entre sí. En la medida de lo posible precisaré los significados de los términos técnicos empleados, aunque habrá que aceptar que muchos de ellos se emplean en la literatura con cierta imprecisión. Por otra parte, en este terreno la reflexión filosófica requiere de manera indispensable, para poder avanzar, la evidencia que procede de las ciencias empíricas. Más aún, en estos temas las investigaciones en filosofía experimental pueden ser decisivas para dirimir entre posiciones en conflicto (Knobe, Buckwalter, Nichols, Robbins, Sarkissian & Sommers, 2012).

Se entiende la atribución psicológica como el conjunto de habilidades, destrezas, mecanismos y estrategias —tanto innatas como adquiridas— que empleamos para dar sentido al comportamiento intencional de un agente —ya sea otro o uno mismo— mediante la adscripción de estados mentales. La expresión *folk psychology* se usa en las discusiones sobre atribución psicológica para referir a la habilidad de lectura de mentes (*mindreading*), es decir, aquella que nos permite reconocer los estados mentales de los demás con la finalidad de predecir, explicar, comprender y describir su comportamiento y sus otros estados mentales, así como también producir generalizaciones acerca de su comportamiento intencional (Stich & Ravenscroft, 1994). Al usar la expresión «psicología folk» se suele subrayar que se trata de una especie de teoría psicológica implícita que usan las intérpretes de manera innata y no consciente, como si fuera un conjunto de conocimientos o hipótesis psicológicas de uso común que la gente emplea sin saberlo y sin haber tenido un entrenamiento formal para ello. De manera análoga, se suele hablar de una *folk physics* —física popular o física folk—, que estaría conformada por un conjunto de conocimientos sobre el funcionamiento de la naturaleza, que el ser humano normal posee sin haberlo aprendido

formalmente y que incluiría el saber que, por ejemplo, todo efecto tiene una causa, la naturaleza es regular y no aleatoria, no pueden coexistir dos objetos en el mismo espacio físico, etcétera. También es probable que tengamos una biología folk, que nos permita distinguir de manera casi inmediata entre criaturas vivas e inertes, y una geometría folk, que nos conduce a creer que el espacio es plano y que la línea más corta entre dos puntos es la línea recta. Hay quienes piensan que también es parte de nuestra dotación innata presuponer un dualismo en la naturaleza entre objetos físicos y mentales, o suponer que todas las especies y clases naturales tienen una esencia. De ser cierto que tenemos una dotación de teorías implícitas folk, estas serían el producto de la evolución de la especie para resolver problemas puntuales y facilitar la supervivencia, incluso si estas teorías resultaran no ser verdaderas o fuesen incompatibles con las teorías científicas mejor justificadas empíricamente en el presente. Ese sería el caso, por ejemplo, de la geometría euclidiana.

Estas teorías folk suelen llamarse conocimientos folk, aunque en sentido estricto no siempre constituyen conocimiento porque no siempre son verdaderas. Serían formas de conocimiento en un sentido amplio e impreciso, en el sentido en que la gente se conduce como si lo fueran.

Al hablar de teorías o conocimientos folk, se suele distinguir entre una dimensión fenoménica y una arquitectónica. La dimensión fenoménica alude al conjunto de conceptos que tienen las intérpretes y que les permite dar sentido al comportamiento ajeno vía un conjunto de predicciones, descripciones, explicaciones, etcétera, con el empleo, principalmente, de los conceptos de creencia y deseo (Davies & Stone, 1995a). Esta dimensión toma el nombre de fenoménica porque, en líneas generales, tenemos una experiencia intuitiva de ella y, en principio, podemos describir los conceptos que empleamos para interpretar a los demás sin requerir de una investigación científica más profunda. La dimensión arquitectónica, de otro lado, tiene que ver con los mecanismos neurológicos —tanto cognitivos como afectivos— que subyacen a la dimensión fenoménica y la hacen posible. No tenemos experiencia fenoménica de estos mecanismos y para tener conocimiento de ellos es necesario un entrenamiento científico detallado. Es usual sostener que la dimensión arquitectónica depende de la configuración biológica del cerebro, de manera que es universal en la especie y es producto de la selección natural. Como consecuencia de ello, deberíamos encontrar versiones rudimentarias de esta dimensión en otras especies de primates. También tendríamos que encontrar en el desarrollo del niño estadios de maduración más o menos homogéneos, los cuales serían, en gran medida, independientes de las culturas, tal como ocurre con la adquisición y el desarrollo de las habilidades lingüísticas.

Aunque la dimensión fenoménica depende de la arquitectónica, es mucho más sensible a la influencia cultural que esta. Así, en principio podríamos encontrar gran variedad de estrategias de atribución psicológica en las distintas culturas, aunque todas ellas serían variaciones de una estructura básica universal. Ello no sería muy diferente de la manera en que las distintas lenguas son variaciones de una gramática universal, si uno cree que esta existe. Asimismo, la diversidad cultural en un nivel fenoménico podría explicar que en las distintas culturas haya conceptos psicológicos algo o muy diferentes entre sí —por ejemplo, de creencia, deseo, acción, etcétera— lo que podría generar la sensación de que no hay nada en común entre ellas.

Las discusiones sobre atribución psicológica han girado usualmente en torno de dos perspectivas, la de tercera y la de primera personas. Los partidarios de la perspectiva de tercera persona consideran que la interpretación de un agente intencional no es fundamentalmente diferente de la explicación de cualquier otro fenómeno de la naturaleza. Según este modelo, la intérprete atribuye estados mentales al agente para explicar su comportamiento en términos causales y sobre la base de regularidades nomológicas. Esto significa que la intérprete operaría como una científica que aplica el modelo nomológico deductivo para investigar el cosmos. Ella se dedicaría a encontrar relaciones causales entre estados mentales y acciones para luego buscar las regularidades que gobiernan esas relaciones causales. Según este modelo, la intérprete podría tener un módulo innato de atribución psicológica que estaría estructurado como una teoría científica. Así como los seres humanos probablemente tengamos una *folk physics* o una *folk biology* que dependerían de sendos módulos cerebrales, podríamos venir al mundo dotados de un módulo de atribución psicológica que funcionaría como una teoría nomológica, lo que nos convertiría en psicólogos natos. Recientemente Alan Leslie (1994, 1999, 2000) ha propuesto la existencia de un módulo de atribución psicológica para explicar nuestras habilidades naturales para dar sentido al comportamiento ajeno mediante la atribución de estados mentales. Esta adaptación evolutiva habría sido fundamental para la supervivencia, pues nuestros ancestros primates no solo estaban obligados a adaptarse a un peligroso entorno natural sino también a un exigente entorno social que progresivamente se hizo más complejo, lo que les habría exigido desarrollar habilidades de atribución psicológica en términos casuales y en varios niveles de intencionalidad.

La tesis central de Leslie es que el autismo puede ser explicado como un daño en este módulo de atribución psicológica, que puede ser visto como un mecanismo innato de procesamiento de cierto tipo de información social, lo que se evidencia en los autistas como una significativa incapacidad para simular escenarios alternativos —lo que incluye juego simbólico, como, por ejemplo, imaginar que una silla

es un caballo sabiendo que no lo es—, comunicación intencional y competencias sociales. Pero Leslie (1992) va más lejos aún. En experimentos realizados a comienzos de la década de 1990, observó que los autistas tienen un déficit en la comprensión de estados mentales. Es decir, los autistas pueden comprender procesos físicos causales, pero su incapacidad radica en entender situaciones cuya explicación requiere inevitablemente de la atribución de estados mentales y, por tanto, del reconocimiento de agencia, incluso comparados con niños con síndrome de Down con un CI mucho menor que el promedio entre los autistas. Según esos experimentos, la formulación verbal de los autistas incluye más conceptos y descripciones físicas y menos conceptos y descripciones mentales que personas con síndrome de Down. Pero cuando se expone a ambos grupos de personas a historias que pudieran ser entendidas sin recurrir a conceptos mentales, los autistas equiparan a las personas con síndrome de Down o incluso los superan en la prueba. Desde entonces los hallazgos de Leslie han sido corroborados por otros experimentos, sugiriendo que el autismo es un daño en la capacidad metarrepresentacional y de simulación, lo que, a su vez, sugeriría un daño en un módulo de teoría de la mente o atribución psicológica. La conclusión de Leslie, por tanto, es que la única manera de explicar el rápido desarrollo de las capacidades de atribución psicológicas de los niños es postulando la existencia de un módulo de atribución psicológica que, cuando sufre un daño en el desarrollo temprano del cerebro da lugar a síntomas que se encuentran dentro del espectro autista. Al respecto, Leslie y Thaiss sostienen que:

> las nociones de estado mental que se desarrollan durante el período preescolar son de dominio específico y están asociados a la comprensión de agentes (1992, p. 246).
>
> Dado un cerebro normal en su desarrollo, por una diversidad de razones las distintas partes de su arquitectura pueden mostrar desarrollos paralelos. Al encontrar cambios paralelos en la realización de pruebas a lo largo del tiempo, la irresistible conclusión es que tales cambios paralelos son evidencia de un solo mecanismo cognitivo (p. 249).

Todo esto sugeriría que existe un módulo de atribución psicológico. Por ello, en este punto será necesario detenerse brevemente en el concepto de modulo mental. El cerebro está organizado en regiones que son el producto de diversos momentos y estadios de adaptación a diferentes medios y necesidades. El cerebro no responde a una planificación previa sino es, más bien, una superposición de estructuras con objetivos adaptativos diferentes que se han acumulado desordenadamente unas sobre otras, dentro de un cráneo más bien reducido en tamaño y que no puede crecer ilimitadamente, porque generaría otros problemas anatómicos de adaptación.

La mente, por otro lado, está organizada en términos de funciones, no de espacios. Las divisiones funcionales de la mente se llaman módulos. La tesis de la modularidad de la mente fue propuesta originalmente por Jerry Fodor (1986), basado en Noam Chomsky. Según Fodor, los módulos son independientes entre sí —están encapsulados— y solo se comunican entre ellos por un sistema central que integra la información de los diversos módulos para la generación de creencias y acciones. Originalmente Fodor propuso dos módulos: la visión y el lenguaje, aunque posteriormente diferentes autores propusieron la existencia de otros. Un módulo se caracteriza por ser:

- Un conjunto de reglas para resolver problemas específicos.
- Un cuerpo de conocimientos complejos sobre la realización de acciones y comportamientos.
- Innato.
- Hereditario.
- Encapsulado: cada módulo está aislado de los otros.
- Una función con especificidad de dominio: tiene una función puntual.
- Una operación fuera del control voluntario: uno no lo controla.
- Un mecanismo que se caracteriza por su rapidez, comparado con los procesos conscientes.
- Un mecanismo implementado en arquitecturas neuronales específicas, aunque no necesariamente localizado en regiones concretas del cerebro.
- Una función que da lugar a daños y patologías muy concretos.

Más adelante, Peter Carruthers postuló la tesis de la modularidad masiva (2006). Esta sostiene que los módulos son construcciones a partir de módulos más pequeños, los cuales están conformados por módulos aún más pequeños y así sucesivamente. Los módulos no estarían encapsulados, sino que se comunicarían entre sí y no existiría un sistema central. Como para Carruthers no hay sistemas centrales, el lenguaje tiene el rol de integrar a los diversos módulos, lo que origina la conciencia. Este autor defiende su tesis empíricamente, mostrando cómo las actividades que involucran el lenguaje son mucho más lentas que las que no lo hacen. Así, de acuerdo con Carruthers, el lenguaje permea las diversas actividades cognitivas humanas.

Podría haber innumerables módulos encargados de la producción de creencias y acciones, si bien cada uno de ellos estaría restringido a un dominio muy específico como, por ejemplo, las creencias físicas, morales, psicológicas, etcétera. Estos módulos

darían lugar a cuerpos de creencias como la psicología o la física folk. En el caso humano, podría haber módulos de:

- El reconocimiento de rostros.
- El comportamiento social, que gobernaría fenómenos como la cooperación y la compasión.
- El lenguaje. Este módulo incluiría como submódulos la sintaxis y la semántica. Es discutible si podría haber un módulo pragmático.
- El conocimiento geométrico.
- La atribución de estados mentales.
- La detección del engaño.
- El apego en el vínculo temprano.
- El comportamiento moral.

Además de la existencia de un módulo lingüístico (Chomsky, 1965, 1988; Fodor, 1986), Elizabeth Spelke (1988) ha propuesto uno de predicción y explicación del movimiento de objetos espaciotemporales de tamaño mediano; Scott Atran (1998), uno de clasificación de animales y plantas; y Leslie, como vimos, uno de atribución psicológica (Hirshfield & Gelman, 1994; Sripada, 2008). Recientemente también se ha propuesto la existencia de módulos morales con el objetivo de explicar principios fundamentales subyacentes a la diversidad moral cultural humana (Stich, 1993; Mikhail, Sorrentino & Spelke, 1998; Harman, 1999; Dwyer, 1999).

Cada módulo estaría estructurado a la manera de un pequeño procesador computacional y resolvería problemas de manera análoga a como lo hace una computadora. Si todo esto es correcto, la mente humana podría ser vista en analogía al conjunto de las funciones producidas por un grupo de computadoras de alto procesamiento, que trabajan de manera conjunta y mantienen comunicación entre sí.

Volviendo a la perspectiva de tercera persona, esta podría depender de la existencia de un conjunto de estrategias neurológicas que darían lugar a un módulo, pero esta perspectiva no está comprometida necesariamente con una teoría de la modularidad. La perspectiva de tercera persona suele recibir el nombre de *theory of theory* o T-T, porque es la teoría que sostiene que la atribución psicológica es posible gracias a que poseemos una teoría de la mente innata, esto es, a que la comprensión es fundamentalmente el proceso de construcción de teorías de la mente, de manera análoga a como producimos teorías naturales para explicar el universo físico.

Es conveniente decir, sin embargo, que la expresión «teoría de la mente» puede confundir. En primer lugar, sus defensores no suelen distinguir entre explicación y comprensión, como sí deseo hacerlo yo. En segundo lugar, en el sentido técnico usado en estos debates esta expresión significa la habilidad que desarrolla una intérprete para poder explicar el comportamiento de un agente mediante la elaboración de un conjunto de hipótesis sobre los contenidos de su mente, es decir, mediante el reconocimiento de que el agente posee una mente dotada de subjetividad que funciona, en líneas generales, sobre la base de estados mentales que causan acciones. La expresión «teoría de la mente» fue acuñada por los primatólogos David Premack y Guy Woodruff (1978), en un clásico artículo en el que se preguntan si hay primates no homínidos que tengan la capacidad para reconocer estados mentales ajenos, esto es, que se den cuenta de que otras criaturas también poseen estados mentales. El criterio para determinar eso fue la prueba de la falsa creencia, que mide la habilidad que puede tener un infante para reconocer que otro individuo tiene creencias falsas, es decir creencias distintas de las que ella tiene, lo que probaría que ella se da cuenta de que la criatura en cuestión tiene estados mentales propios y diferentes.

Los experimentos de la falsa creencia fueron diseñados con el objetivo de determinar a qué edad los niños son capaces de atribuir a otros individuos creencias diferentes de las que ellos mismos tienen (Perner & Wimmer, 1983). Hay muchas versiones de estas pruebas, pero lo que todas tienen en común se puede reconstruir en esta situación: hay dos personajes en un escenario, A y B. A juega con una canica, la deja en una caja de cartón y sale del escenario. B saca la canica de la caja, sin que A lo vea, y la deja en una canasta de mimbre. Se pregunta a dos grupos de niños, el primero de cuatro años en adelante y el segundo de menos de tres años: Cuando regrese A, ¿dónde buscará la canica? Los niños más grandes responden que en la caja de cartón —donde A dejó la canica—, mientras que los niños menores contestan que en la canasta de mimbre —donde los niños saben que la canica está, aunque A no podría saberlo—. Mientras que los niños menores creen que los demás creen lo que ellos creen, es decir, todavía no tienen la noción de que las otras personas tienen perspectivas diferentes de la misma realidad, los niños mayores son capaces de imaginar que otras personas tienen estados mentales que ellos no tienen.

Las pruebas de la falsa creencia fueron diseñadas para determinar el momento en que una criatura puede distinguir entre sus creencias y las creencias de los demás, o entre su perspectiva y las perspectivas de los demás, y durante mucho tiempo lo usual fue pensar que esto ocurría entre los tres y cuatro años de edad. Lo importante en estas pruebas es que aparentemente miden una habilidad que no está conectada con el coeficiente intelectual, porque niños con síndrome de Down frecuentemente

las superan incluso si tienen un CI menor que el de niños con autismo, que no las superan. También se ha solido asumir que la edad en que los niños pasan estas pruebas es universal y que no tiene relación con el sexo, la lengua o la cultura. Los niños que no las aprueban suelen pertenecer al espectro autista, en sus diversos grados, que son niños con una marcada dificultad para imaginar situaciones contrafácticas. De hecho, los autistas no comprenden metáforas ni oraciones contrafácticas, no tienen juego simbólico —no juegan a que algo es diferente de lo que es ni adoptan roles— y no podrían participar en una obra de teatro, cosa que los niños con síndrome de Down sí pueden hacer. Los autistas tampoco establecen relaciones entre ellos mismos, otras personas y un objeto que ambos comparten, y no suelen seguir la mirada de otro ni señalan con el dedo a un objeto compartido con otra persona. En otras palabras, no triangulan.

En 1978, Premack y Woodruff sostuvieron que los chimpancés sí pasan la prueba de la falsa creencia y que, por tanto, pueden llegar a tener una teoría de la mente ajena. A partir de entonces se han discutido incansablemente varias opciones:

(1) Si las pruebas de la falsa creencia, aplicadas durante las décadas de 1980 y 1990, estaban viciadas al presuponer la capacidad verbal del agente.

(2) Si las pruebas de la falsa creencia realmente miden la presencia de una teoría de la mente o solo de un tipo de teoría de la mente, que sería la propia de un humano suficientemente desarrollado y dotado de lenguaje.

(3) Si hay primates no homínidos con capacidad para reconocer estados mentales en otras criaturas.

(4) Si las estrategias de atribución psicológica son mucho más complejas que la pura atribución mecánica de estados mentales estructurados de manera teórica, lo que podría notarse ya en bebes muy pequeños o en mamíferos superiores.

(5) Si no deberíamos abandonar definitivamente la expresión «teoría de la mente», por estar comprometida con errores conceptuales insalvables, para sustituirla por otras expresiones con menos connotaciones, como «mentalización», «metarrepresentación», «simulación», etcétera.

De ser así, deberíamos abandonar también la idea de que la intérprete es una pequeña científica que construye, de manera no consciente, teorías para comprender a los otros. Los críticos de la T-T sostienen que los mecanismos de atribución psicológica son o incorporan estructuras mucho más básicas de tipo precognitivo y preverbal, que se gestan en situaciones intersubjetivas desde el momento mismo del nacimiento (para una revisión crítica del debate véase Wellman, Cross & Watson, 2001).

En un importante artículo, Josep Call y Michael Tomasello (2008) pasan revista a la evidencia reciente para evaluar las conclusiones de Premack y Woodruff (1978). Su propuesta es que los chimpancés sí son capaces de reconocer los objetivos, las intenciones, las percepciones y el conocimiento de los otros, pero que no son capaces de entender el concepto de creencia falsa y, menos aún, de atribuir falsas creencias a los demás. A partir de la evidencia que han recogido, Call y Tomasello concluyen que los chimpancés comprenden a otros en función de un tipo de psicología basada en la percepción y los objetivos, más que en términos de la psicología típicamente humana basada en creencias y deseos. En general, la tesis de Tomasello (2014, 2016) es que los simios pueden simular condiciones contrafácticas, lo que les permite engañar y contraengañar, e incluso pueden llegar a tener un comportamiento que podría verse como si fuera cooperativo sin serlo, porque actúan siempre bajo su propio interés. Por el contrario, los humanos podemos cooperar socialmente teniendo en cuenta el interés del grupo, lo que permite la conformación de un «nosotros», es decir, de una agencia colectiva. Para Tomasello (2016) este sería el origen filogenético del comportamiento moral. La constitución de un «nosotros» que se amplía como un círculo en expansión ha sido visto por Peter Singer (1981) como un modelo para entender el desarrollo moral grupal y también puede ser empleado para estudiar las relaciones de cooperación y competencia (Quintanilla, 2016a).

Siendo este un tema de debate, es razonable suponer que los chimpancés no tienen propiamente un concepto de creencia falsa, porque este presupone el de creencia verdadera, así como otras dicotomías de igual complejidad, como realidad/apariencia, ser/parecer, saber/creer, etcétera. Todo esto sugiere que, aunque los primates no humanos, en este caso los chimpancés, tienen ciertas habilidades de atribución psicológica y hasta podrían llegar a tener una psicología folk, estas podrían ser significativamente diferentes de las humanas. Más aún, aunque podría haber una arquitectura neurológica de atribución psicológica universal en los seres humanos —y hasta quizá un módulo de atribución psicológica— las diferencias culturales sí marcarían una diferencia.

Algunas investigaciones sugieren que, aunque los niños de habla quechua del departamento de Junín, en el Perú, recién pasan la prueba de la falsa creencia cuando se acercan a la adolescencia (Astington, 1996), sí tienen perfecta comprensión de la distinción entre realidad y apariencia (Vinden, 1996). La tesis de Penelope Vinden es que esta incongruencia podría deberse a una de dos causas: (1) en el habla de estos niños no hay términos para referir con suficiente precisión a creencias. Así, no se les pregunta «¿Qué cree A?», sino «¿Qué diría A?». (2) Otra opción —no muy diferente de la anterior— es que estos niños tienen una concepción de la creencia distinta de la concepción habitual occidental. En otras palabras, aunque los niños quechuahablantes de Junín distinguen entre verdad y falsedad, realidad y apariencia y, sin duda,

tienen estrategias de atribución psicológica tan sofisticadas como cualquier otro niño de esa edad, su cultura hace que o no entiendan las indicaciones de la prueba —eso sería lo menos importante— o que tengan una psicología folk con algunos rasgos diferentes de los occidentales en la dimensión fenoménica, lo que sería más interesante desde un punto de vista filosófico. Más aún, otras investigaciones sugieren que si se modifican los términos de la prueba, niños tan pequeños como de 15 meses de edad ya pueden reconocer creencias falsas (Onishi & Baillargeon, 2005).

Naturalmente el debate no está cerrado; y en cierto sentido recién comienza. Ha habido mucha discusión en lo que respecta a esta edad tan temprana para reconocer creencias falsas, incluida una interpretación diferente de los resultados propuesta por Josef Perner y Ted Ruffman (2005). No sería extraño, en todo caso, que hacia los tres años los niños tuvieran procesos metacognitivos asociados a la atribución de creencias falsas, mientras que niños mucho más pequeños pudieran reconocer situaciones de falsas creencias de manera más intuitiva y no consciente.

2.2. La perspectiva de primera persona

Hacia comienzos de la década de 1990 apareció un nuevo enfoque en la discusión sobre atribución psicológica que recibió el nombre de «perspectiva de la primera persona». Este modelo parte de suponer que la intérprete tiene un conocimiento intuitivo, introspectivo e inmediato de sus propios estados mentales. Sostiene que ella atribuye al agente los estados mentales que ella cree que tendría si estuviera en las circunstancias en que ella cree que él se encuentra, imaginando los estados mentales del agente, imaginándose a sí misma en esos estados o imaginando ser el agente.

De acuerdo con esta posición no es necesario postular un módulo de atribución psicológica conformado nomológicamente, sino solo una capacidad innata que permite imaginar los estados mentales ajenos de manera contrafáctica como si fueran propios, o que permite imaginarse a una misma siendo otra persona y en otras circunstancias vitales. Esta capacidad recibe el nombre de «simulación», por lo que a veces se denomina a este modelo «teoría de la simulación» o S-T (*simulation theory*). Desde esta concepción, el punto de inicio de la interpretación es el conocimiento intuitivo que la intérprete tiene de sus propios estados mentales, es decir, la metacognición. A partir de este autoconocimiento, la intérprete simularía ser el otro en condiciones contrafácticas y atribuiría los estados mentales que ella considera que tendría si estuviera pasando por las circunstancias que ella cree está pasando el agente, en el modelo de Alvin Goldman (1992b, 1993, 1995a, 1995b, 2006, 2013) o que ella cree que tendría si ella fuera el otro, en el modelo de Robert Gordon (1986, 1992, 1995a, 1995b, 1995c, 1995d).

La perspectiva de primera persona se ha visto fortalecida con el descubrimiento de las neuronas espejo, lo que sugiere que la interpretación y la atribución psicológica, así como la comprensión en general, pasan por un proceso de identificación automática y no consciente. Las neuronas espejo son un tipo de neuronas que se activan cuando una persona observa a otra teniendo una experiencia o realizando una acción, de manera que la observadora experimenta literalmente la experiencia o acción del observado (Rizzolatti & Craighero, 2004; Goldman, 2006). Se ha encontrado sistemas de neuronas espejo en primates no humanos, de hecho, fueron descubiertos originalmente en macacos por el equipo de Giacomo Rizzolatti en la Universidad de Parma, en las décadas de 1980 y 1990. También es altamente probable que los perros y algunos tipos de aves tengan sistemas neuronales con funciones parecidas a las de las neuronas espejo de los primates. En los humanos se ha encontrado actividad neuronal que sugiere la existencia de neuronas espejo en zonas como la corteza premotora, la corteza primaria somatosensorial y la corteza parietal inferior. Aunque se sigue investigando sobre el tema, es altamente probable que el sistema de neuronas espejo sea la base fisiológica de funciones como la comprensión automática de intenciones, la empatía, la capacidad lingüística, la imitación motriz —*motor mimicri*— y hasta la autoconciencia (Oberman & Ramachandran, 2009). Se sabe que hay déficits de actividad en las neuronas espejo en los niños autistas y se cree que las mujeres tienen sistemas de neuronas espejo más activos que los hombres, en quienes hay muchas más posibilidades de autismo que en las mujeres (Cheng, Lee, Yang, Lin & Decety, 2008). La menor presencia de oxitocina está asociada al autismo. La oxitocina es una hormona que se sintetiza en el hipotálamo, que se encuentra en la base del cerebro, desde donde se distribuye por el cerebro y el resto del cuerpo. Se sabe que esta hormona tiene un rol central en la promoción de cohesión social, pues genera sentimientos de vínculo, solidaridad y cooperación (McCall & Singer, 2012). Recién se está comenzando a entender el importante rol que la oxitocina tiene en la vida social y afectiva humana. Se sabe que las mujeres embarazadas producen mucha de esta hormona y que su presencia genera actitudes de confianza para con los miembros del endogrupo, aunque también de defensa del endogrupo frente al exogrupo, si es que hay algún grado de competencia entre ambos. En situaciones experimentales, cuando un individuo es conducido a esnifar oxitocina, mejora su capacidad de reconocimiento de rostros humanos y, en general, sus vínculos sociales, aunque solo por un tiempo determinado. Se sabe también que la aplicación de oxitocina por vía intravenosa permite que los autistas mejoren en sus habilidades de atribución psicológica, aunque el efecto se va perdiendo en el curso de algunas semanas (Hollander, 2007).

Las teorías de la simulación han sido empleadas para explicar la complejidad psicológica de diversos fenómenos en que una intérprete se pone en el lugar de otra persona.

Esto ocurre, por ejemplo, en las experiencias de compasión y solidaridad que están presentes en el comportamiento moral, pero también en la experiencia estética propia de la literatura, el cine o el teatro, en que uno simula ser un personaje y vive sus experiencias como propias. El caso paradigmático de esto último es la catarsis aristotélica (*Poética* 1449b), en que uno no solo experimenta las vivencias ajenas sino, además, permite que estas experiencias lo transformen interiormente al generar cierto desahogo mediante la liberación de contenidos psíquicos tóxicos, a la manera de una «purga», como dice Aristóteles, o una «purificación». También se ha propuesto la existencia de un «simulador interno», el cual permitiría que uno imagine sus propios estados mentales pasados y sus posibles estados mentales futuros, así como posibles escenarios objetivos y subjetivos alternativos, en caso que uno opte por una acción u otra. Pienso que el simulador interno sería una pieza clave para explicar el libre albedrío, pero he desarrollado esa tesis en otros lugares (Quintanilla, 2011, 2013, 2014e, 2017a).

Aunque el modelo de la simulación es históricamente posterior a la perspectiva de tercera persona, puede notarse en él cierto espíritu cartesiano, además de que pueden rastrearse algunos de sus rasgos en las diversas teorías de la *Verstehen* de la hermenéutica alemana, básicamente con Schleiermacher (1986, 1996) y Dilthey (1989). Para estos autores, la comprensión pasa por la capacidad para revivir las vivencias (*erlebnis*) del otro. En esa concepción, el término «vivencia» sería cercano a lo que hoy entendemos por conciencia de nuestros estados mentales.

El término «conciencia» tiene una gran carga conceptual y ha sido muy elaborado filosóficamente. Actualmente se suele distinguir entre la conciencia nuclear y la conciencia autobiográfica o extendida (Damasio, 2001). La primera alude a la capacidad de experimentar un estado mental —por ejemplo, una sensación, un dolor, el hambre o el sueño— y en principio está presente, aunque con diversos grados de desarrollo, en las especies animales dotadas de sistema nervioso central suficientemente complejo. La conciencia autobiográfica, por otra parte, es una evolución de la conciencia nuclear e incluye niveles metacognitivos, esto es, la capacidad de tener estados mentales acerca de otros estados mentales; la referencialidad, que permite reconocer los estados mentales como integrados entre sí y propios de uno, y la facultad de percibirse a uno mismo percibiendo o la posibilidad de observarse a uno mismo como si fuese desde fuera —que es la definición aristotélica de autoconciencia (*Acerca del alma*, parte III)—. Esto permite verse a uno mismo como una historia que fluye en el tiempo, incluyendo distintos momentos que están causalmente relacionados entre sí y que se suceden a partir de las decisiones que uno toma. La conciencia autobiográfica incluye al yo autobiográfico o *self*, es decir, a la experiencia —tanto implícita como explícita— de que soy una historia que tiene pasado,

presente y futuro, y que fluye en una dirección que depende de las acciones que yo realice voluntariamente. La conciencia autobiográfica también permite resignificar los acontecimientos del pasado según sus resultados en el presente o de acuerdo a las interpretaciones que uno haría de su propio presente. Hasta donde se sabe, solo el ser humano tiene conciencia autobiográfica, aunque podría ser que los chimpancés tuvieran algún grado rudimentario de esta y es muy probable que los australopitecos ya tuvieran un grado significativo de este tipo de conciencia. Es importante notar, sin embargo, que hay un continuo entre la autoconciencia, la conciencia y la vida no consciente. Esa es una tesis que está presente en Charles Darwin y que los pragmatistas clásicos integraron con la filosofía (James, 1983 [1890], 1904; Dewey, 1910; véase también Godfrey-Smith, 2017 y Popp, 2007).

Es posible decir que cuando hay conciencia autobiográfica está presente también el yo o el *self*, también traducido al castellano como el «sí mismo». Cuando eso ocurre, el individuo se ve a sí mismo como agente de acciones, experimentador de afectos, sujeto de creencias y como una historia que fluye en el tiempo. Pero algo fundamental es que, hasta donde sabemos, la conciencia autobiográfica solo está presente donde hay lenguaje, de manera que el agente consciente puede autodescribirse y es capaz de relatar una historia de sí mismo como una identidad que fluye en el tiempo.

Durante la década de 1990 se desarrolló un debate entre las perspectivas de primera y tercera persona que también recibió el nombre de debate entre la teoría de la teoría (T-T) y la teoría de la simulación (S-T). Al día de hoy, hay cierto consenso en que ambas posiciones no son incompatibles y que probablemente deban ser integradas entre sí, lo que ha generado la aparición de modelos híbridos. No hay acuerdo, sin embargo, respecto de cómo podrían estos modelos ser integrados ni sobre qué elementos de cada perspectiva deberían sobrevivir.

2.3. Segunda persona y triangulación

Además de las perspectivas de tercera y primera persona, está la de segunda persona, que pretende ser una superación de las anteriores. Aunque el autor que acuñó la expresión *second person* en filosofía de la mente fue Davidson (2001d [1992]), él no desarrolló esta posición posteriormente, no participó de los debates con las otras perspectivas ni tampoco tuvo como objetivo superar la polémica entre T-T y S-T, y son muchos los autores que sostienen que su posición sigue siendo de tercera persona. No obstante, autores posteriores emplearon la expresión «perspectiva de segunda persona» para aludir a un contacto más real, menos teórico y observacional, y más comprometido e involucrado con el «tú» o el «ustedes», conformando algo así como un «nosotros» que incluye y no excluye al otro.

En el período de madurez de su producción filosófica Davidson (2001b) desarrolló un modelo triangular, algunas de cuyas intuiciones básicas pueden rastrearse hasta la semiótica triádica de Charles Sanders Peirce y la filosofía social de George Herbert Mead (1972 [1934]). En el modelo de Davidson, los vértices del triángulo son el hablante, la intérprete y el mundo objetivo que ambos comparten[1]. Según este autor, una intérprete solo puede atribuir estados mentales a un agente, así como significados a sus proferencias verbales, en relación con un mundo objetivo común a ambos. Esto es así porque ella interpreta las proferencias del agente mediante mecanismos de atribución que conforman oraciones-T con la siguiente estructura lógica:

(T) La oración «s» es verdadera si y solo si p.

En este esquema formal, la oración «s» proferida por un hablante es verdadera en un lenguaje L si y solo si p, donde «s» y p tienen las mismas condiciones de verdad, es decir, son verdaderas ante las mismas circunstancias del mundo, son verdaderas en los mismos mundos posibles o, lo que para todo efecto práctico es lo mismo, hablante e intérprete estarían dispuestos a proferirlas ante los mismos hechos del entorno. Así es como —de manera holista— la intérprete atribuye estados mentales al agente y significados a sus proferencias, y también reconoce como acciones intencionales a los eventos físicos que constituyen su comportamiento, confrontándolos con los hechos objetivos del mundo compartido.

Ahora vamos a detenernos en los aspectos generales de la triangulación. Davidson usa su modelo para explicar cuatro fenómenos puntuales: (1) las condiciones necesarias para que emerja el pensamiento y el lenguaje, (2) la manera en que se aprende una lengua, (3) la forma en que se determinan los contenidos de nuestros estados mentales en la situación interpretativa (Davidson, 2001e [1991], 2001d [1992], 2001c [1997]) y (4) la interdependencia del conocimiento del mundo exterior, el conocimiento de las otras mentes y el autoconocimiento (Davidson, 2001e [1991])[2].

El modelo davidsoniano de la triangulación ha sido muy discutido, con argumentos a favor y en contra. Puede verse, por ejemplo, Yalowitz (1999), Lasonen y Marvan (2004), Verheggen (2006, 2007), Bouma (2006) y Glüer (2006), entre otros. Más allá de Davidson, sin embargo, deseo usar un modelo triangular para explicar otros fenómenos que él no aborda, pues hay importantes conexiones lógicas, epistemológicas, filogenéticas y ontogenéticas que se pueden iluminar mediante este modelo (Quintanilla, 2014b).

[1] En el modelo triangular de Peirce los vértices son hablante, intérprete y signo o lenguaje.

[2] Pedace (2017) menciona las primeras tres como parte del programa davidsoniano, pero creo que la cuarta también debe incluirse.

(1) Desde un punto de vista lógico, los conceptos de yo, otros —o ustedes— y mundo son interdependientes y no se pueden definir si no es en relación a los otros; lo mismo ocurre entre los conceptos de subjetividad, intersubjetividad y objetividad.

(2) Desde un punto de vista epistemológico, el conocimiento que uno tiene de sí mismo —de sus propios estados mentales— solo es posible gracias al conocimiento que se tiene de los otros —de los estados mentales ajenos— y al conocimiento del mundo exterior objetivo y compartido con los demás. Pero lo mismo ocurre con el conocimiento de la objetividad, que presupone el autoconocimiento y el aleoconocimiento, y con el conocimiento de los otros, que presupone nuestro común conocimiento del mundo compartido y de nosotros mismos. En otras palabras, no comenzamos con un conocimiento intuitivo, inmediato e incorregible de nuestros estados mentales —como ocurre en el modelo cartesiano— para luego pasar a las otras dos formas de conocimiento. En esto hay una clara distancia respecto de la perspectiva de primera persona.

(3) Desde un punto de vista filogenético, no parece razonable suponer que primero evolucionara alguno de los tres vértices de este triángulo y que esto acarreara la evolución de los otros dos, es decir, que la selección natural seleccionara el autoconocimiento, el aleoconocimiento o el conocimiento objetivo para que luego alguno de ellos diera lugar a los otros dos. Es más razonable suponer que se trató de una evolución simultánea y que las tres formas de conocimiento fueron seleccionadas simultáneamente y por la misma razón adaptativa. Para una defensa de esta tesis contra las posiciones de Carruthers (1996a, 2009)[3], Nichols y Stich (2003) y Goldman (1995b, 2006), véase Quintanilla (2014b).

(4) Desde un punto de vista ontogenético, es razonable suponer que el infante desarrolla el conocimiento de sus propios estados mentales de manera simultánea al conocimiento que adquiere de los estados mentales ajenos —particularmente el de la madre o el cuidador— y a su reconocimiento de que ellos comparten el mismo mundo (Quintanilla, 2017b). Así es como la subjetividad se constituye en un entorno intersubjetivo y frente a la objetividad compartida. De igual manera, el bebe adquiere los conceptos de «yo», «otros» y «mundo» simultáneamente y mientras lo hace construye dualidades fundamentales como verdadero/falso, ser/parecer y realidad/ficción.

[3] Carruthers cree que primero fue seleccionada la lectura de mentes y que luego esta giró sobre sí misma para dar lugar a la metacognición.

En este terreno, el psicoanálisis ha hecho aportes significativos que pocas veces se integran con los desarrollos en filosofía de la mente (una de las pocas excepciones es Cavell, 1993, 1998, 2006). Por ejemplo, Winnicott (1971) ha acuñado el concepto de «espacio transicional», que también tiene una forma triangular conformada por el bebe, la madre y el objeto transicional, que sería, por ejemplo, un osito de peluche. Este es un espacio que no es puramente subjetivo —como la representación del osito— ni tampoco objetivo —como el osito mismo—, y que está conformado por los objetos transicionales que ayudarán al bebe a crear su mundo subjetivo, pero también a formarse representaciones de lo objetivo, es decir, de aquellos objetos que son independientes de su mente y su voluntad, objetos que comparte —y que sabe que comparte— con otras personas. El espacio transicional permite que se constituya la experiencia psíquica del bebe y con el desarrollo se convierte, en el adulto, en el espacio cultural de lo lúdico, el arte y la imaginación creadora.

Hay un interesante debate en torno a si el triángulo comienza con una relación diádica entre la madre y el bebe, como cree Stern (1991), para luego incorporar al mundo objetivo, o si el proceso nace en su forma triangular. En este último caso, aunque al comienzo el bebe está concentrado en el rostro materno y reacciona principalmente ante sus gestos, también reacciona ante otros objetos del mundo, aunque no de manera coordinada. Es decir, reacciona ante la madre y reacciona ante otros objetos, pero todavía no reacciona ante la relación. Si a esos momentos tempranos debemos llamarlos «relaciones diádicas» o «triádicas» es, en gran medida, un asunto estipulativo, aunque me inclino por pensar que ya se trata de un triángulo, aunque bastante rudimentario. La razón de ello es que el infante reaccionaría de manera diferente si un tercer objeto interactuase con la madre. En todo caso, incluso si fuera cierto que debemos interpretar las primeras reacciones del bebe ante la madre y el mundo como diádicas y carentes de una estructura triangular, de allí no se sigue que cuando el triángulo ya está constituido en el nivel de la atención compartida, la relación diádica con el otro tenga prioridad o sea más fundamental que las otras relaciones que conforman el triángulo, como algunos sostienen. Regresaré sobre este tema hacia el final de este capítulo.

Lo importante es que, en este espacio triangular, el bebe, progresivamente, se descubre a sí mismo, a los demás —a la madre—, y a la realidad que comparten y que es diferente de ambos. Con el desarrollo, podrá atribuir estados mentales diferentes de los suyos a los demás, pero sobre los mismos objetos de la realidad. Para conocer a los demás deberá interpretarlos atribuyéndoles estados mentales semejantes a los que él reconoce en sí mismo. Para conocer la realidad compartida deberá comparar sus estados mentales con los de los demás que, siendo diferentes, son acerca de la misma realidad. Y para conocerse a sí mismo deberá coordinar sus propias experiencias fenoménicas subjetivas con los estados mentales que reconoce en los demás acerca del mundo que comparten.

Un ejemplo de esto último es el concepto de *reverie* desarrollado por Bion (1962, 1963, 1965). La palabra *reverie* procede del francés antiguo que, hacia el siglo XIV, refería a un estado de delirio placentero semejante a la embriaguez, en el que uno se siente en comunión con otros. Bion desarrolló el concepto de «*reverie* materno» para describir el proceso por el cual la madre acompaña al infante, tanto afectiva como cognitivamente, permitiéndole configurar y moldear su aparato psíquico. Así, por ejemplo, ante un posible cólico estomacal del bebe —que para él es una experiencia novedosa y aterradora— la madre lo tranquilizará, le explicará que se trata de algo pasajero e inofensivo y además le pondrá nombre, de manera que, en adelante, cada vez que él tenga esa misma sensación, sabrá de lo que se trata y ese saber le generará menos angustia. Algo semejante ocurrirá con la experiencia de otros estados mentales y el aprendizaje de cómo nombrarlos y qué esperar de ellos.

En el fenómeno de *reverie*, la madre no solo cumple una función de contención sino también permite al bebe aprender a comprender, procesar, elaborar, clasificar y hasta nombrar sus propios estados mentales. De esta manera, el infante aprende a experimentarlos, regularlos, describirlos y hasta predecirlos. Así es como él se incorpora en una forma de vida, la cual está conformada por prácticas sociales compartidas que hacen posible que uno experimente su propia subjetividad de manera análoga a como otros miembros de su comunidad lo hacen, aunque, por supuesto, sobre la base de una estructura biológica y neurológica compartida. Es a través de la percepción no consciente del cuerpo —ya sea por imitación motriz o con la participación de las neuronas espejo— que madre y bebe sintonizan afectivamente.

Todo este proceso es inseparable de la adquisición del lenguaje, pues no es que uno sepa primero referir a los objetos del mundo externo para luego poder nombrar sus experiencias psíquicas. El niño aprende a nombrar los objetos del mundo externo en tanto es capaz de reconocer a las otras personas como agentes diferenciados de los objetos que son solo físicos. Pero al mismo tiempo, experimenta estados internos como dolores, emociones y deseos, y los ordena, categorizando y nombrando.

Pero el punto que deseo subrayar es que no es que el infante tome conciencia de sus estados mentales antes de poder atribuirlos a otras personas. Tampoco ocurre lo contrario. Se trata de un triángulo en el que los tres vértices se dan simultáneamente y así es como se constituye intersubjetivamente la subjetividad, en un entorno objetivo compartido. En otras palabras, lo subjetivo, lo intersubjetivo y lo objetivo emergen en simultáneo pues se necesitan mutuamente. Ninguno de los tres puede darse sin los otros dos.

La idea es que sería un error suponer que el infante vive en un mundo narcisista y egocéntrico antes de descubrir que existe el otro. Eso sería un remanente cartesiano anterior a los desarrollos teóricos sobre la constitución intersubjetiva de la subjetividad.

El infante se reconoce a sí mismo porque puede reconocer al otro, porque puede reconocerse en el otro o porque puede reconocer al otro en sí mismo; y ciertamente la capacidad de *reverie* es fundamental para que este proceso pueda ocurrir.

Algo de esto también reaparece en la vida adulta, siendo un ejemplo paradigmático de ello lo que acontece en el consultorio de la analista o psicoterapeuta, donde también se conforma un triángulo entre el paciente, la analista y el mundo objetivo que ambos asumen compartir, el cual ha sido en gran medida construido por ellos mismos pero anclado en el mundo real. En un proceso terapéutico, este triángulo se amplía y robustece, haciendo posible en el paciente, pero también en la terapeuta, un espacio subjetivo cada vez más rico y valioso.

En el terreno de la filosofía, el rechazo a una subjetividad previa a la intersubjetividad y la idea de que esta se constituye intersubjetivamente comenzó a desarrollarse, hacia fines del siglo XIX, con Peirce (1988) y su cuestionamiento al modelo cartesiano de lo mental, y también con James, en su libro de 1890, *The Principles of Psychology* (1983), en el que describe al yo como constituido interpersonalmente. James va tan lejos, para la época, como para decir que los vínculos son parte de lo que el yo literalmente es y que incluso los objetos del mundo que nos importan son parte del yo. En Husserl (1973a) también hay un valioso desarrollo de estas intuiciones, lo que influyó notablemente en la filosofía europea. En el caso de la filosofía de la mente actual, las ideas de Peirce y James han influido en las discusiones recientes sobre lo que se denomina *the embodied self* y *the extended self,* que tienen a Clark (1999, 2008, 2016) y Chalmers (1996, 2010, 2012) como algunos de sus más interesantes representantes. La idea en esta concepción es que los límites de la mente no son las fronteras del cerebro, pues la mente incluye todos aquellos objetos que permiten y potencian su cognición y su afectividad. Así, por ejemplo, en tanto mi computadora me permite realizar y almacenar procesos computacionales cognitivos que probablemente no podría hacer sin ella, hay un importante sentido en que esta es parte de mi mente. Análogamente, como he insistido, las personas con las que tengo importantes vínculos afectivos constituyen parte de lo que yo soy. Se trata, pues, de debilitar o aligerar la frontera entre la mente y lo exterior a ella, de la misma manera que un pianista profesional podría decir que cuando toca su piano ha suavizado la frontera entre su cuerpo y ese instrumento.

Volvamos a la versión del modelo triangular que deseo defender. Tanto la intérprete como el agente reconocen los hechos de la realidad frente a un marco intersubjetivo que incluye a los interlocutores, por lo que el reconocimiento de los hechos objetivos del mundo requiere como condición necesaria la interacción con otros intérpretes y agentes. De manera análoga, tanto nuestra experiencia del mundo objetivo, de nosotros mismos, como de nuestro reconocimiento de la subjetividad ajena,

depende de que somos parte de una comunidad social. Por eso la intérprete no puede acceder introspectivamente a sus propios estados mentales si no reconoce, en simultáneo, los estados mentales ajenos y los hechos objetivos que ambos comparten. Ella tampoco puede atribuir estados mentales al agente si no reconoce, al mismo tiempo, sus propios estados mentales y los hechos objetivos compartidos. Finalmente, tampoco puede reconocer la objetividad de los hechos del mundo si no reconoce, simultáneamente, sus propios estados mentales y los del agente. Ninguno de los tres vértices del triángulo es el punto de partida, el punto de partida es el triángulo como tal. Además, no hay acceso a cualquiera de estos vértices si no se pasa por los otros dos. Es claro que intérprete y agente están siempre sustituyendo roles, de manera que tanto la interpretación como la comunicación son posibles porque las atribuciones que mutuamente se hacen se afinan ante la evidencia del nuevo comportamiento del otro, en relación con la realidad objetiva que comparten.

Este es un apretado resumen, que ampliaré posteriormente, de la manera en que se puede entender el carácter triangular de la interpretación. Como señalé, Davidson acuñó en este sentido las expresiones «triangulación» y «segunda persona», que están articuladas entre sí, pero no integró sus hipótesis especulativas con información empírica ni tampoco extrajo las consecuencias filosóficas de esta integración. Hacer eso es parte de mi interés en este libro y en artículos publicados previamente.

Ahora bien, antes de regresar a ver cómo este modelo podría superar los modelos de primera y tercera personas, es necesario detenernos un momento en el aspecto ontogenético de la emergencia de la triangulación. Un análisis detallado de la ontogénesis del triángulo puede encontrarse en Quintanilla (2017b).

Aunque hay discusión sobre cuáles son los mecanismos por los que la triangulación emergió y si el triángulo procede de una relación diádica previa, hay cierto acuerdo en que el triángulo —ya constituido de manera robusta— implica procesos cognitivos y afectivos relativamente maduros que se encuentran en el niño hacia los 3 o 4 años. Pero antes de eso, naturalmente, hay todo un proceso de emergencia que conduce hacia él. También hay algún consenso en que el fenómeno anterior a la constitución del triángulo es lo que se suele llamar atención conjunta (*joint attention*)[4]. Tomasello (1999) es uno de los investigadores que más ha ampliado las fronteras de este terreno interdisciplinario en investigaciones con niños y con primates.

La definición clásica de atención compartida alude al fenómeno por el que dos o más personas —en el ámbito ontogenético se trata de un infante y un adulto—

[4] Algunos autores usan de manera intercambiable «atención conjunta» (*joint attention*) y «atención compartida» (*shared attention*). Yo prefiero hacer una distinción entre ambas expresiones: la primera es ontogenéticamente anterior a la segunda y condición de posibilidad de ella. Véase Quintanilla, 2017b.

intercambian gestos y miradas para compartir su atención hacia un objeto o acontecimiento reconocido como común para ambos, pero también hacia las reacciones que cada uno de ellos tiene acerca del objeto y del otro, lo que también es reconocido como una experiencia compartida. La atención compartida es un hito fundamental para el posterior desarrollo de la teoría de la mentalización —teoría de la mente, lectura de mentes, atribución psicológica, psicología folk, metarrepresentación, simulación, etcétera— y para la adquisición del lenguaje. Una manera un poco más breve de definir la atención compartida es esta: se trata del fenómeno por el que dos criaturas reaccionan ante un mismo objeto y ante las reacciones de cada una de ellas —así como ante las reacciones de cada una de ellas ante la otra— de manera intercambiable y sucesiva, con lo que se reconoce que hay algo que están compartiendo.

Los bebes muy pequeños ya se involucran en intercambios de sonrisas y gestos con el cuidador, lo que podría sugerir que se trata de una relación diádica que luego se transformará en triádica. Pero también podría suponerse que se trata de una relación triádica desde el inicio, porque el bebe reacciona ante los gestos de la madre y también ante otros objetos del mundo que no son ella, aunque todavía no reacciona comparativamente ante ambos. Es decir, reacciona ante la madre y reacciona ante otro objeto, pero todavía no reacciona ante la relación que hay entre la madre y el objeto, con lo que tenemos ya un triángulo rudimentario.

Algunos autores han propuesto que la aparición de la atención compartida tiene dos momentos de complejidad (Moll & Meltzoff, 2011; Flavell, Everett, Croft & Flavell, 1981). En un primer nivel, el infante es capaz de reconocer qué objetos están a la vista de otras personas y qué objetos no podrían ser vistos por otras personas, aunque sí por el propio infante. Hay numerosa evidencia de que los chimpancés y los perros alcanzan este nivel. A eso podemos llamar atención conjunta. El segundo nivel aparece cuando el infante no solo reconoce que otras personas ven distintas cosas —si por ejemplo están en un ángulo diferente— sino también que esas personas ven las mismas cosas pero de una manera diferente. Este nivel emerge hacia los tres o cuatro años y, hasta donde sabemos, no se encuentra en otras especies además del *Homo sapiens*. A esto podemos llamar atención compartida.

La atención conjunta aparece en el infante antes que este tenga la capacidad de reconocer y atribuir estados mentales a otros. Entre los once y quince meses, los niños ya están vinculados con otras personas en relación con un objeto compartido. Antes del primer, año los infantes coordinan su atención con otras personas acerca de objetos comunes, aunque todavía no tienen concepto de perspectiva, es decir, ni pueden ponerse en la perspectiva ajena ni se dan cuenta de que el otro tiene una perspectiva diferente; en otras palabras, ven distintas cosas según el lugar donde se encuentren.

Hacia aproximadamente los dos o tres años, el infante adquiere la capacidad de reconocer que los otros tienen perspectivas distintas, lo que es el primer nivel de la atención compartida en la distinción que mantienen autores como Moll y Meltzoff (2011), y Flavell y otros (1981), y que yo prefiero llamar atención conjunta. Entre los tres y cuatro años, el niño pasa al segundo nivel de la atención conjunta (que yo llamaré atención compartida) y no solo reconoce las perspectivas ajenas sino también las puede comparar con las propias, es decir, se da cuenta de que las otras personas ven las mismas cosas que él ve, pero de una manera diferente. Cuando esto ocurre, puede decirse que el niño reconoce que los demás tienen estados mentales diferentes de los suyos, con lo cual está en condiciones de pasar la prueba de la falsa creencia.

Los fenómenos que anteceden a la plena constitución del triángulo, hacia los tres y cuatro años en que el niño ya tiene una teoría de la mente y está en proceso de desarrollar agencia, están siendo muy estudiados. No solo se investiga la aparición de la atención conjunta sino también el rol que otros fenómenos, como paradigmáticamente el apego, cumplen en estos procesos (Mantilla, 2014a y 2014b).

Ahora bien, es claro que la concepción triangular no es una perspectiva de primera persona, porque no hay un acceso introspectivo e incorregible que la intérprete tenga respecto de los contenidos proposicionales de sus estados mentales, ni antes ni independientemente de su reconocimiento de los estados mentales ajenos. Tampoco es un modelo de tercera persona, porque no hay conocimiento de los estados mentales ajenos sin un conocimiento de los propios y de la objetividad de la realidad. Tampoco es una perspectiva de segunda persona —como el que analizaré más adelante—, sino un modelo que se propone integrar las tres perspectivas.

En la triangulación, el niño no solo adquiere los conceptos de tú, otro, él, ellos sino también el de yo, que acompaña a la conciencia de la propia identidad, el saber de sí, el sentido del sí mismo o *self*, que hace posible la conciencia autobiográfica. Cuando eso ocurre, el infante es capaz de diferenciarse de los otros, reconocer que ellos y él pertenecen a una misma red social, e identificar a los demás como semejantes, pero diferentes. La conciencia que tiene un niño de su propia identidad como individuo es inseparable de su conciencia de que hay otros individuos que tienen sus propias identidades. Esta es otra importante idea que desarrolló James (1983 [1890]) hacia fines del siglo XIX: la propia identidad se define en nuestra relación con los otros.

Ahora vamos a concentrarnos en la propuesta de Gomila (2001, 2002, 2003, 2008, 2012; Gomila & Pérez, 2017), uno de los más interesantes representantes actuales de la perspectiva de segunda persona. Él piensa que la comunicación no es solo el proceso de atribuirse mutuamente estados mentales sino también el de atribuirse mutuamente estados mentales sabiendo que estos son compartidos, lo que

implica por lo menos cuatro niveles de intencionalidad. Cree que el modelo estándar de atribución de estados mentales es básicamente cognitivo y racional, por ello, no hace justicia a la complejidad de la comunicación humana cuya base es no cognitiva y no racional. Considera también que intérprete y agente no son observadores desvinculados uno del otro, sino individuos que interactúan entre sí, modificando sus atribuciones y sus estados mentales según las reacciones y el comportamiento del otro.

Sostiene que en el fenómeno de la comunicación e interpretación intersubjetiva se constituye un vínculo que va más allá de las atribuciones explícitas de estados mentales y significados (Gomila, 2001, p. 67). Esto ocurre mediante patrones de interacción que involucran la «atribución implícita y recíproca de estados mentales» (2001, p. 70). Afirma también que la comunicación intersubjetiva «depende de procesos y mecanismos perceptivos más básicos» que no son «cognitivamente penetrables» (p. 68). Por ello, el tipo de «conocimiento» que uno tiene del otro, según la perspectiva de segunda persona, «no es proposicional sino práctico» (p. 68). Considera también que los modelos de primera y tercera persona no dan cuenta del fenómeno del paralenguaje —la comunicación no lingüística que acompaña al lenguaje— que está presente en la intersubjetividad (p. 72). Señala que la selección natural no habría seleccionado criaturas egoístas interesadas solo en sus propias necesidades, sino individuos que forman parte de formas de vida que requieren de mecanismos de reconocimiento social, lo que implica complejos mecanismos de creación de significado y coordinación práctica. Adicionalmente, opina que, en contra de lo supuesto por Freud y Piaget, el niño no parte de una orientación egocéntrica que progresivamente se hace social, sino tiene ya una disposición social desde el inicio.

En este último punto, la tesis principal de Gomila es que el adoptar la perspectiva de segunda persona es condición necesaria para el comportamiento moral. Afirma que «las «actitudes reactivas» implícitas en las emociones morales presuponen constitutivamente la perspectiva de segunda persona y, en esa medida, constituyen la base última de la obligación moral. Dicho de otro modo, la base de la moralidad radica en nuestra capacidad de interactuar con los demás desde la segunda persona, es decir, desde la intersubjetividad» (2009, p. 2).

La tesis general es que la mente se forma y moldea en las interacciones con el entorno físico y social. Pero este proceso comienza, ontogenéticamente —y quizá filogenéticamente también— con mecanismos no conscientes e implícitos de reacción emocional ante la conducta ajena (2009, p. 13), aunque aún no de reconocimiento de sus estados mentales, lo que progresivamente se convierte en mecanismos de atribución psicológica, conscientes y explícitos.

Considero que las tesis de Gomila son básicamente correctas. Lo que me parece equivocado es suponer que «las atribuciones de segunda persona son más básicas y es a través de ellas que se adquieren los conceptos mentales involucrados en las atribuciones de primera y tercera personas» (Gomila & Pérez, 2017, p. 279). También considero equivocado que para la atribución psicológica «no es indispensable la existencia de un mundo compartido» (2017, p. 277).

Por ello, sostengo que el modelo triangular tiene la capacidad de integrar algunas de las virtudes explicativas de las tres perspectivas en aparente disputa. Más aún, aunque considero que las tesis de Gomila son básicamente correctas, creo que las diversas teorías que intentan explicar la atribución psicológica no tienen que negar ninguna de ellas y, menos aún, el modelo triangular. Bastaría con que sostengan que explican el proceso ya maduro, que seguramente contienen elementos no conscientes e implícitos. Al mismo tiempo, estas teorías pueden desear hacer una reconstrucción racional de algo que tiene una estructura menos explícita.

No parece que los argumentos de Gomila sean efectivos contra el modelo triangular. Por una parte, este modelo insiste en que la interpretación es un fenómeno dinámico donde intérprete y agente se transforman mutuamente en tanto interactúan entre sí. Es necesario, sin embargo, investigar conceptual y empíricamente los orígenes de la triangulación, algo que se está desarrollando mucho últimamente desde la filosofía y la psicología.

En mi opinión, aunque la versión de segunda persona de Gomila podría añadir a las perspectivas de primera y tercera elementos que faltarían a estas, no llega a integrar los elementos explicativos que estas tienen, algo que sí hace el modelo triangular y que, por tanto, resulta más completo. Es innegable que la interpretación y la comunicación intersubjetiva requieren de mecanismos afectivos y no conscientes que van más allá de la atribución racional y consciente de estados mentales con contenido cognitivo y proposicional, todo lo cual se expresa en la complejidad del paralenguaje. También es indiscutible que la comunicación infantil tiene esas características, a partir de las cuales emerge otro tipo de comunicación. Pero considero que esto no es incompatible con las tesis explícitas del modelo de primera y tercera personas, y menos aún con el modelo triangular. Por tanto, no me parece que estos sean buenos argumentos en contra de esos modelos que se proponen elaborar una suerte de reconstrucción racional de la estructura lógica de la interpretación y de la comunicación intersubjetiva. No creo que ellos necesariamente asuman que agotan el fenómeno que quieren explicar, sino más bien piensan que están haciendo una suerte de radiografía de la estructura de la intersubjetividad, aunque reconozcan que hay mucho material importante que no sale en la descripción. Parece beneficioso que Gomila subraye aspectos de la intersubjetividad que con frecuencia no son suficientemente iluminados, pero

no resulta claro que sus argumentos lleguen a objetar los modelos que pretende cuestionar. Es también positivo que Gomila subraye el fenómeno de la creación del vínculo, que no se reduce a un conjunto de atribuciones, sino que más bien puede ser visto como la creación de un espacio común compartido. En efecto, he empleado la metáfora de la creación de un «espacio compartido» —que desarrollaré y discutiré más adelante— para analizar una intuición que, me parece, está en Gadamer —la fusión de horizontes— y en Wittgenstein —el compartir formas de vida—.

Otras autoras que han investigado en la línea de la perspectiva de segunda persona son Scotto (2002, 2004), Pedace (2012, 2017) y Pérez (2013). En todos estos casos hay un legítimo y necesario intento por completar un vértice del triángulo que se encuentra poco desarrollado. Las autoras mencionadas subrayan que las tres perspectivas son compatibles y que no se excluyen entre sí, pero afirman la prioridad lógica y anterioridad ontogenética de la perspectiva de segunda persona.

Pedace sostiene que «sin la perspectiva de *segunda* persona no podría tener lugar la *tercera,* esto es, sin transitar una dimensión participativa o de involucramiento con otro próximo, no podríamos llegar a ser intérpretes distantes de la conducta de un tercero» (2017, p. 328). Esto sugiere que antes del triángulo hay una díada. Sin embargo, inmediatamente después señala, correctamente, a mi juicio, que «dado el hecho de que algunas emociones son intencionales, el mundo entra a través de nuestras emociones» (2017, pp. 328-329), cerrando de esta manera el triángulo.

Discrepo con la idea de que la perspectiva de segunda persona sea ontogenéticamente anterior a la de tercera persona, aunque coincido con la tesis de que la díada madre-bebe incluye ya como un tercer elemento el mundo. Pedace señala que esto ocurre gracias al carácter intencional de las emociones, lo que me parece acertado, pero además de ello, no puede haber una díada inicial que involucre solo a la madre y el bebe, porque el bebe está ya en condiciones de distinguir entre la madre y todo lo que no es ella, es decir, el mundo objetivo que comparte con ella. Si el debate, entonces, es si el punto inicial es un bebe en su soledad —como en algún momento creyó Freud— o la díada madre-bebe —como sostienen algunas posiciones defensoras de la perspectiva de la segunda persona, así como Stern (1991)—, me inclino por sostener que lo que tenemos desde el inicio es una relación triangular, ninguno de cuyos vértices antecede a los otros, lo que coincide con la posición de Cavell (2006).

Las perspectivas de primera y tercera persona constituyen una reconstrucción racional de lo que ocurre en el proceso ya maduro de atribución psicológica, mientras que la perspectiva de segunda persona se propone hacer un análisis más fino de la ontogénesis y de los aspectos relacionales y vinculares de la atribución psicológica. En principio no tendría que haber incompatibilidad entre estas tres perspectivas, aunque sería necesario investigar en detalle de qué manera pueden ser integradas.

Hay otros puntos centrales en que me parece que la perspectiva de segunda persona yerra o es insuficiente. Como señalé hace poco, Gomila y Pérez (2017, pp. 277 y 283) sostienen que no es indispensable la existencia de un mundo compartido. Como resultará claro, esta posición es incompatible con el modelo triangular, aunque este último modelo sí puede incorporar algunas características de la perspectiva de segunda persona, como la importancia que asigna al intercambio intersubjetivo afectivo y los elementos no conscientes que participan en la comprensión. Pero lo más importante es que, tal como veo las cosas, la perspectiva de segunda persona no da cuenta del fenómeno de la comprensión, sino de un aspecto de este, que es la empatía. Esto queda claro en tanto esta perspectiva no puede ni desea explicar por qué la gente hace lo que hace o tiene los estados mentales que tiene sobre la base de sus propias razones, de manera que está dejando fuera algo que es fundamental en el fenómeno de la comprensión, que es entender el punto de vista del otro, sus motivaciones y justificaciones, sin por ello perder la propia perspectiva. Por ello, considero que la perspectiva de segunda persona es una buena manera —aunque no la única— de dar cuenta de un aspecto de la complejidad de la interacción humana que a su vez es solo un aspecto de lo que está en juego en el fenómeno de la comprensión. Creo, pues, que se trata de un modelo incompleto.

Al sostener que la perspectiva de segunda persona es anterior y más básica —en sentidos filogenético, ontogenético y lógico— que las otras perspectivas (Gomila & Pérez, 2017, pp. 278-279), sus defensores están dejando de lado otras dos perspectivas que también son necesarias en la constitución de cada una de las perspectivas. Como intento sostener aquí, las tres perspectivas son irreductibles unas en las otras, son simultáneamente básicas y se requieren mutuamente para que pueda darse cualquiera de ellas. Más aún, como dice Escajadillo, «hay hechos o aspectos de la realidad que no están disponibles o no son accesibles desde «cualquier» punto de vista» (2018, p. 264). En efecto, la perspectiva de primera persona puede capturar la experiencia fenoménica subjetiva que no es accesible desde las perspectivas de segunda o tercera personas. La perspectiva de segunda persona puede capturar la experiencia intersubjetiva de la interacción coordinada o de la vida afectiva —cuya forma paradigmática sería, por ejemplo, la experiencia del amor compartido— de una manera que las otras perspectivas no pueden hacerlo. Y la perspectiva de tercera persona logra capturar un poco de la manera como se ve el mundo si uno logra descentrarse de su propia subjetividad para imaginar cómo se vería la realidad desde varios ángulos al mismo tiempo. Todo esto, como resultará claro, proporciona argumentos contra la idea de que la perspectiva de segunda persona sea más básica que las otras y, más aún, que pueda por sí misma dar cuenta del fenómeno de la comprensión.

Mi posición es, entonces, que la perspectiva de primera persona se origina en la introspección; la de segunda persona, en la interacción con otro; y la de tercera

persona, en la extrospección, esto es, en una mirada que pretende ser objetiva, imparcial y externa. Las tres perspectivas son necesarias para la comprensión del otro[5], pero, más aún, las tres deben estar presentes para que puedan constituirse los conceptos de subjetivo, intersubjetivo y objetivo.

En todo caso, las habilidades cognitivas conscientes y conceptuales de la atribución psicológica se desarrollan sobre la base de procesos preconscientes y preconceptuales que han sido propuestos por algunos filósofos desde por lo menos fines del siglo XIX. Estos procesos han recibido diversos nombres, como «prácticas sociales compartidas» —Peirce y el pragmatismo—, «mundo de la vida» —Husserl—, «precomprensión» —Heidegger—, «formas de vida» —Wittgenstein— y «*the background of intentionality*» —Searle—, entre otros.

Es importante resaltar que la comprensión es un fenómeno que incorpora mecanismos conscientes y no conscientes. En el primer caso, la intérprete tiene experiencia fenoménica y metacognición —que podría darse en varios niveles de intencionalidad— de las atribuciones psicológicas que está realizando, y en algunos casos estas atribuciones son incluso el producto de un proceso de deliberación. En el segundo caso, ella registra información de los estados mentales ajenos de manera subliminal, lo que no impide que ese registro genere en ella misma otros estados mentales complejos que la induzcan a tener actitudes emocionales o cognitivas respecto del agente o actitudes conscientes que ella intentará justificar de diversas maneras.

En 1947, Gilbert Ryle distinguió de manera célebre entre dos formas de conocimiento: *know that* —conocimiento proposicional— y *know how*—conocimiento no proposicional o tácito, identificable con habilidades no proposicionales— (1967 [1949]). Según este autor, ambas formas de conocimiento son igualmente valiosas, pero la primera requiere de la segunda, aunque no al revés, siendo el conocimiento de habilidades —*know how*— más básico que el proposicional. En el contexto que nos interesa podría también distinguirse entre comprensión proposicional —*to understand that*— y comprensión basada en habilidades —*to understand how*— siendo esta anterior y más básica que aquella.

La comprensión puede ser lingüística o no lingüística. En el primer caso, la intérprete es capaz de verbalizar los estados mentales que reconoce en el agente y, eventualmente, también podrá formularlos en términos proposicionales. En el segundo caso ella reconoce conscientemente los estados mentales del agente, pero es incapaz de expresarlos verbalmente por no encontrar las categorías léxicas que ella consideraría adecuadas o porque percibe en sí misma una masa caótica de estados mentales que no logra ordenar. Gran parte de lo que ocurre en la comprensión no requiere

[5] Y como sostuve al final del primer capítulo, también para la autocomprensión.

de lenguaje verbal y se expresa a través del paralenguaje —todo lo que acompaña al lenguaje y sirve para la comunicación, como los gestos, las miradas, etcétera—, pero en muchos casos —aunque no en todos— puede ser reconstruido verbalmente.

En un sentido filogenético y ontogenético, la adquisición del lenguaje potencia la atribución psicológica, lo que ocurre tanto en la aleoatribución como en la autoatribución. El poder poner en palabras los estados mentales que reconocemos en los otros y en nosotros mismos nos permite capturarlos con más precisión, así como experimentarlos de maneras más conspicuas, lo que sin duda potencia la autoconciencia, es decir, nos permite no solo experimentar estados mentales sino también tener otros estados mentales acerca de ellos, en varios niveles de intencionalidad. En este punto, las características particulares de las lenguas moldean los estados mentales que se reconocen en los otros e influyen en la manera como uno experimenta los propios. La adquisición del lenguaje también hace posible que se constituyan los contenidos proposicionales. Pero, lo que es particularmente importante, el lenguaje permite reconstruir gran parte de la comprensión no lingüística y no consciente de manera verbal, lo que hace posible que la expliquemos de una manera más precisa.

Estas reflexiones nos conducen a preguntarnos cómo se constituyeron la atribución psicológica y el lenguaje desde un punto de visto filogenético, y cómo se constituyen ontogenéticamente.

El antepasado común entre el *Homo sapiens* y los chimpancés vivió en la sabana africana hace, por lo menos, cinco millones de años y tenía una masa encefálica de alrededor de 450 cc. El cerebro humano moderno tiene una masa encefálica de 1450 cc en promedio. Es necesario preguntarse, por tanto, por qué se produjo ese gran crecimiento en el caso humano, si el cerebro de los chimpancés es prácticamente el mismo en términos de volumen y si ambas especies cohabitaron el mismo entorno.

La teoría más aceptada es la hipótesis del cerebro social de Dunbar (2003, 2009). De acuerdo con esta hipótesis, el aumento de individuos en la tropilla de primates —de un aproximado de 50 a 150, en el caso de nuestros antepasados en los últimos tres millones de años— tuvo como consecuencia que aumentaran también los conflictos sociales, las relaciones de jerarquía y poder, y la necesidad de cooperar y competir para poder sobrevivir. Esto condujo, a su vez, a que aumentase la complejidad social a la que los individuos tenían que adaptarse. En la sabana africana, tanto como ahora, para sobrevivir había que adaptarse a un entorno peligroso, pero también a un ambiente social extremadamente complejo.

Hace tres millones de años, entonces, el aumento de individuos de la tropilla originó la necesidad de comprender, interpretar y predecir esa complejidad social. Esto implicaba saber, por ejemplo, quién es amigo o enemigo de quién, quién nos traicionó o podría nuevamente traicionarnos, quién es más astuto para engañar y

contra engañar, etcétera. La necesidad de entender y predecir esa complejidad condujo al desarrollo de estados mentales en varios niveles de intencionalidad metacognitiva. No bastaba con creer que A era peligroso. Era necesario saber si B creía que A es peligroso. También si B creía que siendo A peligroso, era menos fuerte que C. Y que siendo C más fuerte que A, B creía que era más fácil de cooperar con A que con C. La tesis de Dunbar, por tanto, es que el aumento de la complejidad social condujo al aumento de los niveles de intencionalidad metacognitiva, lo que condujo a un rápido crecimiento del cerebro, concentrado sobre todo en el neocórtex del lóbulo frontal, específicamente en la corteza prefrontal, que es la zona del cerebro que controla la cognición social (Dunbar, 2003, pp. 164-168).

De esta manera, Dunbar correlaciona el número de individuos de la tropilla con los que uno tenía que interactuar, con el aumento de niveles de intencionalidad y el crecimiento del cerebro. Según este autor, los *Australopithecus* llegaban a dos grados de intencionalidad —como los chimpancés modernos—, el *Homo erectus* alcanzó tres grados, los humanos arcaicos cuatro grados y lo humanos modernos cinco grados en promedio. Dunbar cree que no se puede alcanzar ni cuatro ni cinco grados de intencionalidad sin lenguaje.

Todo esto ocurrió en los últimos 3 millones de años. Pero los aspectos semánticos del lenguaje deben haber evolucionado hace 500 000 años y la sintaxis probablemente hace unos 200 000, esto es aproximadamente 100 000 años antes de la salida de África de nuestros antepasados. Los lenguajes conformados por semántica, pero sin sintaxis son llamados protolenguajes por Bickerton (1990, 2009, 2014) y no son muy diferentes de la manera como habla un niño de aproximadamente 2 años de edad.

El punto es que la aparición del lenguaje plenamente desarrollado —dotado de semántica y sintaxis— permitió construir oraciones recursivas muy complejas para poder entender estados mentales de varios niveles de intencionalidad, tanto ajenos como propios (Dunbar, 2003, pp. 173-7). Esto potenció notablemente las habilidades de atribución psicológica, y complejizó aún más los procesos sociales de los homínidos, lo que aumentó todavía más sus habilidades de cognición social. Aquí se inició, por tanto, un vertiginoso incremento de la cognición social, el crecimiento del cerebro, la complejización de los procesos sociales de las comunidades homínidas y también de las características de la comunicación y del lenguaje[6]. Hasta aquí llega la propuesta de Dunbar.

Como ya vimos, todas las especies dotadas de sistema nervioso central suficientemente complejo tienen conciencia nuclear, es decir, experiencia fenoménica básica, como dolor, placer, sensaciones de hambre y sed, etcétera. Pero con los mamíferos

[6] Algunas buenas reconstrucciones de la evolución del lenguaje son: Deacon, 1997; Burling, 2005; Tallerman, 2005; Tecumseh Fitch, 2010; Everett, 2017.

aparecieron las emociones más básicas, precisamente porque estas especies necesitan vínculos afectivos fuertes para garantizar la supervivencia de las crías y establecer lazos sociales entre ellos. Sin embargo, la conciencia autobiográfica —también llamada autoconciencia, conciencia reflexiva o simplemente subjetividad— requiere de varios niveles de intencionalidad, de manera que debió haber evolucionado, en sus formas más elementales, con la aparición de tres o cuatro grados de intencionalidad, lo que coincide con el surgimiento del lenguaje por lo menos en el nivel semántico, es decir, con la aparición de las formas más básicas de protolenguaje.

En el nivel ontogenético ocurre algo parecido. Los niños comienzan a aprender a comunicarse con otras personas desde el momento mismo de su nacimiento, pero recién es hacia los nueve meses que establecen relaciones de atención compartida con la madre o el cuidador (Quintanilla, 2017b). Hacia el año y ocho meses aprenden las primeras palabras, y hacia los tres años ya tienen una semántica y una sintaxis rudimentarias. Entre los tres y cinco años adquieren por lo menos dos niveles de intencionalidad, lo que les permite no solo representarse el mundo sino también representarse las representaciones ajenas. Hacia los seis años alcanzan tres grados de intencionalidad; hacia los nueve años, cuatro grados; y hacia los once años, cinco grados (Henzi, Souza Pereira, Hawker-Bond, Stiller, Dunbar & Barrett, 2007), que es el promedio en nuestra especie. Con el aumento de los niveles de intencionalidad se potencian las habilidades pragmáticas en la comunicación, lo que no termina de madurar antes del inicio de la adolescencia.

Es obvio que los infantes humanos tienen conciencia nuclear desde antes de su nacimiento, pues sabemos que oyen y tienen sensaciones desde por lo menos las 35 semanas de gestación, cuando su sistema nervioso central está suficientemente maduro para procesar sensaciones. Pero la conciencia autobiográfica emerge muy lentamente y su desarrollo coincide con la aparición de los diversos niveles de intencionalidad y del lenguaje, hacia los dos años aproximadamente, cuando también pueden reconocerse en el espejo, lo que suele ser considerado un indicador de que el infante reconoce su cuerpo como suyo y, por tanto, ya tiene algún nivel de subjetividad. Pero como hay especies no humanas que parecen tener por lo menos dos grados de intencionalidad y se reconocen en el espejo —como los grandes simios y algunos otros mamíferos no primates—, no sería absurdo suponer que también podrían tener formas muy básicas de conciencia autobiográfica, la que, sin embargo, no sería potenciada por la ausencia de lenguaje[7].

[7] No hay duda de que algunas especies de animales tienen formas complejas de comunicación, pero eso no es suficiente para decir que tienen lenguaje. En esos casos se habla de comunicación no lingüística y se reserva la expresión «lenguaje» para un sistema que tenga semántica y sintaxis, y que permita la comunicación de contenidos proposicionales.

El punto por resaltar es que tanto en la evolución de la especie como en el desarrollo del niño hay una interacción constante entre el incremento de habilidades metacognitivas y de atribución psicológica —lo que incluye el aumento de niveles de intencionalidad—, y la adquisición del lenguaje. Pero no solo eso, la aparición del lenguaje en la especie y su adquisición en el niño potencian notablemente la posibilidad de hacer atribuciones finas y de dar razón de ellas, así como permiten hacer reconstrucciones racionales de procesos interpretativos no conscientes y no lingüísticos. Más aún, el lenguaje también tiene un rol en la posibilidad de hacer consciente lo no consciente y lo inconsciente[8]. Cuando todo esto está presente podemos hablar de una subjetividad madura en el niño.

Hasta ahí podemos ver importantes analogías entre lo filogenético y lo ontogenético (para un mayor desarrollo de este tema véase Quintanilla, Mantilla & Cépeda, 2014), pero en la línea temporal de la dimensión cultural también hay codesarrollo entre lo lingüístico y lo cognitivo. Hace aproximadamente 5500 años, los sumerios inventaron una forma de escritura silábica que reflejaba la fonología y la sintaxis de su lengua. Esto ocurrió en Mesopotamia, entre los ríos Tigris y Éufrates, en lo que actualmente es Irak. Este tipo de escritura, sin duda basada en formas más rudimentarias de registro, permitió que el ser humano pudiera almacenar fuera de su cerebro información valiosa para su supervivencia. También hizo posible que esa información se conservara y se transmitiera de generación en generación, lo que produjo una acumulación continua de conocimiento y cultura.

El desarrollo del lenguaje y la invención de la escritura facilitaron la aparición de conceptos abstractos, algo que en las lenguas tradicionales suele ocurrir a partir de conceptos concretos. Así, por ejemplo, del verbo griego *phyein*, que significa brotar, crecer, generar o producir, surgió el sustantivo *physis*, que significa lo que brota, lo que crece, lo que genera o lo que produce, y que fue traducido por los latinos como *natura* y pasó al castellano como «naturaleza». Lo que es muy importante para los fines de nuestra investigación es que, de esta manera, aparecieron los primeros conceptos psicológicos, por lo menos en la lengua griega. Así, por ejemplo, *thymós,* que originalmente significaba pecho comenzó a significar ánimo o voluntad, y se asoció con el lugar del alma en el cuerpo. *Fren,* que tenía el significado de diafragma, empezó a significar el lugar de las emociones. Las palabras *noos* y *nous,* que tenían el sentido de ver y observar, comenzaron a significar comprensión, inteligencia y conocimiento. *Psychein*, que significaba respirar, dio lugar al sustantivo *psyché,* que adoptó el significado de lo que respira, en primer lugar, y luego el principio de la vida,

[8] Recuérdese que «no consciente» alude a procesos cognitivos de los que no tenemos experiencia fenoménica e «inconsciente» a contenidos reprimidos en un sentido psicodinámico freudiano.

el movimiento y la sensación. Posteriormente esta palabra fue traducida al latín por *anima* y condujo al castellano «alma». En muchas lenguas tradicionales es usual que primero aparezcan verbos y luego estos sean sustantivados. Es el caso de *phyein*, que dio origen a *physis* y *psychein* que produjo *psyché*. La lengua quechua también es un ejemplo de ello. El quechua es, en general, una lengua de verbos y acciones, pero algunos de ellos se han sustantivado. Ese es el caso de *yachay,* el verbo para conocer o estar seguro, que dio origen al sustantivo *yachay,* conocimiento.

Naturalmente, mucho tiempo antes de la aparición de conceptos psicológicos, los seres humanos hacían atribuciones psicológicas —a los demás y a ellos mismos— y tenían autoconciencia. Pero, cuando estos conceptos aparecieron, los individuos pudieron reflexionar sobre ellos con más facilidad, es decir, pudieron tener estados mentales no solo acerca de sus estados mentales o de los de otros, sino acerca del concepto mismo de estado mental. Así, por ejemplo, es claro que, desde tiempos ancestrales, los griegos arcaicos podían ser invadidos por la ira o el enojo, para lo que tenían la palabra *menis,* que es el término que emplea Homero al inicio de la *Ilíada* para hablar de la cólera de Aquiles. Pero, con la progresiva aparición de los conceptos psicológicos, resultó posible relacionar la ira con otros estados mentales y los lugares del cuerpo donde esta se experimenta. Así también se desarrollaron concepciones sobre las maneras en que uno puede experimentar estos estados mentales, ya sea si son enviados por los dioses y, por tanto, nos toman por asalto, o si son causados por nuestra interpretación de la realidad circundante, o si, finalmente, son otras personas las que los inoculan en nosotros.

El punto, entonces, es que con la aparición de conceptos abstractos y, en este caso, con la aparición de conceptos psicológicos, fue posible concebir y experimentar estos conceptos —en uno mismo y en las atribuciones que se hacía a los demás— de maneras más conspicuas, lo que facilitó la aparición de momentos de introspección más profundas. Al operar con estos conceptos uno podía analizar los procesos sociales que estaban contenidos en ellos, y comprenderlos con mayor nitidez. Todos los conceptos son resúmenes miniaturizados de complejos procesos sociales, pero los conceptos abstractos lo son de procesos mucho más complejos aún. Considérese, por ejemplo, los conceptos «conocimiento», «mente» o «comprensión». La aparición de estos conceptos, el ponerles nombre y el ser capaz de escribirlos para transmitir información sobre ellos permitieron que sus usuarios pudieran formarse representaciones operativas de fenómenos extraordinariamente complicados, lo que no solo facilitó el intercambio social sino también hizo posible tematizarlos teóricamente. Esto dio lugar, posteriormente, a las primeras reflexiones sobre lo que hoy llamaríamos «psicología», como el diálogo *Fedón*, de Platón (2010), en el que este defiende una concepción inmortal del alma humana, y el tratado *De anima*, de Aristóteles (1983),

en el que se encuentra una teoría naturalista de los procesos psíquicos y de la conciencia humana. Es razonable suponer, aunque los detalles de esta tesis son debatibles, que todo esto permitió ampliar aún más la autoconciencia y la subjetividad.

A lo largo de los dos primeros capítulos he discutido algunas de las características principales de las ciencias, tanto naturales como humanas, cuando se proponen explicar la naturaleza y comprender a los agentes intencionales y, en lo concerniente al segundo caso, nos hemos concentrado en el fenómeno de la atribución psicológica. En el siguiente capítulo analizaré a la ciencia como práctica social al interior de comunidades epistémicas que conforman paradigmas. Esto nos conducirá, por tanto, a analizar la posición de Thomas Kuhn y su legado.

Capítulo tres
Explicar dentro de paradigmas y comunidades epistémicas

3.1. El concepto de paradigma

El proyecto epistemológico de Thomas Kuhn nos ha obligado a repensar el fenómeno de la comprensión al interior y fuera de la práctica científica, y tiene consecuencias importantes en torno a la pregunta sobre la traducción entre distintos lenguajes o esquemas conceptuales.

En efecto, desde aproximadamente la década de 1950, la filosofía de la ciencia sufre un giro copernicano que ha transformado muchos de sus principales presupuestos. Aquella disciplina enmarcada en el contexto de la justificación y poco interesada en su propia historia vuelve sus ojos hacia sus problemas y sus conceptos, viéndose a sí misma como un tipo de discurso en la pluralidad de discursos que constituyen la tradición.

Los escritos de Kuhn, especialmente *La estructura de las revoluciones científicas*, publicado originalmente en 1962 (1971)[1] y los textos posteriores a ese (1982), han ocupado un papel central en ese desplazamiento y han modificado notablemente muchas concepciones tradicionalmente asociadas a la investigación científica; nociones tales como «progreso», «acumulación» y «objetividad», entre otras. Kuhn es principalmente un teórico de la historia de la ciencia y su cometido es estudiar esa historia con la finalidad de esclarecer las relaciones genéticas y causales que guardan los momentos anteriores y posteriores de la investigación científica. Las contribuciones de Kuhn se han desarrollado básicamente en tres áreas: la sociología del conocimiento, la epistemología y la ontología. En este capítulo me ocuparé principalmente

[1] La primera edición en inglés es de 1962; la segunda, mejorada con Post Data, es de 1970. La traducción castellana, de Agustín Contín, es de la segunda edición y ha sido publicada por el Fondo de Cultura Económica en México, en 1971. Todas las citas corresponden a esta edición.

de las dos últimas, con lo cual dividiré este capítulo en tres partes: en las dos primeras haré una presentación general del pensamiento de Kuhn en sus puntos más polémicos y en última intentaré una evaluación de sus consecuencias ontológicas a partir de su análisis de las relaciones entre teoría y observación.

El aspecto ontológico es indudablemente aquel en el cual Kuhn se mueve con menos comodidad y sus afirmaciones son en ocasiones ambiguas y requieren de esclarecimiento, pero Kuhn ha tocado en sus trabajos una de las fibras más importantes de la filosofía contemporánea, aquella que trata acerca de nuestra teorización sobre el mundo y lo que pretendemos que el mundo realmente es. En este contexto, la pregunta por la existencia de una realidad anterior e independiente de nuestras teorías se vuelve especialmente pertinente. Interesantes presentaciones sobre este tema se pueden encontrar en Brown (1984) y Newton-Smith (1987).

Uno de los aportes centrales de Kuhn proviene de extraer las consecuencias de la tesis de la discontinuidad de la tradición científica en períodos inconmensurables entre sí. Antes de Kuhn había ya en la tradición filosófica francesa una larga polémica en torno a las nociones de «ruptura» o «corte» epistemológicos, polémica en la que participaron autores como Duhem (1962), Bachelard (1985) y Koyré (1979). Sin embargo, el mérito de Kuhn, quien reconoce la influencia de por lo menos Koyré (Kuhn, 1971, p. 10), es el haber desarrollado ampliamente estas consecuencias.

Kuhn publicó en 1957 (1979) un análisis sobre la revolución copernicana en que planteaba, aunque de manera muy seminal, muchas de las ideas que después habría de desarrollar en 1962 en *La estructura de las revoluciones científicas*. El aporte más importante que llegó a configurarse en este último texto gira en torno del concepto de paradigma. En la primera edición de 1962 definió «paradigma» como un conjunto de «realizaciones científicas universalmente reconocidas que, durante cierto tiempo, proporcionan modelos de problemas y soluciones a una comunidad científica» (Kuhn, 1971, p. 12).

Mastermann (1970) encontró, sin embargo, que en *La estructura de las revoluciones científicas* había veintidós sentidos diferentes del término, lo que condujo a Kuhn a precisar este concepto principalmente en dos lugares: los «Post Data» a la edición de 1970 y en «Algo más sobre paradigmas» (1982). En esos textos Kuhn distingue dos sentidos principales de la noción de paradigma. En un primer sentido, un paradigma es un ejemplo concreto de realización científica de solución de problemas. Este es el sentido original, porque fue el que condujo al autor a elegir el término y hacerlo extensivo a los otros conceptos vinculados (1971, p. 286). En otro sentido, que es el que más nos interesa aquí, un paradigma es un conjunto de creencias, valores y técnicas que comparten los miembros de una comunidad científica en un momento de su desarrollo. A este sentido Kuhn lo denominó «matriz disciplinar»

y alude al conjunto de presupuestos compartidos en torno a la determinación del objeto de estudio, el método que será empleado en la investigación, la pertinencia de los problemas que dirigirá y las formas que servirán para organizar y seleccionar los datos observacionales. Pero, sobre todo, a la definición de la noción misma de explicación y a valores que permitirán que los científicos determinen cuando una teoría es más explicativa que otra.

El programa central de Kuhn es mostrar que la observación y selección de datos fácticos depende de un paradigma y es imposible fuera de él, de manera que toda observación está siempre cargada de presupuestos teóricos compartidos. Este postulado resultaba especialmente provocador —aunque no del todo novedoso, pues ya había sido discutido entre los positivistas lógicos— porque ponía en cuestión la tesis empirista según la cual la observación es el terreno neutro y estable que permite la elección objetiva entre teorías. Si la observación está cargada teóricamente, la corroboración de teorías se convierte en un círculo vicioso. La tesis de Kuhn es que, en efecto, es imposible corroborar una teoría si no es desde los presupuestos de un paradigma, el cual no puede ser justificado racionalmente, pues eso supondría una estructura estable y anterior al paradigma que sea el criterio objetivo de elección racional. El paradigma no puede ser justificado racionalmente según los criterios de racionalidad imperantes entre los positivistas, lo que conduce a Kuhn a sugerir que lejos de considerar a la actividad científica como irracional, lo que debe cambiar es nuestra concepción misma de racionalidad.

Ante la crisis de fundamentación de la racionalidad autónoma moderna, la filosofía de la ciencia de las primeras décadas del siglo XX volvió sus ojos a ese «suelo neutral de observación» como el único criterio que podía preservar la racionalidad en la elección y justificación de teorías. Así pues, según Kuhn, sería una ilusión toda pretensión de justificación o fundamentación extraparadigmática, de manera que sería imposible elegir, según esos criterios de racionalidad, entre teorías pertenecientes a paradigmas diferentes, tales como, por ejemplo, la teoría aristotélica de la caída de los cuerpos y la teoría gravitacional de Newton, pues pretender elegir entre ellas supondría erróneamente que ambas comparten el mismo concepto de explicación. De acuerdo con Kuhn, lo primero que cambia cuando cambia un paradigma es la noción misma de explicación. Por eso, según él, solo tiene sentido comparar, en el contexto de justificación, teorías que pertenecen al mismo paradigma. Se justifica una teoría científica por su consistencia y efectividad intraparadigmática, pero no es posible justificar un paradigma, como tampoco escoger entre dos paradigmas en conflicto. Esta es la posición inicial y básica de Kuhn, que después irá matizando y precisando progresivamente.

Como ya mencioné, en los «Post Data» y en «Algo más sobre paradigmas» (Kuhn, 1971), precisa el sentido de paradigma como matriz disciplinar y lo hace afirmando que incluye los siguientes contenidos. En primer lugar, se trata de «generalizaciones simbólicas compartidas». Estas son presupuestos generales que tienen en común los miembros de una comunidad científica en torno a la determinación del objeto de estudio y su método. Estas generalizaciones incluyen también presupuestos en torno a los significados de los términos teóricos, con lo que debemos entender el abandono de las generalizaciones simbólicas, que marca el inicio de una revolución científica como variación de los significados de los términos teóricos. Las propiedades naturales de los objetos de observación y los rasgos semánticos de los términos teóricos que los designan están íntimamente entrelazados, y en ocasiones es imposible discernir cuándo una teoría está describiendo una propiedad de los objetos y cuándo simplemente está proponiendo un significado diferente para las expresiones que refieren a ellos. Así, por ejemplo, cuando Einstein afirmó el carácter relativo de la simultaneidad, ¿alteró una propiedad de la simultaneidad o varió el significado de ese concepto? Mientras para un físico newtoniano el enunciado «la simultaneidad es relativa» es simplemente una contradicción semántica, para Einstein es un enunciado analítico que se sigue de los significados que en su teoría se atribuye a los conceptos de simultaneidad y relatividad (Kuhn, 1971, p. 282). Algunos autores (Newton-Smith, 1987) han señalado la similitud entre las generalizaciones simbólicas compartidas y lo que Lakatos denomina el «núcleo central» de un programa de investigación, es decir, los presupuestos que la comunidad científica alteraría solo en última instancia ante la presencia de anomalías.

Un segundo aspecto de la matriz disciplinar tiene que ver con la noción de «modelo». Se trata de compromisos compartidos con creencias específicas (1971, p. 282) que cumplen funciones heurísticas y ontológicas al asignar a la comunidad problemas pertinentes y lo que será considerado una buena solución para estos.

Una tercera faceta de la matriz disciplinar está dada por los «valores», también denominados por algunos autores «criterios epistémicos» o «virtudes epistémicas». Estos son aquellos presupuestos que permiten la elección de teorías, como la exactitud, el hecho de que las predicciones cuantitativas sean preferibles a las cualitativas, la coherencia interna, la consistencia de las teorías con otras teorías usualmente aceptadas, etcétera.

3.2. Variación de significado e inconmensurabilidad

Cuando hay una comunidad estable de científicos que comparte un paradigma se produce un período de ciencia normal. En este el acuerdo es mayor que el desacuerdo en torno a la pertinencia de los problemas y a lo que significa resolver adecuadamente

un enigma. No obstante, en ocasiones la tensión entre los presupuestos de la matriz disciplinar y los datos de observación producen anomalías en las predicciones. Estas no constituyen contraejemplos a las teorías, son simplemente casos rebeldes que aún no han sido explicados. Sin embargo, depende de la comunidad científica y, por tanto, es convencional el asignar a un fenómeno inexplicado el carácter de anomalía o contraejemplo. Cuando las anomalías son suficientes en número como para cuestionar las teorías y los presupuestos que a ellas subyacen, se produce un período de crisis donde el acuerdo disminuye notablemente y el discurso científico se torna anormal, lo que conduce a una revolución científica. De esta manera, el nuevo paradigma que se constituye produce un período de ciencia normal inconmensurable con el anterior.

La inconmensurabilidad tiene básicamente dos orígenes: en primer lugar, está lo que se suele denominar «variación de valor» (Kordig, 1971; Newton-Smith, 1987, pp. 166 y ss.). Al cambiar el contenido valorativo de la matriz disciplinar, es decir aquellas que se consideran virtudes explicativas de las teorías, cambia también el concepto mismo de explicación científica. Mientras, por ejemplo, para Aristóteles explicar un fenómeno significa ubicarlo en una cadena causal gobernada por esencias inmutables, para el paradigma moderno instaurado por Galileo y continuado por Newton explicar algo significa encontrar, en el curso de la naturaleza, una regularidad en la cual subsumir el fenómeno en cuestión. Esa regularidad será descrita mediante una ley universal y, así, explicar el fenómeno será mostrar la correspondencia entre su comportamiento y la descripción nomológica. Ahora bien, como es imposible justificar de manera concluyente un juicio de valor explicativo, si se contraponen dos concepciones diferentes —no excluyentes— de explicación, no sería posible elegir de manera conclusiva una.

Aunque la posición inicial de Kuhn era que la inconmensurabilidad de valores explicativos tiene como consecuencia la imposibilidad de elección racional, en «Algo más sobre paradigmas» ha sugerido que hay criterios extraparadigmáticos que pueden colaborar con la elección racional y suavizar la tesis de la inconmensurabilidad. Estos criterios, que se confunden con los valores mencionados en los «Post Data», son los siguientes: precisión, consistencia, amplio alcance, simplicidad y fertilidad. Kuhn reconoce, sin embargo, que la manera como se entienden estos criterios es materia de discusión, con lo cual deja abierta la posibilidad de que, eventualmente, varíen entre paradigmas. Esto conduce a que posteriormente Kuhn abandone este primer origen de la inconmensurabilidad y acepte la posibilidad, por lo menos en el nivel de valores, de una parcial conmensurabilidad entre paradigmas. Kuhn termina, sin embargo, incurriendo siempre en una posición convencionalista, porque los criterios extraparadigmáticos de elección de teorías siguen siendo un producto de la tradición.

El segundo origen de la inconmensurabilidad, que ha tenido una importancia decisiva en la historiografía de la ciencia, es lo que se denomina «variación radical de significado» y se encuentra, aunque de una manera muy elemental, ya en Campbell (1957). Vamos a detenernos en una breve explicación histórica. Los positivistas lógicos solían distinguir entre los términos observacionales y los términos teóricos. Los primeros son aquellos que refieren a objetos espacio-temporales, empíricamente observables, tales como «roca», «estrella» o «célula». Estos términos obtienen su significado extensionalmente por medio de la experiencia. Así, conocer el significado de un término observacional sería saber qué tipos de experiencia supondrían casos de aplicación del concepto observacional en cuestión. Los enunciados observacionales, cuyos términos son exclusivamente observacionales, se contrastan directamente con la experiencia y son significativos en la medida que se conozca su método de verificación. Por el contrario, los términos teóricos tales como «gravedad», «inconsciente», «clase social» o «adverbio», son aquellos que no refieren a objetos espacio-temporales y, por tanto, su significado no se fija extensionalmente sino en conexión con términos observacionales. Los enunciados teóricos, cuyos términos son únicamente teóricos, o teóricos y observacionales, solo se hacen significativos en el contexto de una teoría determinada. Un término teórico que no pueda ser definido en relación con un conjunto de observaciones carecería de significado, con lo cual los términos teóricos resultarían siendo entidades inferidas a partir de la observación. Para Russell (1957) estas entidades inferidas se expresarían, en un lenguaje depurado, en términos de construcciones lógicas. Algunos positivistas postularon que el significado de los términos teóricos se encuentra contenido en los postulados de la teoría, los cuales conectan el significado de los términos teóricos con la observación. En este nivel, resulta difícil distinguir entre los postulados de significado de los términos teóricos y simples enunciados teóricos, de manera que estos se convertirían en enunciados analíticos. Como es claro, cuando se produce un cambio en la teoría, cambian los postulados de significado y, en consecuencia, cambian los términos teóricos mismos. Una de las constataciones más antiguas de esta situación se debe al propio Carnap. En un primer momento (1953, pp. 47-92) este autor intentó definir los términos teóricos como disposicionales, los cuales, a su vez, se definen por referencia a términos observacionales. Posteriormente (1956, pp. 38-76) abandonó esa posición.

Para los positivistas lógicos, Carnap incluido, era vital mantener la distinción, de suerte que, aunque se admitiese variación de significado en los términos teóricos, el significado de los términos observacionales se mantendría estable y constituiría el criterio de elección y conmensuración entre teorías. Los positivistas lógicos pretendieron que había un conjunto de términos observacionales básicos que, constituyendo un lenguaje observacional, no presupondría teoría alguna. Algunos de ellos

se dedicaron a la búsqueda de ese lenguaje y denominaron «proposiciones protocolares» a sus proposiciones básicas. Estas constituirían el cimiento de las teorías y serían ese suelo de la observación al cual aludía Herbert Feigl. Las teorías serían excluyentes y, al serlo, serían conmensurables no cuando sus enunciados teóricos se muestren diferentes, pues eso sería simple desacuerdo, sino cuando sus enunciados observacionales resulten contradictorios, pues supuestamente las observaciones deberían ser comunes entre todas las teorías en disputa.

Uno de los primeros en cuestionar la distinción teoría/observación fue Otto Neurath en su famoso artículo «Proposiciones protocolares» (1932-1933), en el que comentaba y criticaba el artículo de Carnap, «El lenguaje físico como lenguaje universal de la ciencia» (1931). En ese texto, Carnap pretendía demostrar que los enunciados protocolares —u observacionales— no requieren de confirmación, pues son los enunciados que conectan las teorías con el mundo. Para Carnap lo único que debemos modificar en nuestra investigación científica son los enunciados teóricos. Neurath propuso, por el contrario, que cualquier enunciado de una teoría puede ser eliminado ya que todos poseen el mismo estatuto epistemológico, con lo cual, en última instancia, la verdad de un enunciado sería su coherencia con la totalidad o con el conjunto consistente máximo de enunciados de un sistema. Como veremos en el próximo capítulo, esta tesis influiría mucho en el holismo semántico y epistémico que Quine postuló más adelante.

Popper también rechazó la noción de enunciado observacional, aunque introdujo, en cambio, la noción de enunciado básico, al cual denominó «falsador potencial» y cuya función es corroborar las teorías (1983, p. 388). La diferencia entre los enunciados protocolares y los enunciados básicos de Popper es que la aceptación de estos últimos depende de una convención en la comunidad científica y, por tanto, su significado no es inmutable (1982).

La propuesta de Kuhn es que cuando cambia el significado de los términos teóricos cambia también el de los términos observacionales, con lo cual no hay cómo elegir entre teorías de paradigmas rivales si se recurre a la experiencia. El significado de los términos, tanto teóricos como observacionales, está determinado por el lugar que estos ocupan en la teoría. Cuando cambia la teoría cambian los significados de los términos, de manera que los enunciados de dos teorías en disputa se tornan intraducibles entre sí y los enunciados cuyas formulaciones eran aparentemente excluyentes resultan siendo compatibles si se muestra que los significados de los términos son diferentes. Así, por ejemplo, el enunciado teórico «la masa es invariable» tendrá interpretaciones diferentes según se formule en un contexto newtoniano o einsteiniano. Para la mecánica newtoniana, la masa es una constante que se define como la cantidad de materia producto de la densidad por el volumen. Para Einstein, por el contrario,

la masa es una variable que está en relación con la velocidad, en la que su mínimo valor es el mismo de la masa newtoniana, mientras que su máximo valor aumenta infinitamente mientras el móvil se acerca a la velocidad de la luz. Así pues, el enunciado newtoniano «La masa es invariable» y su contraparte einsteiniano «La masa es variable» no son contradictorios, porque tienen significados diferentes.

Israel Scheffler (1967) ha intentado eliminar el problema de la inconmensurabilidad aduciendo que las revoluciones cambian el significado de los términos, pero no alteran los objetos a los cuales ellos refieren. Su argumento, que en ocasiones es empleado también por otros autores, es por lo menos discutible, porque cuando cambia el significado cambia también la extensión del concepto, ya que es imposible discriminar a qué refiere un término si no es escudriñado su uso. Una grave limitación de la tesis de Kuhn es que no aclara cuál es el grado preciso de variación de la teoría que origina variación de significado. De otro lado, si el significado de un término es su uso, lo cual evidentemente implica diversidad y pluralidad, ¿cuándo diremos que un término ha sufrido un cambio importante de significado como para producir inconmensurabilidad? Kuhn no ofrece ninguna respuesta a estas interrogantes.

Por eso, una buena crítica a esta posición es la de Shapere, quien lo expresa en estos términos:

> Este relativismo y las doctrinas que surgen en él, no son resultado de una investigación de la ciencia verdadera y de su historia; más bien, son consecuencia puramente lógica de una preconcepción estrecha acerca de lo que es el «significado (1981, p. 109).

Lo primero que Kuhn preguntaría es cuál es el significado de la expresión «ciencia verdadera». Es cierto que Kuhn no incluye una teoría del significado ni del cambio semántico, lo cual es una grave limitación, pero no sería difícil aplicar alguna de las teorías ya existentes a su modelo. Por ejemplo, si decimos que el significado de un término está fijado por el conjunto de enunciados que lo contienen aceptado por un hablante competente, es posible decir que hay variación de significado cuando la clase de enunciados aceptables se altera de un modo considerable. Es necesario poder determinar qué cuenta como una alteración considerable, pero finalmente ese es un problema de convención y decisión terminológica. Así como hay variación de significado en conceptos internos a las teorías, tales como «simultaneidad», «selección natural» o «inconsciente», también la hay en conceptos pertenecientes a la filosofía de la ciencia —o a la filosofía en general— como «explicación», «racionalidad», «interpretación», «significado», etcétera, lo que naturalmente también conduciría a un margen de intraducibilidad que solo se puede afrontar al comparar diacrónicamente los matices de los conceptos.

Una consecuencia importante de la tesis de la inconmensurabilidad es la imposibilidad de encontrar una línea de progreso a lo largo de la historia de la ciencia. En tanto los criterios valorativos son intraparadigmáticos, no hay manera de evaluar distintos conjuntos de criterios valorativos. En sus últimos trabajos Kuhn ha suavizado su posición, pero la fuerza del argumento se mantiene independientemente de su grado de radicalidad. En los «Post Data» ha afirmado que el único sentido que él considera legítimo de progreso científico es aquel según el cual las teorías científicas posteriores son mejores que las anteriores en la medida en que resuelven enigmas presentes que son distintos de los del pasado. Solo en este sentido, si es que es un sentido no trivial, se puede hablar de progreso en la ciencia. En todo caso, más allá de Kuhn, sí cabe hablar de progreso entre paradigmas, en el sentido de que un paradigma posterior supone a los anteriores y de esa manera los contiene. En otras palabras, hay progreso entre las teorías de paradigmas posteriores respecto de las anteriores si aquellas pueden resolver los problemas que estas podían resolver, además de los nuevos problemas producto de la nueva evidencia e investigación. Evidentemente, las revoluciones científicas no implican rupturas radicales en las que se prescinde del pasado.

Normalmente la metodología posterior supone y absorbe a las anteriores o, como en el caso del desplazamiento de la explicación causal aristotélica a la explicación moderna en virtud de causas eficientes, implica la profundización de un aspecto de la metodología anterior. Comparamos ambas teorías sin salir de nuestro paradigma. Analizamos retrospectivamente la doctrina aristotélica sobre la caída de los cuerpos como contenida en nuestra propia forma de ver el mundo, la cual reserva un lugar importante para Aristóteles. Es en ese sentido diacrónico que es posible la conmensurabilidad. Probablemente el propio Kuhn no tendría inconvenientes en aceptar esta noción comprehensiva de progreso, siempre que no se postule que en un paradigma posterior hay una descripción más exacta de la realidad o un mayor acercamiento a la verdad de la naturaleza, ya que esto es algo que rechaza tajantemente:

> Quizá haya alguna manera de salvar la idea de «verdad» para su aplicación a teorías completas, pero esta no funcionará. Creo yo que no hay un medio independiente de teorías, para reconstruir frases como «realmente está allí»; la idea de una unión de la ontología de una teoría y su correspondiente «verdadero» en la naturaleza me parece ahora en principio una ilusión; además, como historiador, estoy impresionado por lo improbable de tal opinión. Por ejemplo, no dudo de que la mecánica de Newton es una mejora sobre la de Aristóteles, y que la de Einstein es una mejora sobre la de Newton como instrumento para resolver enigmas. Pero en su sucesión no puedo ver una dirección coherente de desarrollo ontológico. Por lo contrario, en algunos aspectos importantes, aunque, desde luego, no en todos, la teoría general de la relatividad, de Einstein, está más cerca de la de Aristóteles que ninguna de las dos de la de Newton (1971, p. 314).

Ya que, según este autor, no hay un conjunto dado y estable de hechos que sea abordado por las distintas teorías de los paradigmas en disputa, son los propios paradigmas los que configuran los hechos al seleccionar y determinar lo que será considerado un dato observacional pertinente. En este nivel, el mayor mérito de Kuhn, que puede retrotraerse hasta otros autores, es el haber llamado la atención sobre la imposibilidad de comparar directamente las teorías con los hechos y haber subrayado el carácter teórico de todo proceso de observación.

3.3. Teorías y realidad

En el capítulo X de *La Estructura de las revoluciones científicas* Kuhn arriesga, como una consecuencia ontológica de sus planteamientos epistemológicos, una tesis poco desarrollada y explotada en el libro, pero, sin lugar a dudas, plenamente enmarcada en la discusión ontológica contemporánea. Afirma que:

> [L]os cambios de paradigmas hacen que los científicos vean el mundo de investigación que les es propio de una manera diferente. En la medida en que su único acceso para ese mundo se lleva a cabo a través de lo que ven y hacen, podemos desear decir que, después de una revolución, los científicos responden a un mundo diferente (1971, p. 176).

En este párrafo se evidencia una ambigüedad que atraviesa la obra de Kuhn. De un lado, afirma que los científicos ven el mismo y único mundo de investigación de maneras diferentes; de otro lado, sostiene que la revolución da lugar a un mundo distinto. Uno podría suponer que lo que quiere decir es que si bien el mundo, en tanto conjunto de entidades espacio-temporales y de regularidades que gobiernan dichas entidades, es permanente, lo que cambia es el tipo de explicación que los científicos emplean, con lo que cambia la interpretación que los científicos tienen de la realidad. Kuhn rechaza, sin embargo, esta posibilidad (1971, p. 190) y afirma que «lo que sucede durante una revolución científica no puede reducirse completamente a una reinterpretación de datos individuales y estables» (p. 291). En la misma página insiste con una afirmación críptica: «aunque el mundo no cambia con un cambio de paradigma, el científico después trabaja en un mundo diferente» (p. 291). Lo que podemos recoger hasta aquí es que después de una revolución científica, en un sentido el mundo sí cambia y en otro sentido no. Lo que hay que aclarar es cuáles son esos dos sentidos y si la afirmación se justifica. En los «Post Data» dice que «… dos grupos cuyos miembros tienen sensaciones sistemáticamente distintas al recibir los mismos estímulos *en cierto sentido* viven en diferentes mundos» (p. 295). Lo que podría entenderse de esta afirmación, y que es consistente

con la crítica kuhniana a la distinción entre enunciados teóricos y observacionales, es que toda afirmación que pretenda corresponder a un dato observacional depende de la aceptación o no de presupuestos teóricos, lo que incluye la comprensión que tenemos del significado de los términos. Si es así, lo que hemos de considerar el mundo de nuestra observación varía en la medida en que cambian los presupuestos teóricos con los que lo abordamos. Evidentemente existe un mundo independiente y anterior a nuestro hablar de él, pero es igualmente evidente que solo tenemos acceso a este tal como es descrito por nuestras teorías, no en el sentido de que estas sean una instancia intermediaria entre nosotros y el mundo, sino más bien en tanto toda percepción de la realidad involucra inevitablemente una interpretación de ella. No es posible comparar las creencias directamente con el mundo, las creencias solo se comparan con otras creencias que han sido producidas causalmente por el mundo.

La posición de Kuhn es ambigua en torno a si:

(1) Existe una realidad interpretada múltiplemente, por lo que es imposible distinguir cuál de estas representaciones científicas es la que se adecua más a ella.

(2) Los diversos paradigmas dan lugar a diversas percepciones y, en tanto la realidad es inseparable de nuestras interpretaciones, se trata de realidades múltiples.

Kuhn acepta que la experiencia que tenemos de un objeto puede variar por la sola presencia de ciertos presupuestos. Así, por ejemplo, dice que «cuando Aristóteles y Galileo miraron piedras oscilantes, el primero vio una caída forzada y el segundo un péndulo» (1971, p. 191). Barnes (1986, p. 140) sugiere que, si por «mundo» entendemos la experiencia ordenada por un grupo de hombres, es correcto lo que Kuhn dice, pero si «mundo» es el medio físico «tal como es» anterior a su percepción y expresión, lo que Kuhn dice es falso. Sin embargo, precisamente lo que este último rechazaría es la pertinencia e inteligibilidad de hablar de un mundo «tal como realmente es» (1971, p. 314).

Según esto, no es solo que los datos sensibles se interpretan de varias maneras, es también que la constitución misma de los datos sensibles varía entre comunidades y paradigmas. Afirma Kuhn:

> Concluyo que, aunque los datos son los elementos comunes mínimos de nuestra experiencia individual, tienen que ser también respuestas compartidas a un estímulo dado, solo entre los miembros de una comunidad educativa, científica o lingüística relativamente homogénea (p. 333).
>
> A los miembros de comunidades diferentes se les presentan datos diferentes mediante los mismos estímulos (n. 18).

Así pues, según Kuhn un cambio de paradigma es como un cambio gestáltico: el objeto que es observado —en tanto fuente de estímulos sensoriales— es el mismo, sin embargo, la experiencia que tenemos de él puede variar. Es como ver una figura como un pato o como un conejo, según la célebre imagen de Wittgenstein (1988).

Cuando Tolomeo y Galileo miraban el sol, el segundo veía un cuerpo estático en torno del cual gira la Tierra y el primero un cuerpo en movimiento alrededor de la tierra. El objeto de percepción de ambos es el mismo en el sentido en que es la misma fuente de datos sensoriales. Sin embargo, el objeto que cada uno de ellos veía, en tanto se convierte en un objeto de discurso significativo por el lugar que ocupa en su sistema de creencias, era un objeto diferente. Como atinadamente anota Brown, una cosa es «ver» y otra «ver algo como algo» (1984, p. 111). Tolomeo y Galileo veían el mismo objeto, pero lo «veían como» dos fenómenos diferentes. Nuestra ventaja histórica nos permite saber que la hipótesis de Galileo tuvo mayor vigencia que la de Tolomeo, pero eso no lo hubiera podido saber un hipotético observador neutral de la época.

Cuando se ve un objeto se tiene simplemente una imagen retiniana, cuando «se ve como» «el objeto es identificado y, por tanto, la percepción está teóricamente cargada» (Brown, 1984, p. 111). Pero, como describir un objeto es definirlo, delimitarlo y categorizarlo, es imposible describir un objeto si no es cargándolo teóricamente. La única forma de describir un objeto preteóricamente sería como un conjunto asociado de datos sensibles; no obstante, como es imposible definir con claridad los límites de estos datos sensibles, no sería propiamente un objeto aquello a lo que estaríamos refiriendo, sino un mero caos de sensaciones, para aludir a Kant. Sin embargo, hablar de un caos de sensaciones es ya elaborar una descripción, lo que conduce a la paradoja de suponer que podemos describir una supuesta realidad previa a nuestras descripciones de ella. Por ello el simple ver es ininteligible, solo el «ver como» es significativo.

Brown ha llegado al extremo de postular que lo que observamos son significados (1984, pp. 114-117) y que los datos sensibles «suponiendo que existan tales cosas y que seamos conscientes de ellos, no pueden ser los objetos primarios de nuestro conocimiento […] El científico no registra todo lo que observa, sino más bien solo aquellas cosas que las teorías que observa indican que son significativas» (p. 114). Algunos autores han creído encontrar matices kantianos en esta posición. La diferencia pertinente con Kant sería que para este la estructura categorial es universal y necesaria, mientras que en el caso de los autores a quienes comentamos serían los esquemas conceptuales o sistemas de creencias, que no son ni universales ni necesarios, los que participan en la síntesis.

Esta postura está muy cerca del holismo y relativismo ontológicos defendidos por Quine (1969), para quien atribuimos existencia a una entidad en la medida en que pueda canjearse por una variable ligada en una teoría. Las creencias acerca de la existencia de entidades, así como aquellas acerca del mundo externo, «afrontan el tribunal de la experiencia sensible no individualmente sino como un todo integrado» (1963, p. 41).

Más allá de Kuhn, en un sentido, el mundo físico es un conjunto de objetos espacio-temporales y las leyes que los gobiernan. Sin embargo, nuestra categorización de lo espacio-temporal depende, en alguna medida, de la manera en que nuestras teorías, nuestros sistemas de creencias y nuestras lenguas clasifican sus objetos de discurso. Es materia de discusión entre lingüistas y filósofos si las lenguas tienen compromisos ontológicos (Rivarola, 1991) y aunque durante algún tiempo la hipótesis del relativismo lingüístico de Sapir (1921) y Whorf (1956) se consideró abandonada, esta ha gozado de cierto renacimiento en lo que se suele denominar las «teorías neowhorfianas». Estas suelen ser versiones débiles del whorfismo clásico. La versión tradicional fuerte del whorfismo afirma que la lengua de una comunidad —sus categorías lingüísticas— determina la percepción y las posibilidades cognitivas de los miembros de esa comunidad, es decir, moldea y limita lo que estos pueden percibir, pensar y concebir, de manera que, en principio, podría impedir que los miembros de una comunidad lingüística dada pudieran percibir ciertos objetos o comprender determinados conceptos. Otra manera de formular esta forma de relativismo conceptual es señalando que postula que lo real es relativo a un esquema conceptual y que los diversos esquemas conceptuales son o pueden ser intraducibles e incomparables entre sí. Esta versión llegaría al extremo de sostener que los hablantes de diferentes lenguas «viven» en mundos diferentes, que es lo que dice Kuhn respecto de quienes «viven» en paradigmas y teorías científicas diferentes. Así de radical, esa tesis tiene poca evidencia en su favor y mucha en contra. Pero la versión débil es más aceptable. Esta sostiene que las categorías de una lengua influyen —en mayor o menor medida, es algo aún por determinar empíricamente— en la cosmovisión de la comunidad que la usa, lo que favorece ciertas posibilidades cognitivas y no otras. Esto implicaría que los hablantes de esa lengua tendrían cierta tendencia a categorizar el mundo y razonar de cierta manera y no de otra. La versión débil acepta, sin embargo, que los hablantes de una lengua pueden concebir la realidad de una manera diferente a la favorecida por su propia lengua si son expuestos a otra visión del mundo (Boroditsky, 2001, 2003), de manera que no está presente el determinismo lingüístico de la versión más radical. La diferencia entre ambos tipos de whorfismo, débil y fuerte, radica en el grado en que la lengua afecta nuestra cosmovisión, modelando la percepción

y dificultando la posibilidad de comprender visiones del mundo «constituidas» por otras lenguas. Mientras más fuerte sea la posición implicará mayor determinismo.

Las posiciones neowhorfianas, a su vez, pueden ser aplicadas a la realidad objetiva exterior al hablante o a su propia vida mental. En el primer caso, la lengua influiría en nuestra categorización, percepción y cognición acerca del mundo. En el segundo caso, influiría en nuestra categorización, experimentación y conciencia de nuestros propios estados mentales.

Es ampliamente plausible que las lenguas influyan en nuestra cosmovisión —en la línea del neowhorfismo débil— y que tengan un rol en la manera como categorizamos, describimos y experimentamos los objetos del mundo externo, así como también nuestros propios estados mentales. Pero es mucho menos probable que también afecten nuestra percepción sensorial, permitiéndonos percibir ciertos objetos —o rasgos de objetos— y no otros. Sin embargo, nadie discutiría que el lenguaje, en tanto sistema semántico y sintáctico, influye en nuestra categorización y concepción del mundo objetivo y subjetivo. En ese punto, el debate radica en si el lenguaje está constitutivamente involucrado en el pensamiento, es decir, si puede haber pensamiento sin lenguaje o no. Si por pensamiento entendemos la pura experiencia fenoménica, es obvio que esta se puede dar sin lenguaje, pero esta sería una concepción tan general de pensamiento que resultaría trivial. Si, más apropiadamente, por pensamiento entendemos la capacidad de concebir proposiciones y tener actitudes proposicionales, esto es, la posibilidad de predicar conscientemente de objetos, hay dos posiciones en disputa. Algunos autores como Locke (1982) y buena parte de los filósofos modernos aceptarían que el pensamiento no requiere de lenguaje, pues este solo tiene el rol de transmitirlo y hacerlo público. Otros, como Fodor (1987, 1990, 1994), sostendrían que las lenguas adquiridas no cumplen ese rol, pero sí el lenguaje interno, denominado «mentalés» o *the language of thought.* Finalmente, varios otros autores como Wittgenstein (1988), Davidson (1980b, 1984c) o Carruthers (2006) afirmarían que la vida en comunidad y la capacidad para ser intérprete de un lenguaje son necesarias para poder pensar proposicionalmente. Mi posición es que, aunque el pensamiento proposicional requiere de comunicación intersubjetiva sistemática, es decir de alguna forma de lenguaje, en principio este no tiene que ser una lengua natural. Sin embargo, el pensamiento proposicional complejo que involucra conceptos abstractos sí requiere de ser hablante de una lengua natural (Quintanilla, 2000)[2]. Además, tanto en términos filogenéticos como ontogenéticos, el lenguaje parece ser el elemento que detona el pensamiento proposicional complejo.

[2] En Quintanilla, 2018, sostengo que, a diferencia de la manera como suele interpretarse a Aristóteles, este autor pensaba que el lenguaje es condición necesaria para el pensamiento.

Pero volvamos brevemente al whorfismo. La idea de que las lenguas condicionan nuestra visión del mundo se remonta hasta por lo menos Humboldt y el romanticismo alemán de comienzos del siglo XIX. Esta intuición, a su vez, provenía de desplazar hacia las lenguas naturales algo que Kant había afirmado acerca del entendimiento en general. Kant sostenía que podemos conocer la realidad porque la constituimos en una síntesis formada a partir de los datos de la sensibilidad y las categorías ordenadoras del entendimiento, que es universal a todos los seres humanos. Kant no podía explicar por qué existen categorías universales del entendimiento, pero sí pretendía mostrar cuáles y cuántas son, cómo operan y también que son las mismas para todos. Si hoy quisiéramos abordar esa problemática tendríamos que preguntarnos acerca de la existencia de categorías estructuradas neurológicamente y que podrían ser universales, aunque no necesarias sino contingentes, porque serían el producto de la evolución del cerebro del *Homo sapiens*. Pero, naturalmente, para formular las cosas de esa manera fue necesario esperar hasta 1859, cuando Darwin publicó *El origen de las especies* (2009).

Ahora bien, algunos críticos de Kant sostuvieron que no hay razón para suponer que existen tales categorías universales, de manera que los datos sensibles de la realidad serían categorizados por las lenguas. Así es como apareció la hipótesis de que las distintas lenguas categorizan de diferente manera la realidad y que, en un importante sentido, uno vive en un mundo estructurado por su lengua, es decir, la hipótesis Sapir-Whorf, en sus versiones fuerte y débil.

Pero recientemente ha surgido otra opción, propuesta por Guy Deutscher (2010, p. 159) y bautizada por él como «principio Boas-Jakobson». Según Deutscher, la diferencia entre lenguas no radica en lo que ellas permiten a los hablantes expresar, sino en lo que ellas los obligan a expresar. La idea es que algunas lenguas compelen a los usuarios a explicitar ciertas relaciones o contenidos que otras lenguas no obligan. Pero ninguna lengua impide que uno exprese un contenido que otra lengua permite expresar. De manera más clara aún: algunas lenguas exigen a sus usuarios a pensar en contenidos que otras lenguas no exigen pensar. Puede verse este principio, por tanto, como una inversión de la hipótesis del relativismo lingüístico de Sapir-Whorf.

Las lenguas nativas del Perú facilitan la ejemplificación de esta idea. Como es conocido, tanto el quechua como el aimara emplean marcadores de evidencialidad. En el caso del quechua, por ejemplo, hay dos sufijos de evidencialidad, *-mi*, que es de validación, y *-si*, que es reportativo. Uno añade *mi*, al verbo para expresar convicción o certeza y *si* para expresar algo de lo que hemos sido informados y que, por tanto, no nos consta. En el primer caso, hay responsabilidad del hablante, en el segundo no. Si uno no usa marcadores de evidencialidad en una lengua que tiene evidenciales, indica que no es un hablante nativo o que no se compromete con lo que dice. En otras palabras, en estas lenguas es mandatorio usar evidenciales.

Pero hay en la Amazonía peruana un caso aún más interesante. Los matsés o mayoruna son un pueblo de aproximadamente 2200 individuos que viven en la frontera entre el Perú y Brasil, cerca del río Yaquerana. Su lengua, estudiada por Fleck (2006), obliga a los hablantes a hacer distinciones que no suelen hacerse en otras lenguas. Por ejemplo, hay tres tipos de pasado en el que puede haber ocurrido un acontecimiento: pasado reciente (aproximadamente hace un mes), pasado distante (aproximadamente de un mes a cincuenta años) y pasado remoto (aproximadamente de cincuenta años a más). El sistema de evidencialidad es el más complejo conocido. Cuando un matsés se refiere a un acontecimiento, está obligado a reportar con precisión de qué manera tomó conocimiento de los hechos que está narrando. Hay cuatro evidenciales: el que reporta algo experimentado directamente por uno (por ejemplo, «Yo vi un jaguar»); el que reporta algo inferido («Vi las huellas de un jaguar»); el que reporta una conjetura («Los jaguares pasan siempre por aquí»); y el que reporta algo que se escuchó («Me dijeron que hay jaguares en esta zona»). Una descripción puede ser muy compleja si uno combina los tres tipos de pasado con los cuatro marcadores de evidencialidad. La complejidad puede ser extrema si además uno itera niveles de intencionalidad (por ejemplo: «X me dijo que Y había dicho que Z pensaba que…», etcétera). Puede ser que ese grado de precisión sea sumamente ventajoso para una sociedad de cazadores-recolectores, pero es claro que no toda sociedad de cazadores-recolectores tiene esa complejidad gramatical. El punto es que si uno es un matsés está obligado a pensar en detalles que no tiene que pensar un hablante de otras lenguas, pero el ser hablante de otras lenguas no impide pensar lo que piensa un matsés.

En la *Estructura de las revoluciones científicas*, Kuhn defiende una versión fuerte del whorfismo, aunque aplicada a los paradigmas científicos. Parece claro que si el whorfismo fuerte tiene poca evidencia a su favor, aplicado a los paradigmas científicos es aún menos plausible.

Sin embargo, el valor de Kuhn radica en mostrar que el empirismo radical de los positivistas lógicos no puede sostenerse seriamente pues lo termina conduciendo a su propia disolución, ya sea en el idealismo o en el nominalismo, en particular en lo concerniente a la distinción fuerte entre términos teóricos y términos observacionales. Si los términos teóricos solo son significativos si refieren a objetos de experiencia —y como finalmente la experiencia es un fenómeno mental— los objetos de nuestro conocimiento terminan siendo contenidos mentales, ya que entonces se torna imposible distinguir entre la privacidad o intersubjetividad de nuestras imágenes mentales. Si solo tienen significado los términos que refieren a objetos perceptibles, es obvio que los términos generales carecen de significado. Luego, los únicos enunciados observacionales serían aquellos que describen una experiencia particular, tal como Neurath

mostró en su comentario a Carnap (Neurath, 1932-1933). Pero, como todo lenguaje requiere de términos generales, el lenguaje observacional básico solo podría estar constituido por deícticos, de los cuales nunca se podría pasar a enunciados teóricos.

Experimentamos el mundo físico a partir de categorías cognitivas que son parcialmente universales, pues proceden de la evolución de nuestros cerebros, y parcialmente culturales, porque dependen de las características peculiares de nuestras lenguas y culturas. Por tanto, la experiencia no es unívoca y en alguna medida está influida por la interpretación que nuestras creencias previas le confieren.

En «On the Very Idea of a Conceptual Scheme», Davidson (1984d) esgrimió un argumento célebre —y según algunos definitivo— contra la tesis de la inconmensurabilidad, que también ha sido aplicado contra algunas versiones de relativismo conceptual. La inconmensurabilidad podría resumirse como la tesis que sostiene que dos teorías —o paradigmas— son inconmensurables si, pretendiendo describir o explicar los mismos fenómenos, no existe criterio objetivo que permita decidir entre ambas, pues al ser intraducibles resultan incomparables entre sí. Estas teorías serían inconmensurables porque no comparten una estructura conceptual que permita que la evidencia empírica favorezca a una de ellas o a ninguna. El relativismo conceptual, por otra parte, podría resumirse en la afirmación según la cual no existen criterios objetivos que nos permitan decidir epistémicamente entre dos esquemas conceptuales en disputa.

El argumento davidsoniano es el siguiente[3]: si sostenemos que A y B hablan sobre el mismo objeto —por ejemplo Tolomeo y Galileo sobre el Sol— es porque asumimos que tenemos criterios para determinar que A y B comparten un número elevado de creencias sobre ese objeto, lo que nos permite definir a ese objeto como común para ellos y para nosotros. Si decimos que A, B y nosotros mismos compartimos un gran número de creencias, no podemos decir también que A y B son inconmensurables. Podemos afirmar que son parcialmente inconmensurables, pero eso equivaldría a postular que son parcialmente conmensurables, que es todo lo que necesitamos para rechazar la inconmensurabilidad.

Una interpretación del argumento antirrelativista de Davidson, aunque es incierto si el propio autor lo hubiera aceptado, afirmaría que el significado de una oración es relativo a una teoría o sistema de creencias y, dado que la verdad de una oración depende de su significado, entonces la verdad de una oración también sería relativa a una teoría o sistema de creencias. Hasta ahí tenemos un argumento contextualista y podemos aceptar que tanto el significado como la verdad son sensibles al contexto.

[3] Véase Duica, 2014, para un interesante análisis del argumento davidsoniano contra la dualidad entre esquema y contenido que subyace a la tesis de la inconmensurabilidad.

Pero ya en este punto el enfoque cambia por completo. El relativista conceptual diría que la aceptación del sistema de creencias es arbitraria e irracional, y que no hay criterios para preferir nuestro sistema de creencias por sobre cualquier otro diferente.

Mi interpretación diría que la elección entre sistemas sigue siendo racional, aunque desde los criterios de nuestro inevitable sistema de creencias. La fuerza del argumento radica en que, para que reconozcamos a los sistemas de creencias diferentes como tales, estos deben ser traducibles entre sí, con lo cual la elección de criterios de verdad sigue siendo posible y racional. Como resultará claro, la disolución del relativismo conceptual depende de reconocer que toda afirmación acerca de la verdad o falsedad de una creencia, o acerca de la intertraducibilidad o no de sistemas de creencias, se da desde el punto de vista de una intérprete que, a su vez, pertenece a algún sistema de creencias, y no desde un punto de vista privilegiado. Solo se puede afirmar que dos sistemas de creencias son intraducibles entre sí, si después de haberlos comprendido, afirmamos que entre ellos no podrían comprenderse.

En lo concerniente a una forma más básica de relativismo, si no hay criterios objetivos para establecer la verdad o falsedad de ninguna oración, tampoco es posible establecer el valor de verdad de la oración que afirma que «No hay criterios para establecer el valor de verdad de ninguna oración». En esta discusión estoy planteando las cosas en términos de la relatividad de la verdad, porque cualquier otra forma de relativismo puede ser formulada a partir del problema de la verdad. Por ejemplo, la pregunta sobre si hay valores morales universales, estéticos o de cualquier otra índole puede plantearse en términos de si hay oraciones valorativas —morales, estéticas o de cualquier otra índole— cuyos valores de verdad sean universales.

Si bien esto es cierto, como también es cierto que la oración «La masa es invariable» es verdadera al interior de la mecánica newtoniana, pero no de la teoría de la relatividad de Einstein, eso no impide que pueda haber discusión racional acerca de las conveniencias y ventajas por las que es posible aceptar ciertos enunciados, incluso en comunidades donde no han sido planteados originalmente.

Que la verdad de una oración depende de un contexto es algo bastante obvio, ya que, como acabamos de ver, la verdad de una oración depende de su significado y su significado depende de un contexto. Lo que añade un elemento de relativismo conceptual es la posibilidad de que los distintos contextos sean diferentes a tal punto que no puedan ser integrados en un contexto común, es decir, que no puedan ser traducidos ni comparados entre sí. Esto es lo que normalmente se entiende por inconmensurabilidad y esta es la posibilidad que el argumento davidsoniano rechaza. Lo que es discutible es si inconmensurabilidad e intraducibilidad apuntan a lo mismo. Aunque Davidson sostiene que sí, hay quienes no están de acuerdo (Doppelt, 1978; Rorty, 1989; Bernstein, 1983), por lo que el debate continúa abierto.

Kuhn ha motivado una reflexión más profunda sobre las relaciones entre las ciencias naturales y las ciencias humanas, así como replanteado la discusión ontológica en la epistemología. Muchas de sus tesis son actualmente difíciles de aceptar, pero su trabajo no solo nos ha servido para superar presupuestos obsoletos de la filosofía de la ciencia tradicional sino también para ver el mundo de la ciencia de una manera diferente.

Este capítulo está dedicado a Kuhn por tres razones. En primer lugar y para los efectos de nuestro tema central, él nos ha mostrado que la explicación nunca es aséptica ni neutral, pues siempre está cargada teóricamente y procede de comunidades epistémicas que tienen características sociales particulares. Aquello que vale para la explicación se aplica también, y con mayor razón, a la comprensión. En segundo lugar, se sigue de lo anterior que la distinción entre ciencias naturales y ciencias humanas, tan rígida hasta fines del siglo XX, debe cualificarse. En tercer lugar, Kuhn está entre quienes más han reflexionado sobre los problemas que surgen cuando necesitamos criterios para elegir entre teorías en disputa que pretenden explicar el mismo fenómeno. Este punto nos concierne especialmente porque, como vimos en el primer capítulo y volveremos a ver, también necesitamos criterios para determinar cuándo una interpretación de un agente intencional es preferible a otras porque permite comprenderlo mejor que las demás. Como señalé en el primer capítulo, los criterios que empleamos en las ciencias humanas para explicar no son muy diferentes de los que empleamos en las ciencias naturales, con la diferencia de que la comprensión, a diferencia de la explicación, pretende compartir una perspectiva subjetiva. Más aún, los criterios que empleamos para determinar cuándo una interpretación es preferible a otra son también básicamente los mismos que usamos cuando queremos saber si una teoría científica es más explicativa que otra. En cuarto lugar, Kuhn nos ha llamado la atención sobre el problema de la variación de significado y las consecuencias que eso tiene, lo que nos conduce a nuestro próximo tema.

Para comprender a alguien es importante conocer los significados de sus acciones intencionales en general y, en particular, los de las oraciones y expresiones que él profiere. Es necesario conocer tanto los significados que él cree que las expresiones tienen como las que uno podría creer que tienen, en caso que considere que el agente está equivocado. Por ello, en la segunda parte de este libro me concentraré en la naturaleza del significado y sostendré que tanto el significado como la referencia se fijan y emergen al interior de situaciones interpretativas y comunicativas de tipo triangular.

Segunda parte
Interpretar significados y metáforas

Capítulo cuatro
¿Qué es y cómo emerge el significado?

4.1. Significado e interpretación radical

En la mayor parte de situaciones comunicativas comprender a un hablante supone saber los significados que él atribuye a sus palabras. Pero, ¿qué son los significados, cómo se constituyen y de qué sistemas complejos emergen?

Los significados son propiedades que tienen algunas acciones intencionales, y las proferencias verbales o las expresiones escritas son un subconjunto de ellas. Las acciones, a su vez, son eventos físicos dotados de intencionalidad y causados por los estados mentales de un individuo dotado de agencia. Las acciones son también eventos físicos, pero no todos los eventos físicos son acciones. Por otra parte, no todas las acciones intencionales tienen significado. ¿Qué hace, pues, que una acción signifique? ¿Qué es el significado o qué es un significado? ¿Qué significa «significar»?

Los significados no son entidades que existan en sí mismas ni tampoco propiedades platónicas intemporales de las expresiones. Hay quienes creen que sí lo son, pero esa es una posición imposible de defender empíricamente y muy difícil de justificar conceptualmente, a menos que se sucumba a algún tipo de circularidad. Adicionalmente esa tesis no tiene ninguna ventaja para explicar situaciones interpretativas o comunicativas y resulta irrelevante si se tiene interés en dialogar con la lingüística, la psicología o las ciencias cognitivas. Asumiré, por tanto, que los significados no son eternos y que se constituyen en el tiempo, por lo que será necesario preguntarse cómo emergen.

Hay muchas teorías del significado y no parece que la producción de estas se esté deteniendo, de manera que no voy a revisarlas todas sino tomaré posición por una familia de ellas. Tampoco entraré en discusiones técnicas sobre los detalles de alguna teoría del significado, a menos que sea relevante para mi objetivo principal que es aclarar la naturaleza de la comprensión.

Sostendré que el significado es una propiedad que emerge en una situación comunicativa triangular que vincula a un hablante, una intérprete y el mundo que comparten, que asumen compartir y acerca del cual se expresan. Las tesis que defenderé están influidas por Davidson (1984c), aunque muchas de ellas no proceden de él y es discutible si hubieran sido aceptadas por ese filósofo, de haberlas conocido[1]. Para ilustrar la posición comenzaré con la adaptación de un famoso experimento conceptual llamado por Quine «traducción radical» (1960, capítulo 2) y elaborado por Davidson (1984e [1973]) con el nombre de «interpretación radical». Este experimento conceptual nos permite analizar la estructura formal de la interpretación sin que esté afectada por los contenidos que pudieran o no compartir los interlocutores.

Imaginemos la siguiente situación. Un equipo internacional de filósofas, psicólogas, lingüistas y antropólogas descubre una nueva comunidad no contactada en la Amazonía sudamericana. Ellas se ponen en contacto con los miembros de esa comunidad y advierten que no hablan ninguna lengua conocida y no parecen tener una visión del mundo de la que tengamos noticia. Sin embargo, ellas y ellos necesitan comunicarse entre sí. La historia que narraré es una versión ficticia de casos que realmente ocurrieron. Por ejemplo, se produjo una situación de interpretación radical cuando Cristóbal Colón se encontró con un grupo de nativos en una isla del Caribe, hacia 1492; también cuando Francisco Pizarro conoció a Atahualpa en Cajamarca, una tarde de noviembre de 1532, y todas las veces que aventureros, navegantes, descubridores y conquistadores encontraron pobladores con quienes no compartían nada y con los cuales —por cualquier motivo— deseaban comunicarse. La tesis de Davidson es que si examinamos lo que ocurre en estos casos de interpretación radical, notaremos que es una versión más extrema de cualquier otro caso de comunicación más familiar, como cuando converso con mi vecino o con un viejo amigo, de manera que analizar sus características nos permitirá entender fenómenos menos radicales.

Interpretación, comunicación y comprensión conforman un continuo que tiene, en un borde, lo que ocurre entre quienes no comparten lengua ni visión del mundo y,

[1] Aunque el pensamiento de Davidson ha tenido una gran influencia en la discusión filosófica mundial desde la década de 1970, sus detalles técnicos también han sido ampliamente cuestionados. En muchos casos, esos cuestionamientos se han concentrado en cuestiones accesorias más que en las intuiciones filosóficas centrales, lo que ha conducido a debates sobre detalles que considero poco fértiles. Asimismo, muchos cuestionamientos no entienden los detalles porque se concentran en un aspecto de la filosofía de este autor y pasan por alto el contexto más general de su obra. Por ello, es conveniente tener una visión de conjunto del pensamiento de Davidson antes de cuestionar aspectos específicos. Este libro no ofrece esa visión de conjunto, pero para ello puede verse Ramberg, 1989; Evnine, 1991; Malpas, 1992; Stüber, 1993; Engel, 1994; Caorsi, 1999 y 2001; Hernández Iglesias, 2003; Glock, 2003; Smith y Silva Filho, 2005; Lepore y Ludwig, 2005 y 2013; Caorsi y Silva Filho, 2008; Glüer, 2006; Duica, 2014; Pedace, 2017; entre otros.

cerca del otro, lo que acontece cuando doy sentido a las palabras de mi mejor amigo. Según las características de cada quien y de su hábito de autoexaminarse, en algún punto está la interpretación que uno hace de sí mismo, pues uno también trata de dar sentido a sus propias acciones y estados mentales, las que con frecuencia le resultan tan ajenas y extrañas como las de los demás. Así como la interpretación del otro se llama «aleointerpretación», la de uno mismo es «autointerpretación». En distintos puntos de este continuo están todas las relaciones de alteridad entre civilizaciones, culturas, clases sociales, poblaciones, parejas e individuos.

Las investigadoras de nuestro experimento conceptual se propondrán reconstruir las creencias y deseos de los nativos. Desearán saber cómo reaccionarán y con qué objetivos, y si quieren comunicarse o manipularlas. Quizá en algunos momentos de lucidez autocrítica ellas también se pregunten hasta qué punto están allá para aprender de ellos o para ejercer su poder y con qué objetivos ulteriores. También querrán saber, por ejemplo, si ellos creen que cualquier foráneo es un enemigo y debe ser preparado para la cena, o si desean alimentarse de carne humana por considerarla de gran valor nutritivo.

Ellas desearán confeccionar un diccionario de la lengua nativa, para lo cual elaborarán un «manual de traducción», es decir, apuntarán en un cuaderno las expresiones que escuchan con más frecuencia y anotarán hipótesis sobre sus posibles significados. Ciertamente habrá mejores y peores intérpretes, es decir, algunas serán particularmente hábiles en el arte de conjeturar arriesgadas e innovadoras hipótesis para dar sentido al comportamiento más extraño de los agentes, atribuyéndoles estados mentales y significados a sus acciones y proferencias verbales. Las intérpretes más creativas serán aquellas capaces de alejarse de sus propios estados mentales para concebir formas de vida muy distintas de las suyas. A las intérpretes menos creativas, en cambio, les resultará difícil apartarse de sus propios estados mentales y formas de vida, para concebir lo diferente. Ellas tenderán a atribuir a los demás sus propios estados mentales o asumirán que ellos reaccionarán ante las diversas situaciones como ellas reaccionarían. Sin embargo, el ejercicio en la elucidación de lo insólito y desafiante las convertirá en intérpretes más creativas y audaces.

Una primera cosa importante que las investigadoras de nuestro experimento conceptual descubrirán es que construir un sistema de creencias, deseos, afectos y significados implica elaborar una red holista —del griego *holós*, totalidad —en la que cada estado mental y significado se define en relación con los otros elementos del sistema, pues su contenido no es independiente ni se puede aislar del sistema al que pertenece. En otras palabras, fuera de un sistema no tiene sentido hablar de significados ni estados mentales, pues no sabríamos qué contenidos asignarles. Esa es, por supuesto, una idea que se puede rastrear hasta los orígenes mismos de la lingüística

como ciencia a comienzos del siglo XX (Saussure, 1980) y que también se aplica al ámbito de lo psicológico.

Una segunda cosa que ellas notarán, es que el sistema que construirán para hacer inteligibles a los nativos estará conformado por un conjunto de atribuciones que dependerá en gran medida de las creencias que ellas tengan acerca de ellos, las que cambiarán en la medida en que interactúen más fluidamente. Estas creencias se dan de manera entrecruzada y en varios niveles de intencionalidad, de manera que ellas tendrán creencias acerca de ellos, acerca de las creencias de ellos, acerca de las creencias que ellos tienen de ellas, acerca de las creencias que algunos de ellos tienen respecto de otros y sobre las creencias que ellas tienen acerca de las de ellos. Aunque lógicamente estos niveles son infinitos, en la práctica no lo son, pues se encuentran limitados por la capacidad cognitiva humana que, como hemos visto, llega a un máximo de cinco niveles de intencionalidad en promedio. Así es como ellas modificarán sus hipótesis interpretativas según observen que estas logran dar sentido al comportamiento de los nativos. Pero, en tanto noten que la evidencia conductual no es consistente con sus hipótesis o que estas no pueden predecir exitosamente un nuevo comportamiento, deberán hacer afinamientos en los estados mentales atribuidos.

Una tercera cosa que ellas advertirán es que creencias, deseos, afectos, valores, acciones y significados se definen mutuamente, de manera que ninguno de ellos puede ser el punto de partida. Esto implica que para saber qué creen ellos, se deberá saber primero qué desean, qué sienten y qué significan sus acciones y palabras. Para saber qué desean se tendrá que determinar qué creen, qué sienten y cuáles son sus significados. Y así en adelante. En otras palabras, se tendrá que comenzar con la atribución de sistemas de estados mentales ya conectados y no individuales.

Supongamos, por ejemplo, que ellas ven a los nativos moviéndose velozmente alrededor de una palmera, gesticulando y pronunciando repetidamente las que parecen palabras de gran importancia ritual. Lo que ellas en principio están observando son solo movimientos físicos que, asumen, también pueden describirse como acciones. Pero para averiguar qué movimientos son acciones y cuáles no lo son, así como qué significan esas acciones, y para saber si los ruidos que salen de sus labios son palabras y qué significan, deberán imaginar qué pretenden hacer con esos movimientos y sonidos, es decir, cuáles son sus estados mentales. Las investigadoras sospecharán que se trata de una danza sagrada mediante la que desean honrar a sus dioses y querrán saber cómo conciben a tales dioses y cómo creen que estos desean ser honrados. Pero supongamos que después de observarlos por varias semanas, llegan a la conclusión de que no se trata de ninguna danza ritual sino de la actividad comunal de hacer círculos en la tierra alrededor de las palmeras para poder regarlas con más facilidad. En ese caso, las atribuciones originales de estados mentales y significados habrán sufrido

muchos cambios. Es innecesario decir que el nuevo conjunto de hipótesis no será el definitivo, pues también está abierto al cambio. Por tres razones las investigadoras nunca completarán su interpretación de los nativos:

(1) Ellas siempre podrán encontrar hipótesis interpretativas más finas y precisas, desde su propio punto de vista, naturalmente.

(2) Ellos están siempre cambiando y las hipótesis exitosas del pasado podrían no funcionar para el futuro.

(3) Ellas están siempre cambiando y, dado que sus hipótesis están elaboradas sobre sus propios estados mentales —acerca de ellos y acerca de ellas mismas—, sus interpretaciones serán siempre procesos inacabados e inacabables.

Ahora supongamos que cada vez que aparece un conejo los nativos emiten el sonido «*gavagai*» (este es el célebre ejemplo en Quine, 1960). Las investigadoras apuntarán en su cuaderno «*Gavagai* significa conejo» y creerán que ya tienen un primer significado de la lengua nativa. ¿Pero, cómo podrán ellas saber que *gavagai* es conejo y no conejo macho, conejo blanco, conejo vivo, conejo corriendo, conejo en estado de libertad, etcétera? No importa cuántas y qué variaciones puedan hacer las investigadoras, jamás agotarán todos los posibles escenarios, pues *gavagai* podría significar «parte no separada de un conejo» o «La idea de la conejidad se ha manifestado localmente». El punto es que no importa cuánta evidencia tengan las investigadoras, nunca podrán detectar el «verdadero» significado de *gavagai*. Pero no porque ese significado esté siempre oculto para ellas, como una especie de *noúmenon* incognoscible, sino porque no existe tal cosa como «el verdadero y real significado de algo». El significado de una acción, una oración, una palabra —y en este caso de *gavagai*— es siempre una construcción cambiante entre una intérprete, un agente y un objeto que comparten y al que ambos asumen están refiriendo. Además, como veremos en un momento, en principio, distintas investigadoras podrían atribuir correctamente diferentes significados a las mismas acciones o proferencias verbales.

Mi tesis es que el significado de una expresión —o de una acción en general— es una propiedad relacional triádica que emerge a partir de un conjunto de secuencias regulares de situaciones comunicativas exitosas. Pero para explicar esta posición será necesario hacer un breve recuento metafísico. Desde Aristóteles (1982) es común afirmar que la realidad se compone de dos tipos de entidades: sustancias y propiedades. Esta no es la única descripción que se puede hacer de la realidad ni tampoco es la única correcta. La realidad puede describirse de múltiples formas, y varias de estas pueden ser simultáneamente válidas. Sin embargo, esta descripción es particularmente útil para explicar la naturaleza del significado. De acuerdo con ella, entonces,

el universo puede ser descrito en términos de sustancias y propiedades. Las primeras existen en sí mismas y no requieren de otros objetos para existir. Una persona, un electrón, una mesa o una galaxia son sustancias. Las propiedades, por el contrario, también llamadas atributos o cualidades, no existen en sí mismas sino solo instanciadas en una sustancia. Por ejemplo el color azul, la bondad, la belleza, la magnitud o la velocidad. Pero las propiedades pueden ser monádicas o relacionales, y si son relacionales pueden ser diádicas, triádicas y n-ádicas. La forma circular de una mesa es una propiedad monádica, mientras que el ser más grande que un alfiler es una propiedad relacional diádica que tiene cualquier sustancia que cumpla con esa condición.

Hay un sentido en que toda propiedad es relacional y, por tanto, la idea de propiedad monádica es una idealización. Esto es así porque toda propiedad de una sustancia se define en relación a otras propiedades. Por ejemplo, la forma circular se define en relación a formas que no lo son y la masa de un objeto se mide en relación a la de otros objetos. Más aún, hay quienes dirían que toda propiedad contiene una implícita comparación con una teoría o con un observador, de manera que las supuestas propiedades monádicas están en la relación entre la sustancia y una teoría que la describe. El primero en desarrollar la intuición de que toda propiedad es relacional fue Berkeley (1974) cuando objetó la distinción misma entre cualidades primarias y secundarias que había desarrollado Locke (1982), con lo que propuso su concepción inmaterialista, que después dio lugar a lo que se denominó idealismo y antirrealismo. Este es un tema complejo en el que no necesito detenerme en esta ocasión; bastará con decir que la distinción entre propiedades monádicas y relacionales es metodológicamente útil incluso si uno piensa que las propiedades monádicas son un subconjunto idealizado de propiedades relacionales. El punto es que en el caso de las propiedades relacionales es más obvia la comparación entre sustancias y la propiedad no reside en la sustancia sino en la interacción entre ellas. Por ejemplo, la propiedad «ser más grande que» es una sustancia relacional porque ningún objeto podría tener esta propiedad si no se lo compara explícitamente con otro, mientras que la propiedad «estar hecho de madera» puede estar instanciada en una mesa sin mayor comparación con otros objetos del mundo, aunque está claro que solo se puede ser de madera si no se es de metal ni ladrillo, y que uno entiende la propiedad «ser de madera» solo si tiene una concepción de esta.

No debe suponerse que la sustancia tiene una realidad previa e independiente de sus propiedades, como si ellas fueran aditamentos accesorios de una entidad que preexiste a estas, como sugiere el latín *substare* —del que procede «sustancia»—, que literalmente significa lo que está por debajo o lo que subyace, de manera análoga a *subjectum*, del que procede «sujeto».

Las sustancias son conjuntos de propiedades y cambian en tanto cambien las propiedades. Aristóteles sostuvo que las propiedades pueden ser esenciales o accidentales, de manera que el cambio de las primeras implica una transformación esencial en la que la sustancia deja de ser lo que es para convertirse en otra, mientras que el cambio de las propiedades accidentales solo supone una transformación del mismo objeto. Creo que este es un compromiso muy radical que no es necesario hacer, por lo que me inclino por asumir que todo cambio de propiedades implica un cambio de la sustancia, en mayor o menor grado. Una modificación masiva de las propiedades de una sustancia puede conducirnos a sostener que ahora tenemos una nueva sustancia, aunque no hay una línea fronteriza clara que marque la diferencia.

Un ejemplo paradigmático de propiedad relacional triádica es el color. Por ejemplo, el color azul del mar que tengo frente a mí es una propiedad relacional triádica, porque no reside en las moléculas de agua, en la luz que es refractada por tales moléculas, ni en mi retina que recibe la luz y envía información a través del nervio óptico a mi cerebro, con lo que se produciría la experiencia fenoménica del azul. El azul está en la relación entre esas tres sustancias, de manera que si alguna de ellas faltara el azul no existiría. Así, cuando cae la noche el mar ya no es azul sino negro, que por otra parte es la ausencia de color. No es que siga siendo azul pero que no lo podamos ver, simplemente dejó de ser azul. Lo mismo ocurriría si faltara alguna de las otras sustancias que permiten que el azul emerja como una propiedad relacional triádica. De hecho, el mar está compuesto por moléculas de agua que son incoloras, de manera que su color azul es una suerte de ilusión óptica producida por la luz en la retina. Pero esto no significa que los colores no sean reales, son tan reales como cualquier otra propiedad solo que su realidad depende de que son propiedades relacionales triádicas. Así, la oración «El mar que tengo frente a mí es azul» es perfectamente verdadera y describe un hecho de la realidad, pero un hecho que se constituye triangularmente con la participación de la luz, las moléculas de agua y un cerebro sano.

Las propiedades no existen en sí mismas sino instanciadas en sustancias. Pero eso no significa que las propiedades no existan, ciertamente existen, solo que en una o más sustancias y no independientemente de ellas. Como las sustancias son objetos naturales espacio-temporales, en última instancia todas las propiedades existen porque existen objetos naturales, de manera que si estos no existieran tampoco existirían las propiedades. El número 2, por ejemplo, es una propiedad que tiene una pareja de objetos que conforman un conjunto, y el 2 no tiene ninguna otra realidad que ser esa propiedad instanciada en las diversas parejas de objetos que hay en el universo (para esta concepción de la ontología de los números véase Russell, 2017). Pero, naturalmente, podemos hablar de objetos que no existen, como Zeus y las sirenas, y de propiedades que no son instanciadas en ningún objeto, como el círculo cuadrado o el actual rey de Francia.

Ahora bien, las propiedades existen en distintos niveles de emergencia respecto de las sustancias. Cuando se produjo el Big Bang, hace aproximadamente 13 800 millones de años, surgió un conjunto de sustancias físicas, básicamente partículas subatómicas, y un conjunto de propiedades que se instanciaron en ellas, por ejemplo, la masa y la velocidad. En ese punto de la historia del universo no había más que propiedades físicas. Hace aproximadamente 3500 millones de años algo ocurrió en nuestro planeta que permitió que surgieran propiedades biológicas que antes no existían. Estas son propiedades emergentes de sistemas complejos que requieren tanto de sustancias como de propiedades físicas para existir, aunque no se reducen a propiedades físicas ni en un sentido ontológico ni en un sentido epistémico. Con el paso del tiempo, surgieron en nuestro planeta especies sociales y, así, a partir de nuevos sistemas complejos, emergieron propiedades sociales como el poder, la autoridad y el parentesco. Estas propiedades surgieron solo porque existían sustancias con propiedades físicas y biológicas. Más adelante aún, emergieron propiedades psicológicas como la sensación y las emociones, la conciencia autobiográfica y el razonamiento lógico. También surgió el *modus ponens*, el principio de no contradicción y el del tercio excluido, que son propiedades del razonamiento humano, aunque, en tanto son propiedades emergentes, son parte de la realidad y tienen tanta existencia como las clases sociales, los adverbios y las mesas de madera, a pesar de que tienen sentidos diferentes de existencia porque son propiedades que emergieron de distintos sistemas complejos. Así también surgieron los distintos estados mentales y los significados de las palabras, oraciones y otras acciones intencionales. Ciertamente existen los significados, pero solo en tanto propiedades relacionales que emergen en situaciones comunicativas regulares y predecibles, que son prácticas sociales compartidas por una comunidad de hablantes. Por ello mismo, los significados —y también las proposiciones, que son los significados de las oraciones— solo pueden existir porque existen comunidades de hablantes que se comunican e interpretan mutuamente, y existen «en» esas comunidades, como propiedades emergentes.

Mi tesis, entonces, es que el significado de una expresión —y más adelante diré lo mismo acerca del contenido proposicional de los estados mentales— es una propiedad relacional triádica emergente, que se constituye cuando se dan las circunstancias comunicativas que lo hacen posible. Se constituye el significado de una expresión —o en general de una acción— cuando se conforma una relación comunicativa entre una intérprete, un hablante y el mundo que ambos comparten; y esta relación comunicativa se hace tan regular y frecuente que incluso puede ser predicha, de manera que los interlocutores no tendrán mayores problemas para usar o interpretar la expresión porque asumirán que su interlocutor la entenderá correctamente.

Formulado de otra manera, podemos decir que una palabra «tiene» un significado, en el sentido de que esa palabra se usa comúnmente de cierta manera en una comunidad de hablantes, es decir, que ese uso ha sido convencionalizado. Pero esa propiedad a la que llamamos «significado» está conformada por múltiples relaciones triangulares entre hablantes, intérpretes y la realidad objetiva que, naturalmente, incluye el objeto del que hablan y al que probablemente la palabra refiere. Pero esto no basta para que exista el significado. También es necesario que esta situación comunicativa conformada por varias relaciones triangulares se convierta en una regularidad, es decir, en una práctica social compartida, de manera que es factible sostener —en una línea wittgensteiniana— que el significado de x es una regla de uso compartida dentro de cierta comunidad de hablantes o de agentes intencionales.

Podemos decir, por ejemplo, que el significado de la palabra «caballo» está conformado por las prácticas sociales compartidas que entrañan reglas de uso de esa expresión en una comunidad de hablantes determinada, y que hace que la oración «Iré a buscarte montado en mi caballo» sea aceptable y genere ciertas situaciones comunicativas, aunque no así la oración «El supremo caballo adormeció con sus finas imágenes las ligaduras de la Luna». La mayor parte de hablantes de nuestra comunidad comprendería la primera oración pero no la segunda, y eso es porque en el primer caso la palabra «caballo» está siendo usada según ciertas reglas convencionales, mientras que en el segundo caso eso no ocurre. Naturalmente hay casos imprecisos como los siguientes: «El supremo caballo adormeció la noche con el fino movimiento de sus cascos», o «El poderoso caballo atravesó la noche galopando bajo la Luna». En todos los casos hay reglas que gobiernan el uso de las palabras, conformadas por prácticas regulares ancestrales, pero esas reglas pueden ser más o menos flexibles, y dar lugar a significados más o menos precisos. En ocasiones esas reglas son deliberadamente transgredidas y así tenemos oraciones que, careciendo de significados precisos, pueden generar situaciones comunicativas muy valiosas. A algunas de esas oraciones las llamamos «metáforas». Ese es el terreno de las relaciones entre semántica y pragmática, sobre lo que volveré sobre más adelante.

Ahora bien, si el significado es una propiedad relacional triádica que emerge en la relación triangular entre intérprete, agente y mundo objetivo, la intérprete tiene un rol activo en la construcción del significado de las acciones y su tarea no es solo desocultar o revelar las «verdaderas» intenciones del agente. La posición que defiendo está en la orilla opuesta respecto de la hermenéutica intencionalista del siglo XIX, la cual postulaba que, dado que el significado de un texto o de las palabras de un hablante dependen de las «verdaderas» intenciones de su autor, la interpretación es el proceso de descubrir aquellas reales intenciones que se ocultan en el interior de las mentes de las personas. Esa posición es representada paradigmáticamente

por el teólogo luterano Friedrich Schleiermacher (1986, 1996), quien pensaba que la comprensión correcta de las Sagradas Escrituras depende de entender las reales intenciones de sus autores. Actualmente el más interesante representante de esta posición es Eric Donald Hirsch (1967, 1978), quien la ha aplicado a la teoría literaria. Pero la hermenéutica intencionalista presenta varios problemas. En primer lugar, quizá no exista tal cosa como las «verdaderas» intenciones del autor, pues estas están siempre cambiando y con frecuencia coexisten varios niveles de intenciones diferentes, con diversos grados de conciencia de parte del autor. En segundo lugar, muchas veces el autor desconoce sus propias intenciones o, por lo menos, las que bajo la interpretación de otra persona tienen mayor rol causal o, incluso, las que resultan simplemente más interesantes, creativas u originales. En tercer lugar, la hermenéutica intencionalista parece una versión de la teoría ideacional del significado y, por tanto, presenta los mismos inconvenientes.

La teoría ideacional considera que el significado es una representación mental o una idea. Así, por ejemplo, el significado de la palabra «casa» sería la representación mental que los hablantes de castellano solemos tener cuando escuchamos o proferimos esa palabra. Dado que no todos hemos aprendido a usar esa palabra ante los mismos estímulos sensoriales —algunos están familiarizados con casas de madera y tejas, otros con casas de concreto y habrá quienes entiendan «casa» como una construcción de carrizo— los significados de esa palabra variarán ligeramente entre los hablantes. Esta posición fue sostenida por los empiristas británicos (Locke, 1982, libro III) y probablemente presupuesta por Descartes (2011 [1641]). La objeción clásica a esta posición es que, si el significado fuese un estado mental, los hablantes estarían en un mismo estado mental al entender la misma expresión, lo que no es probable. Hay quienes piensan que los significados están «encriptados» en tanto representaciones mentales en los cerebros de las personas, pero eso es algo que tendría que demostrarse empíricamente y hasta ahora no se ha hecho. Sin embargo, incluso si llegara a demostrarse esto último, no sería incompatible con la tesis de que los significados de una expresión son las prácticas regulares de su uso en una comunidad de hablantes.

Otra teoría tradicional del significado es la llamada «referencialista». Planteada en su forma más cruda, esta sostiene que el significado de una expresión es el objeto o conjunto de objetos referidos por ella. Así, por ejemplo, el significado de «caballo» sería el conjunto de caballos que hay en el universo. Según el tipo de referencialismo, podría tratarse del conjunto de caballos que han existido y existirán alguna vez, o solo de los que existen ahora. Es claro que esta tesis es absurda porque implica que las palabras que carecen de referente —como «sirena»— no tienen significado. Por ello, en esta versión ingenua, ningún filósofo ha defendido explícitamente el referencialismo.

Sin embargo, esta posición suele ser presupuesta por las personas no familiarizadas con la filosofía y ha recibido el nombre de la teoría «Fido»-Fido, porque sostiene que el significado de «Fido» es precisamente Fido.

Ahora bien, aunque no hay ningún filósofo que haya sostenido explícitamente que el significado de una expresión sea su referente, sí hay quienes sostienen que el significado procede o se constituye a partir de la relación de referir, como es el caso de Frege (1984), Russell (1905) y el primer Wittgenstein (1975). Por ejemplo, en el *Tractatus Logico-Philosophicus* Wittgenstein pensaba que una oración es significativa si representa un estado de cosas posible de la realidad, de manera que su significado sería la representación que figura ese estado de cosas. Así, entonces, la oración «Atahualpa viajó a Castilla» es significativa —aunque falsa— mientras que «Ideas verdes incoloras duermen furiosamente» no tiene significado y, por tanto, tampoco valor de verdad. La concepción del *Tractatus* es una versión del referencialismo porque una oración es significativa en tanto puede representar un estado de cosas posible. Pero posteriormente en las *Investigaciones filosóficas* (1988), Wittgenstein abandonó esa tesis para acercarse a una concepción pragmática, según la cual el significado de una expresión estaría dado por sus reglas de uso en una comunidad de hablantes que comparte formas de vida y juegos de lenguaje. Es muy probable que para llegar a esta intuición haya sido influido por Peirce, a través de Frank Peter Ramsey, y de William James, a quien sí sabemos que leyó directamente, en particular *The Principles of Psychology* (1983) y *Las variedades de la experiencia religiosa* (1994).

Pero hay otros casos peculiares. Aparentemente algunos filósofos griegos presupusieron una teoría referencialista del significado, aunque no de manera consciente ni explícita, lo que explica que hayan sostenido algunas tesis que de otra manera serían ininteligibles. A eso volveremos ahora nuestra mirada.

4.2. El presupuesto referencialista en los filósofos griegos

Cuando se lee algunos textos de filósofos griegos antiguos que tienen en cuenta el problema del significado, se tiene la sospecha de que ellos comparten un presupuesto que no ha llegado a nosotros y sin el cual resulta extremadamente difícil entender la naturaleza de sus preguntas y el sentido de sus reflexiones.

Kerferd (1985) sostiene una tesis controversial, aunque ciertamente interesante, que también ha sido adoptada parcialmente por Andreas Graeser (1977) y Nicholas Denyer (1991; véase Quintanilla, 1994). La tesis afirma que, en la filosofía griega temprana, justo con la llegada de los jonios, es posible encontrar un presupuesto en torno del significado que atraviesa la filosofía griega y origina diversos problemas y paradojas. El presupuesto es que los nombres refieren a los objetos y las oraciones

refieren a los hechos, y que el significado de un nombre o de una oración está dado por los objetos o hechos a los cuales el nombre o la oración refieren.

Según esta posición, podemos entender las palabras «Sócrates» o «árbol» porque existe Sócrates y hay árboles, y el significado de estas palabras está fijado por una relación de referencia entre las palabras y los objetos. Así la oración «Sócrates es músico» es significativa porque refiere al hecho de que Sócrates toca algún instrumento. Kerferd lo pone de esta manera:

> El punto de partida puede ser planteado simplemente. Para los jonios y para virtualmente todos aquellos que llegaron después de ellos, las palabras obtienen su significado de los objetos a los cuales supuestamente nombran. Es porque las palabras *refieren* a objetos, que ellas poseen el significado que tienen. De aquí se sigue que hablar significativamente acerca del mundo es nombrar sus constituyentes (1985, p. 16).

Aceptar la tesis de Kerferd implica afirmar que los griegos presuponían, aunque nunca lo discutieron explícitamente, una teoría referencialista del significado. Esto obviamente suscita un gran número de problemas hermenéuticos como, por ejemplo, si es legítimo adscribir a un presocrático una opinión implícita acerca de un problema que aparece posteriormente, como el problema del significado. La respuesta de Kerferd a esta posible objeción sería que él no pretende decir que los presocráticos tuvieran una teoría del significado, lo que constituiría un obvio anacronismo. Él desea simplemente mostrar que ellos discutían algunos problemas filosóficos que se suscitaron porque presuponían lo que hoy nosotros llamamos, para nuestros intereses y curiosidades actuales, una teoría del significado.

Analizaremos la tesis de Kerferd para discutir algunas de sus consecuencias. Esta tesis se verá falsada si su aplicación oscurece los textos de los autores pertinentes, si suscita más contradicciones y paradojas de las que puede disolver o si, finalmente, entra en confrontación directa con textos explícitos. De otro lado, la tesis se corroborará si ilumina y hace inteligibles textos que de otra manera serían incomprensibles, demasiado arbitrarios o simplemente absurdos; si disuelve paradojas y contradicciones en mayor número de las que crea o si, finalmente, es confirmada por textos explícitos.

Pero aceptar la tesis de Kerferd implica enfrentarse a un problema: ¿qué ocurre con aquellos nombres que carecen de referente como, por ejemplo, «sirena», «vacío» o más gravemente «nada»? ¿Y qué ocurre con aquellas oraciones que carecen de referente, pues no refieren a ningún hecho, como cualquier oración falsa, las que versan acerca de objetos inexistentes o simplemente las oraciones acerca del futuro?

En efecto, un problema discutido entre los griegos era el de cómo es posible proferir oraciones falsas y hablar acerca de lo que no existe. Recordemos el célebre capítulo IX del *De interpretatione* de Aristóteles (2015), en el que este filósofo discute

el problema de si las afirmaciones acerca del futuro tienen un valor de verdad fijado antes de la ocurrencia del acontecimiento descrito o si carecen de valor de verdad, lo cual atentaría contra el principio de la bivalencia.

Hay otro problema que aparece a todo lo largo de la tradición griega y latina, desde Parménides hasta Agustín, y luego también en la tradición mística: ¿cómo es posible hablar sobre aquello que afirmamos está más allá de las posibilidades del discurso mismo? En otras palabras, ¿cómo es posible hablar acerca de lo inefable? Podemos ponerlo aún de otra manera: ¿Cómo es posible trazar la frontera, desde el lenguaje significativo, entre el significado y la ausencia de significado?

Este problema, abordado por Parménides y Platón, deriva en los neoplatónicos y los místicos. Parménides afirmó la imposibilidad de hablar acerca del no-ser. Pero decir que del no-ser nada puede predicarse es ya hacer una predicación sobre el no-ser, aunque quizá en un nivel metalingüístico. Afirma Platón, en la *Carta séptima* (1992) y en el *Fedón* (100a), la imposibilidad de poner en palabras sus contenidos mentales y considera la posibilidad de superar el lenguaje hacia un silencio significativo. Los filósofos cristianos tempranos, influidos por los neoplatónicos, desarrollaron la concepción mística según la cual lo único que podemos decir de Dios, además de afirmar su existencia, es que no hay nada que podamos decir de él, pues Dios, que es Uno, se encuentra más allá del lenguaje, que es múltiple. Mucho más recientemente, en la sétima proposición del *Tractatus* (1975) Wittgenstein incurre en la misma paradoja al afirmar que «De lo que no es posible hablar es mejor permanecer en silencio», que es probablemente una paradoja constitutiva de la filosofía occidental. Para aclarar un poco más esta problemática volvamos a la teoría referencialista del significado.

Si el significado de una oración está fijado por el hecho al que refiere, como sería absurdo admitir hechos negativos, las oraciones negativas tendrían que carecer de significado. Esa es precisamente la tesis sostenida por algunos filósofos antiguos como, por ejemplo, Menedemo de Eretria (Cicerón, 1944, libro 2, parágrafo 129, citado por Denyer, 1991, p. 37). quien afirmó que toda oración negativa debe ser convertida en positiva para ser significativa. Lo mismo ocurre con las oraciones falsas pues, dado que ellas no nombran nada, debemos admitir que carecen de significado y que, por tanto, es imposible decir algo falso. Esto es precisamente lo que Eutidemo sostiene en el diálogo platónico del mismo nombre (283e7-284c6). Aunque este argumento contradice el sentido común, no deja de ser interesante notar que Platón lo consideraba digno de ser discutido y registrado.

De igual manera, para que un nombre sea significativo debe referir a algo. No obstante, es necesario preguntar qué ha de contar como un referente aceptable. ¿Entidades espacio-temporales? ¿Objetos mentales? ¿Conceptos públicos e intersubjetivos? Esta es una debilidad en la tesis de Kerferd, pues no responde a esta pregunta fundamental.

En todo caso, si la tesis de Kerferd es correcta, podría explicarse de una manera nueva la afirmación parmenídea de la imposibilidad de hablar acerca del no-ser. Porque si la expresión «ser» refiere a la totalidad de lo existente, la expresión «no-ser» debe referir a la nada. Pero la nada no existe, luego «no-ser» no refiere, no es un nombre en absoluto y no hay nada significativo que se pueda decir al respecto. Ahí nacería una paradoja que atraviesa la filosofía occidental: la pretensión de hablar acerca de aquello de lo cual no es posible decir nada pues, en cierto sentido, se encuentra más allá del lenguaje y el significado. Esta pretensión va acompañada del intento por hablar acerca del no-ser como la única manera de decir algo sobre el ser.

El problema que llega hasta Platón es el de la posibilidad de hablar acerca de conceptos generales tales como «el bien», «lo azul», «la unidad», «la virtud», etcétera, como conceptos generales que son; sus nombres no refieren a entidades físicas, con lo cual —según el presupuesto referencialista— deberían carecer de significado. Por ello, Platón se habría visto obligado a postular la hipótesis de las Formas universales, con el objetivo de otorgar referencia —y por tanto significado— a los términos comunes. De otra manera, y tal como los sofistas sostuvieron, el discurso sobre estos tópicos carecería de significado.

El problema que Platón no resolvería y que Aristóteles abordó mediante la doctrina de las categorías, es el de la multiplicidad de la predicación, esto es, el que aflora cuando hacemos diversas predicaciones acerca del mismo sujeto y producimos diferentes hechos. Así, Simplicio cuenta, en su comentario a la *Física* de Aristóteles (Simplicio, 2012, citado por Denyer, 1991, pp. 41 y ss.), que los miembros de la escuela de Eretria entraban en pánico cuando tenían que afrontar el problema de que un hombre es realmente dos hombres si decimos que es blanco y que camina.

En el diálogo *Eutidemo* de Platón, el sofista del mismo nombre elabora un argumento según el cual es imposible proferir oraciones falsas. El argumento se extiende, en su forma resumida, de 28e7 a 284c6, y en concreto sostiene lo siguiente: Si uno profiere una oración falsa entonces está hablando acerca de algo; para que uno pueda hablar acerca de algo ese algo tiene que existir; pero si aquello de lo cual uno está hablando no existe, entonces uno no está diciendo nada falso. En conclusión, es imposible proferir una oración falsa. En su *Comentario al Cratilo de Platón*, Proclo (2007) resume esta argumentación de la siguiente manera: «Todo enunciado es verdadero, porque el que dice, dice algo; el que dice algo, dice lo que es, y quien dice lo que es, dice verdad». Proclo, como Aristóteles, afirma que esta posición pertenece a Antístenes, pero Platón la remonta hasta los protagóricos (*Eutidemo,* 286c2-3). En todo caso, es obvio que su origen más remoto está en Parménides.

El problema es, en suma, el siguiente: ¿Cómo podríamos hablar de algo si ese algo no existiese en ningún sentido? Y, naturalmente, el problema que se deriva es el de

convertir lo inefable en objeto de discurso: ¿Cómo es posible hablar sobre aquello que consideramos imposible de ser expresado? Si Kerferd tiene razón, el presupuesto no tematizado que atraviesa todas estas discusiones es que una palabra y una oración solo tienen significado si refieren a algo: un objeto, en el primer caso, y un hecho, en el segundo. Aparentemente muchos filósofos griegos encontraron la idea de que hablar o pensar sobre lo que no existe, o hablar sobre aquello que afirmamos no puede ser objeto de discurso, era intolerablemente paradójica, al extremo que no tuvieron reparos de llegar a consecuencias controversiales con tal de no admitir esa posibilidad. Este es un interesante caso de incompatibilidad entre intuiciones, en el que termina primando una de ellas.

Como vimos, la tesis de Kerferd explicaría la necesidad de Platón de postular la existencia de Formas universales para asignar referentes a los conceptos tratados por el discurso moral y epistemológico pues, de otra manera, no se podría decir nada mediante ellos. El presupuesto referencialista habría llegado hasta Platón representado por el personaje Cratilo, en el diálogo del mismo nombre, quien sostiene que hay una relación de necesidad entre el nombre y la cosa, de manera que a los objetos les correspondería un nombre en virtud de su propia naturaleza. Hermógenes, por el contrario, representa en ese diálogo una posición convencionalista entre el nombre y la cosa. El *Cratilo* es un diálogo aporético, porque Platón no concluye en él nada y solo presenta a Sócrates objetando ambas posiciones en debate, lo que sugiere que Sócrates, y probablemente también Platón, estaban en desacuerdo con el presupuesto central que originó el debate y que Cratilo y Hermógenes compartían.

Una consecuencia que Cratilo extrae de su propia posición es que la falsedad es imposible, porque decir algo falso es hablar acerca de lo que no es. Lo pondré de esta manera: decir algo falso es atribuir a un sujeto una propiedad que no le corresponde. Pero atribuir a un sujeto una propiedad que no le corresponde equivale a no hablar de él; de manera que decir algo falso equivale a hablar acerca de lo que no es, es decir, equivale a no decir nada en absoluto. De otro lado, utilizar equivocadamente un nombre, es decir, emplearlo para intentar referir a un objeto al que no le corresponde, es lo mismo que no emplearlo. Esto implicaría que los nombres no son nombres si no refieren al objeto que les corresponde por naturaleza.

La posición de Cratilo puede verse como una versión del presupuesto referencialista, por eso dice, en 435d, que quien conoce el nombre conoce la cosa. La idea es que si yo conozco el significado de una palabra conozco sus rasgos semánticos, y si conozco sus rasgos semánticos conozco los atributos o propiedades del objeto referido por el nombre. Luego, conocer el significado de una palabra implica conocer la realidad o la esencia del objeto referido por esa palabra. Así hay un sentido en que existe una relación de necesidad entre el nombre y la cosa, porque conocer el nombre es conocer la cosa.

Al inicio de su poema *El Golem*, escrito en 1958, Borges (2018) formula esta posición de Cratilo de manera magistral.

> Si (como el griego afirma en el Cratilo)
> el nombre es arquetipo de la cosa,
> en las letras de «rosa» está la rosa
> y todo el Nilo en la palabra «Nilo».
> Y, hecho de consonantes y vocales,
> habrá un terrible Nombre, que la esencia
> cifre de Dios y que la Omnipotencia
> guarde en letras y sílabas cabales.

A diferencia de Cratilo y Hermógenes, la posición socrática —no planteada en el diálogo— es que para poder referir a un objeto yo tengo que estar previamente familiarizado con su esencia, es decir, necesito tener un conocimiento previo de la Forma a la cual el objeto corresponde, de ahí que las palabras no refieran directamente a las cosas sino por intermedio de las Formas. Como Sócrates no cree que las relaciones de significar o referir se den entre el nombre y la cosa, sino más bien entre el nombre, la Forma y la cosa, es que discrepa tanto de Cratilo como de Hermógenes y, por ello, objeta a ambos en el diálogo. Así el presupuesto referencialista se mantendría presente en Cratilo y en el propio Sócrates, de maneras diferentes, aunque ya no en Hermógenes.

En el caso de Aristóteles, el presupuesto referencialista habría desaparecido casi totalmente y solo se mantendría en el caso de los nombres comunes que refieren a las esencias de las clases naturales, pues para él un predicado significativo debe referir a una esencia, la cual, a su vez, existe como el atributo común de un género de sustancias primeras. En los *Analíticos posteriores* (II, 7) dice Aristóteles que un nombre significa una esencia solo si es de clases naturales. Esto es, el significado de un nombre común puede ser considerado como correspondiente a su esencia que, a su vez, sería su referente solo si es el nombre de un género natural (Gómez Lobo, 1976). En todos los otros casos —nombres de clases vacías o ficticias, conceptos abstractos, objetos no naturales, etcétera—, como ellos no poseen una esencia, debemos admitir que tienen un significado fijado por convención. De esta manera, si la tesis de Kerferd es correcta, el presupuesto referencialista se mitigaría con Aristóteles hasta casi desaparecer.

4.3. Interpretación radical y sistemas de estados mentales

Dejemos ahora la teoría referencialista del significado y volvamos a la concepción del significado que surgiría del experimento conceptual de la interpretación radical. En el apartado 4.1 hemos visto que el análisis de lo que ocurre en esos casos extremos muestra aspectos presentes en situaciones menos improbables. Pero, en todos los casos,

la intérprete se ve obligada a modelar sistemas de estados mentales que atribuye al agente para poder hacerlo inteligible. Eso es lo que veremos en este apartado, al discutir cuatro principios que surgen del análisis del experimento conceptual de la interpretación radical, los cuales fueron planteados por Quine (1960) y desarrollados por Davidson (1984c).

4.3.1. La indeterminación del significado

Este principio se plantea en estos términos: cuando se produce la traducción de las proferencias verbales del lenguaje extraño al familiar, la intérprete elabora un conjunto de hipótesis de interpretación que constituirán una suerte de manual de interpretación. La indeterminación se plantea porque siempre es posible que existan varios manuales de interpretación diferentes entre sí que den cuenta de la misma información, de manera que es imposible determinar cuál de los manuales —esto es, cuál de los conjuntos de hipótesis— es el correcto, porque no existe tal cosa como un único manual correcto, ya que varios podrían ser al mismo tiempo correctos, en tanto pudieran ser capaces de iluminar distintos aspectos del agente. Esto ocurre porque la evidencia siempre subdeterminará los manuales de interpretación. Otra manera de decirlo es afirmar que no existe un único significado real o correcto de una proferencia verbal que sea nuestra tarea desentrañar, pues siempre será posible asignar diversos significados —en diversos manuales— a una acción o proferencia, dando sentido al tipo de conducta que pretende ser interpretado. La idea es, entonces, que puede haber varios manuales simultáneamente correctos que den sentido a distintos aspectos del comportamiento del agente. No es una posición escéptica sino pluralista acerca de la posibilidad misma de la interpretación.

Para Quine no hay un hecho o un *fact of the matter* de cuál es el significado de una expresión. Hay quienes piensan que al hacer esto Quine continuó el proceso de desconstrucción del significado que inició el segundo Wittgenstein y que Davidson posteriormente radicalizó. En efecto, se trata de una desconstrucción del significado en la medida en que rechaza la idea de que el significado de una expresión sea una entidad oculta que debamos desvelar, pero eso no tiene que conducir a suponer que los significados no existen. Tal como yo lo veo, los significados son propiedades relacionales emergentes de sistemas complejos conformados por comunidades de hablantes que se interpretan mutuamente. De no existir estas comunidades ciertamente no existirían los significados, pero en una comunidad de hablantes aquellos son tan reales como cualquier otra propiedad emergente como, por ejemplo, los adverbios y las preposiciones.

Quine aborda el problema del significado al utilizar la noción de «significado estimulativo». Piensa que es imposible distinguir el significado de una proferencia verbal

—*verbal utterance*— del conjunto de estímulos no verbales que causarían que la intérprete, ante la presencia de tales estímulos, asienta a la proferencia. Por ejemplo, sería imposible distinguir entre el significado de *gavagai* y los diversos estímulos no verbales que conducirían a la intérprete a asentir ante la pregunta: ¿es aquello *gavagai*? La indeterminación surge porque el asentimiento de la intérprete depende en parte de sus propias creencias previas respecto del objeto y el contexto, y en parte porque ella no podrá distinguir entre la información relevante y la información colateral.

4.3.2. La inescrutabilidad de la referencia

La indeterminación del significado conduce al principio de la inescrutabilidad —o indeterminación— de la referencia. Si no se puede fijar con precisión el significado de una expresión —o, lo que es lo mismo, diversas interpretaciones podrían ser igualmente válidas— tampoco se puede fijar con precisión a qué refiere la expresión o, lo que es lo mismo, podrían ser válidas diversas interpretaciones del referente. Si no existe manera de determinar el significado de un enunciado, porque la evidencia siempre subdetermina nuestras teorías, tampoco hay manera de determinar el referente de nuestras expresiones. Así la referencia no es una relación natural o independiente de una teoría. Los referentes de las expresiones son relativos a un manual de traducción, una teoría o un sistema de creencias. Así como los significados, los referentes no son *theory-independent*.

Más allá de Quine y Davidson, aunque en la bibliografía en lengua inglesa es común plantear el problema de la referencia en términos del objeto al que una expresión refiere —pues en inglés a *Word refers to*— en castellano no son las palabras las que refieren sino los hablantes, y además el verbo «referir» es reflexivo. En castellano, por tanto, diría «Yo me refiero a esto o a aquello» y «Tú te refieres a lo de más allá». Un modelo de interpretación y referencia como el que defiendo está en mayor sintonía con el uso castellano que con el del inglés, pues la intérprete intentará precisar —con un inevitable grado de indeterminación— los objetos a los que el hablante se refiere con sus palabras, no los objetos a los que las palabras refieren, porque las palabras no refieren por ellas mismas a nada.

La indeterminación del significado y la indeterminación de la referencia tienen dos consecuencias importantes: la relatividad ontológica y la inseparabilidad entre significado e información.

4.3.3. El principio de la relatividad ontológica

Este principio, planteado originalmente por Quine (1969), sostiene que no podemos fijar con precisión de qué objetos consta la realidad sino de qué objetos habla una teoría, es decir, cuáles son los referentes de las expresiones que tienen significado

en una teoría dada. Así, la relatividad ontológica establece que, en el mejor de los casos, podemos fijar cuáles son los objetos de los cuales habla una teoría, esto es, podemos fijar los «compromisos ontológicos» de una teoría o sistema de creencias. Lo que no podemos hacer es establecer una ontología, decir cuáles son los objetos que existen, independientemente de una teoría. De esta manera, toda teoría está comprometida ontológicamente y toda ontología es relativa a una teoría.

4.3.4. El principio de la inseparabilidad entre significado y creencia

Se interpreta simultáneamente creencias y significados, y la distancia entre creencias es también distancia entre significados. Tener creencias diferentes acerca de un objeto es atribuir significados diferentes a las expresiones que describen el objeto, y atribuir a una palabra cierto significado equivale a tener creencias semejantes acerca del objeto al que la palabra refiere. Esta intuición —o una parecida a ella— estaba presente en Cratilo cuando decía que conocer el nombre es conocer la cosa, esto es, que tener creencias correctas acerca del significado de la cosa es tener creencias correctas acerca de la cosa. Mucho más recientemente Quine lo planteó de esta manera:

> El problema de separar significado de creencia es uno que me ha impresionado como central. He sentido que no hay esperanza, en general, de separar las creencias de una comunidad en verdades que pertenecen a los significados de las palabras y verdades que uno quisiera ver como información colateral compartida (1974, p. 325).

La manera en que Davidson plantea la inseparabilidad entre creencia y significado es esta: al interpretar a un hablante es imposible conocer cuáles son sus creencias si no conocemos previamente cuáles son los significados que él atribuye a las palabras y, de otro lado, es imposible saber cuáles son los significados que él atribuye a las palabras si no conocemos previamente sus creencias.

Pero la interdependencia entre significado y creencia desarrollada por Quine y Davidson es, en realidad, mucho mayor y requiere de una ampliación. Incluye otras actitudes proposicionales como deseos, temores, propósitos, aspiraciones, acciones, etcétera, y la única manera de atribuir cualquiera de estas actitudes proposicionales a un agente es atribuyéndole una red de muchas otras actitudes proposicionales, es decir, de actitudes que uno puede tener respecto de una proposición. Así, frente a una proposición dada, por ejemplo, que «Estambul es una hermosa ciudad», uno puede creerla, desearla, esperarla, temerla, añorarla, recordarla, etcétera. Como ya hemos visto, esto es lo que se suele llamar el holismo de lo mental: la tesis de que lo que llamamos «la mente» o «el ámbito de lo psicológico» está constituido por una red grande y compleja de actitudes proposicionales, en la que todas están interconectadas y no puede haber una actitud proposicional o un estado mental desconectado o independiente de los otros.

No obstante, es importante recordar que los conceptos de «creencia», «referencia», «significado», etcétera, son construcciones culturales que tienen como objetivo comprender la conducta de los agentes. En última instancia, interpretar a un agente es producir una descripción lo más compleja y precisa posible de un sector de esta red de estados mentales que, a su vez, está conectada causalmente con sus acciones. Como esa red es tejida desde los propios estados mentales de la intérprete, siempre hay cierto grado de indeterminación, porque, en principio, es posible elaborar distintas descripciones alternativas —eventualmente excluyentes, pero nunca inconmensurables— de las acciones y los estados mentales de un agente en las que se atribuyen estados mentales distintos, pero conservando la coherencia del agente y de la interpretación.

Cuando interpretamos a un agente le asignamos actitudes frente a proposiciones, lo que nos obliga a preguntarnos qué son estas. Una proposición es el significado o contenido informativo que una oración —o un conjunto de oraciones sinónimas— transmite. Pero no debe suponerse que las proposiciones tienen existencia independiente de los hablantes, como algunas posiciones de inclinación platónica sostienen. Expresar una proposición mediante una oración es solo representar el mundo de una manera parecida a como muchas otras personas también lo harían. Y hablar de una proposición —por ejemplo de la proposición «Estambul es una hermosa ciudad» y decir de ella que es verdadera, justificada o incomprensible— es solo una manera resumida de describir el comportamiento regular de muchas personas que interactúan entre sí y que se comportan respecto de Estambul de una particular manera, que sería muy diferente si creyeran que Estambul es una ciudad anodina.

Las proposiciones, en tanto son los significados que las intérpretes asignan a las acciones y proferencias de los agentes en relación con la realidad compartida, no son entidades que existan independientemente de ellos —intérpretes, agentes y realidad objetiva—; tampoco son entidades que existan en las mentes de las personas. Son propiedades relacionales triádicas conformadas en una situación interpretativa triangular que incluye a intérpretes, agentes y el mundo que habitan. Análogamente, el lenguaje psicológico o mental es una forma resumida de describir relaciones sociales e intersubjetivas estructuradas pero muy complejas. Así también, los estados mentales con contenido proposicional son solo actitudes que las personas tienen respecto de las formas de comportamiento regular de otras personas.

Los estados mentales, en general, tampoco son entidades que existan dentro de los cerebros de las personas, aunque ciertamente un estado mental cualquiera es una descripción psicológica de un evento neurológico que existe en el cerebro de alguien.

Pero en tanto estado mental y no físico —es decir, cuando describimos el aspecto mental de un evento que también tiene un aspecto físico— lo que hacemos es describir una manera en que una persona puede «estar», no una cosa que la persona «tiene». En efecto, un estado mental, como lo dice la expresión, es una forma de «estar en el mundo» para aludir a la célebre traducción del *Dasein* heideggeriano (1986), un «modo de ser» o, mejor expresado —gracias a una distinción que el castellano tiene, pero no otras lenguas europeas—, «un modo de estar».

El punto, entonces, es que es imposible distinguir entre lo que creemos acerca de los objetos —las representaciones y disposiciones para actuar que tenemos respecto de ellos— y nuestro uso de las expresiones que pretendidamente refieren a estos. En general, nuestro conocimiento de los objetos es inseparable de la manera en que los describimos y nos comportamos en relación con ellos.

Podríamos formularlo de esta manera: el proceso de adquisición del lenguaje, es decir, el proceso del aprendizaje de aquellas prácticas sociales que gobiernan el uso que hacemos de ciertas expresiones o acciones, es inseparable del proceso del conocimiento de la realidad, esto es, del proceso de desarrollo de creencias acerca del mundo y de disposiciones para actuar en el mundo. Dicho de otra manera: aprendemos el lenguaje, desarrollamos una visión del mundo y aprendemos a actuar en él al mismo tiempo; se trata de un mismo y único proceso. Así, el aprendizaje del significado de las palabras, de nuestras creencias acerca del mundo, de nuestras creencias acerca del lenguaje y de nuestras acciones al interior de la comunidad a la que pertenecemos constituyen un solo proceso que es la conformación de nuestro sistema de estados mentales.

Podría ocurrir que ontogenéticamente los niños vayan adquiriendo conceptos o habilidades en diferentes estadios de su desarrollo. Por ejemplo, hay evidencia que sugiere que a los dos años los niños comprenden conceptos de deseo; a los tres, de conocimiento y pensamiento; y a los cuatro, de creencia. No está plenamente demostrado que así sea pero es factible. Lo importante es que en todo momento la vida mental de las personas está hecha de tal manera que cada estado mental define sus contenidos en relación con los otros estados mentales en ese momento existentes.

Quine fue uno de los primeros en acuñar la noción de «sistema de creencias» (1963, 1970), porque su interés prioritario era sostener que las teorías científicas se enfrentan a la evidencia empírica como un todo y no cada oración con un hecho del mundo. Con ello estaba defendiendo un modelo epistémico holista y no atomista, y de la misma manera como el segundo Wittgenstein transitó hacia el holismo desde el atomismo del *Tractatus*. Por ello, Quine sostenía que los enunciados internos al sistema de creencias fijan los significados de los enunciados externos, aunque todos los enunciados están integrados por relaciones inferenciales.

Sin embargo, si utilizamos esa noción para entender procesos psicológicos, es claro que no debemos hablar solo de sistemas de creencias sino de sistemas de estados mentales en general, pues la vida psíquica de una persona no se compone únicamente de creencias, y los estados mentales no epistémicos también influyen en la determinación de los contenidos y la fijación de las creencias. En lo que sigue, por tanto, expondré la concepción quineana de sistema de creencias para luego defender una concepción más amplia de sistemas de estados mentales, que me parece preferible por ser más comprehensiva.

Usaré un segundo experimento conceptual también inspirado en Quine (1963). Imaginemos que pudiéramos hacer algo que ciertamente no es posible hacer ahora y que probablemente nunca lo sea, y no solo por un asunto empírico sino también conceptual. Supongamos que pudiéramos conocer las creencias que un individuo o un grupo de individuos tienen en un momento de su historia. Al hacer la lista de estas creencias descubriríamos que ese conjunto está conformado como una esfera. En el centro de ella están las creencias en las que el individuo tiene más convicción y en la periferia aquellas en las que el individuo tiene menos. Las creencias internas —aquellas en las cuales uno tiene más convicción— implican lógicamente un gran número de creencias externas, de suerte que la eliminación o reubicación en el sistema de una creencia interna conduce a la eliminación o reubicación de un grupo numeroso de creencias externas, lo que implica la transformación y, eventualmente, desestabilización del sistema. Un análisis parecido a este y que podría haber influido en Quine se puede encontrar en el artículo «The Fixation of Beliefs», escrito en 1877 por Peirce (*CP* 5.358-387).

Para Quine las creencias internas pueden verse como enunciados analíticos, es decir como aquellos cuyo valor de verdad solo depende del significado de sus términos; y las externas como enunciados sintéticos, es decir, como aquellos cuyo valor de verdad depende del significado de sus términos y de la manera en que es el mundo. Naturalmente, no hay una línea fronteriza clara sino un continuo entre ambos tipos de creencias y oraciones, pues siempre habrá oraciones más analíticas y sintéticas que otras. El núcleo de creencias más internas son los principios de la lógica y el principio de no contradicción. Las más periféricas, son enunciados acerca del mundo externo que nos resultan suficientemente triviales como para poder ser eliminados sin que se altere gravemente el sistema. En el caso de una teoría científica, los enunciados observacionales suelen ser periféricos y los teóricos suelen ser centrales.

Otro punto importante es que las oraciones analíticas, análogas a aquellas a las que Carnap llamaba «reglas de significado», fijan el significado de las oraciones sintéticas. Las oraciones analíticas son, en realidad, explicaciones de los significados de los conceptos. Por ejemplo, la oración analítica «Los cuadrúpedos tienen cuatro patas»

fija y explica el significado de la oración sintética «Los caballos son cuadrúpedos», con lo cual las oraciones internas entrelazadas con las oraciones externas constituyen la estructura de nuestro lenguaje y nuestra visión del mundo.

En los casos individuales, la misma creencia puede ser interna para una persona y externa para otra, y una creencia puede cambiar su lugar en el sistema en tanto cambien de lugar las otras creencias que le dan contenido. Así, por ejemplo, en mi caso la creencia «La democracia es mejor que la dictadura» es interna, pues necesitaría mucha evidencia para abandonarla, además de que su abandono implicaría una transformación marcada de mi sistema y, por tanto, de mi comportamiento, porque hay que recordar que una creencia también es una disposición para actuar. De otro lado, mi creencia que «El *Homo sapiens* abandonó África hace aproximadamente cien mil años» es intermedia, porque nueva evidencia podría hacerme corregirla, aunque eso me obligaría a revisar, y tal vez abandonar, muchas otras creencias que son importantes para mí, es decir en las que creo y me resulta importante creer porque, a su vez, implican otras creencias que tengo. Finalmente, mi creencia que «La bandera mexicana es más antigua que la italiana» es relativamente periférica. Aunque la creo, podría aceptar evidencia que mostrara que es falsa, con lo que tendría que abandonarla. Pero ese abandono no sería particularmente traumático en mi caso, aunque sí podría serlo para un historiador mexicano. Análogamente la creencia en que «La democracia es mejor que la dictadura» es asumida por muchas personas, pero varias de ellas estarían dispuestas a abandonarla fácilmente ante lo que podrían considerar evidencia relevante. Y la oración «El *Homo sapiens* abandonó África hace aproximadamente cien mil años» simplemente no constituye ninguna creencia para mucha gente. Esas personas podrían escuchar la oración en un documental y olvidarla rápidamente, de manera que la creencia no se llegaría a fijar porque no estaría conectada con otras creencias importantes para esos sujetos. También podría haber casos en que la oración genere suficiente sorpresa como para que la creencia se fije de manera periférica, pero en el curso de dos horas las personas podrían confundirse y creer que «El *Homo sapiens* abandonó África hace aproximadamente cien millones de años», creencia que sería rápidamente modificada ante nueva información pertinente sin que el sistema de creencias sea mayormente afectado.

Hasta ahí tenemos la concepción quineana de sistema de creencias. Pero como señalé hace un momento, no es posible separar creencias de deseos, afectos, valores y otros estados mentales, de manera que usaré la expresión «Sistema de estados mentales». En un sistema de estados mentales, estos conforman una estructura holista que tiene relaciones lógicas, epistémicas y semánticas, de suerte que cada estado mental depende de los otros para la constitución de su contenido y participa en la constitución del contenido de los demás. Veamos ahora algunas formas de holismo.

El holismo lógico sostiene que los estados mentales tienen relaciones inferenciales, de suerte que cada uno implica lógicamente a muchos otros y es implicado por otros tantos. Por ejemplo, desde un punto de vista formal, si creo que p y también creo que p implica q, es altamente probable que crea que q. Si no lo creo es porque no me he dado cuenta de la relación que hay entre esas creencias o porque hay otro estado mental que está ejerciendo mayor presión. Desde un punto de vista material, la creencia «Álvaro es presidente del Perú» implica que el Perú es una república y no un virreinato, y esto último implica que el Perú no tiene un virrey. Esto es independiente, por supuesto, de si «Álvaro es presidente del Perú» es una oración verdadera o no. Mi creencia que «El Perú es una república y no un virreinato» es central a mi sistema, es decir, sería muy difícil que la abandonara porque eso implicaría modificar todo mi conocimiento de la historia del Perú; pero un neozelandés podría abandonarla sin mayor problema, lo que probaría que para él es una creencia periférica. De igual manera, mi deseo de que los países latinoamericanos vivan en democracia implica que deseo que un país latinoamericano en particular viva en democracia. No podría desear lo primero y no lo segundo, a menos que tenga otros deseos, creencias o afectos que ejerzan presión para que así sea. Asimismo, mi emoción de repulsión por los dictadores implica que, si creo que x es un dictador, entonces tengo una emoción de repulsión por x. Si no la tengo es porque hay otros estados mentales que no lo permiten, por ejemplo, tengo un particular afecto personal por x, a quien conozco desde pequeño, o deseo que x termine un importante proyecto bibliográfico en el que está empeñado, o creo que en este particular momento x es preferible a cualquier otra opción disponible.

El holismo epistémico mantiene que no es posible poner a prueba —es decir corroborar, justificar o falsar— estados mentales de manera aislada, sino solo dentro de un subsistema mayor. Así, por ejemplo, si creo que las aves evolucionaron a partir de reptiles, deberé comprobar si mi creencia es consistente con las otras creencias que conforman la teoría de la evolución y no preguntarle a un oráculo, pues yo creo en la validez de las teorías de Darwin y no en los oráculos. Análogamente, si estoy irritado ante x porque creo que es autoritario, pero alguien me prueba que no lo es, mi irritación desaparecerá, pues consideraré que no tengo evidencia para mi irritación. Si no desapareciera, probablemente será porque mi emoción tiene una causa diferente de la que yo creo que tiene. Supongamos que considero que la democracia es un valor y deseo que esta se mantenga en cierta institución, pero al mismo tiempo creo que eso implica hacer ciertas modificaciones estructurales. Se sigue que desearé hacer tales modificaciones siempre que ellas no entren en conflicto con otras creencias, deseos, afectos y valores que también tengo. Reconoceré la evidencia a favor o en contra de mis estados mentales a partir de los otros estados mentales que tengo y que me permitirán interpretar la evidencia de que dispongo.

El holismo semántico afirma que el contenido proposicional de una oración y el significado de una expresión dependen de sus relaciones con el resto del sistema pues, como sostengo, el significado no es una propiedad monádica de las expresiones sino relacional. Supongamos, por ejemplo, que tanto A como B creen que Aristóteles escribió la *Metafísica.* Pero mientras A cree que «Aristóteles» fue un filósofo nacido en Estagira en 384 a.C., B cree que fue un armador y multimillonario griego del siglo XX. Adicionalmente, A cree que la *Metafísica* versa sobre la naturaleza del ser, mientras que B cree que es un poema. Ambos tienen la creencia que Aristóteles escribió la *Metafísica*, pero no creen lo mismo y, para los efectos del holismo semántico, ambas oraciones tampoco significan lo mismo. Así, cuando A y B dicen que «Aristóteles escribió la *Metafísica*» sus oraciones no portan la misma proposición, es decir, no significan lo mismo, y A y B no tienen la misma creencia.

Para algunos autores el estado mental de creer tiene cierta prioridad epistémica sobre los demás. La razón de ello es que la creencia involucra representaciones del mundo y es portadora de valores de verdad. Adicionalmente, se puede formular otros estados mentales en términos de creencias, como cuando Aristóteles en el segundo libro de la *Retórica* (2002) define e individualiza las emociones en términos de creencias diciendo, por ejemplo, que el miedo es la creencia en que un daño puede sobrevenirnos. Por otra parte, como ya mencioné, las creencias suelen moldear a otros estados mentales.

Examinemos ahora brevemente el concepto de creencia. ¿Qué decimos de una persona cuando afirmamos que cree algo? Sostengo que S tiene la creencia que p si y solo si:

(1) S tiene la disposición para comportarse como si p[2] y para aceptar proposiciones que se siguen de p. Es decir, S tiene la disposición para comportarse como si p fuera verdad y para creer otras proposiciones que se deducen de p asociadas a las otras creencias de S. Este es el elemento disposicional o conductual.

(2) S tiene la representación mental que p es verdadera, es decir, su cerebro está en una configuración tal que S se figura que p. Este es el elemento representacional o cognitivo.

(3) S sabe que p podría ser falsa, es decir, S tiene el concepto de creencia y sabe que es consustancial a tener una creencia el hecho de saber que esta pueda ser verdadera o falsa. Este es el elemento conceptual, pues S no solo tiene

[2] La definición de creencia como disposición para actuar es común entre los filósofos pragmatistas y se puede encontrar ya en Peirce, pero su primer defensor fue el filósofo y psicólogo escocés Alexander Bain (1859, pp. 569-570).

> la creencia que p, sino además tiene el concepto de creencia. Este rasgo es particularmente importante, porque si uno lo asume tendrá que aceptar que solo es posible atribuir creencias a una criatura que tenga por lo menos dos niveles de intencionalidad, de tal manera que tenga la creencia que p y también la creencia que p podría ser falsa. Así, en muchos casos, S cree que p y S sabe que cree que p. Aunque también hay casos en que S cree que p y no lo sabe, pues podría tratarse de una creencia no consciente o inconsciente —en un sentido psicodinámico—. Pero, incluso en esos casos, S sabe qué es tener una creencia, pues tiene el concepto de ella.

El cuestionamiento que hizo Quine (1963) a la distinción entre oraciones analíticas y sintéticas conduce a sostener que no se puede separar con claridad entre cuestiones de significado y cuestiones de hecho o, expresado de otra manera, entre nuestras creencias acerca de las cosas y nuestras creencias acerca de los significados con que describimos las cosas. A esa distinción llamó Quine el segundo dogma del empirismo. Davidson (1984d [1974]) fue más lejos y bautizó como el tercer dogma del empirismo a la distinción entre esquema conceptual y contenido que, como vimos en el tercer capítulo, hace posible la inconmensurabilidad, el escepticismo y el relativismo. John McDowell fue aún más lejos y afirmó que el cuarto dogma del empirismo es seguir manteniendo la distinción entre hecho y valor, de manera que no solo cuestionó la posibilidad de hablar sobre hechos sin hacer algún tipo de valoración de ellos sino, en general, sostuvo que toda descripción implica algún tipo de valoración de lo que se está describiendo. En efecto, eso es lo que ocurre en un sistema de estados mentales, pues los juicios de valor están integrados y entrelazados con las representaciones del mundo.

Teniendo en cuenta las características holísticas que tiene un sistema de estados mentales, analicemos ahora cómo se constituyen y cómo cambian tales estados, de manera integrada, a lo largo del tiempo. Pensemos en Galileo y examinemos su creencia, durante su época geocéntrica, de que «La Tierra es el centro del universo». A lo largo de un período de su vida, esta creencia fue central a su sistema y constituía una oración analítica, pues en este se definía «Tierra» precisamente como «el centro del universo». Esta creencia, además, estaba probablemente asociada a ciertos afectos —por ejemplo, el cariño y gratitud que tenía a los profesores que se la enseñaron—, ciertos deseos —como el de continuar la obra de sus maestros— y algunos valores —como el suponer que vivir en el centro del universo nos hace más importantes al interior de la creación—. Esta creencia, a su vez, implicaba y estaba implicada en otras creencias del sistema de estados mentales de Galileo. Por ejemplo, implicaba la creencia que uno camina en un objeto inmóvil que está bajo sus pies y que es el abajo absoluto. Imaginemos también que Galileo sentía particular rechazo por los modelos

heliocéntricos de su época porque tenía un conflicto personal con el defensor de uno de ellos.

El estudio de modelos heliocéntricos y la observación cosmológica condujeron a que Galileo considerara datos empíricos que eran incompatibles con su sistema. Al comienzo seguramente pensó que tales datos eran solo anomalías o casos rebeldes a la explicación. En otras palabras, su sistema no admitió la nueva información o, lo que es lo mismo, tales datos observacionales no fueron suficientemente fuertes como para generar creencias que contradijeran algunas otras centrales de su sistema. Que un estado mental no sea suficientemente fuerte significa que no está asociado a un número suficiente de estados mentales adicionales, como para que pueda eliminar fácilmente a un estado mental más débil.

Progresivamente, sin embargo, los datos fueron tantos en número y tan consistentes entre sí, que parecía inevitable modificar el sistema geocéntrico. De esta manera, después de un largo proceso de deliberación consciente y de reordenación no consciente de su sistema, llegó el momento en que Galileo se convirtió al heliocentrismo. Cuando eso ocurrió, su sistema ya había sufrido modificaciones sustanciales para que pudiera admitir las nuevas creencias. Muchas creencias, tanto internas como externas, fueron abandonadas para dar lugar a creencias nuevas y contradictorias con las anteriores. Algunas internas pasaron a ser externas y viceversa.

También hubo modificación en sus deseos, afectos y valores. Pero no debe creerse que el cambio de creencias ocasionó el cambio en los otros estados mentales. El cambio se da de manera integral, de manera que lo que pudo haber comenzado con un cambio en el nivel de deseos o de afectos pudo haber generado un cambio epistémico.

En el nivel epistémico es importante notar que todo este proceso de transformación modificó la ubicación y el contenido de casi todas las creencias cosmológicas de Galileo. Pero, no solo eso: si antes la oración «La Tierra es el centro del universo» era verdadera y analítica, ahora es considerada por él como falsa y sintética. En relación con los otros estados mentales de Galileo, es ampliamente probable que también haya habido variación en sus deseos, afectos y valores.

Como ya hemos visto, la manera en que las creencias están interconectadas entre sí varía entre individuos y también varía a lo largo de la vida de un individuo, pues los sistemas de estados mentales están siempre cambiando, con lo cual una misma oración puede ser interna para un sujeto y externa para otro, lo cual hace que, en cierto sentido, el significado atribuido a esa oración y sus consecuencias pueda diferir entre sujetos. De igual manera, uno puede tener deseos, valores o afectos internos —esto es, difíciles de eliminar— o periféricos —fáciles de eliminar—. Los primeros son fuertes y los segundos débiles. Así, por ejemplo, mi deseo de agradar a mi madre es mucho más importante que el de agradar a mi vecina, de manera que si ambos

deseos se contraponen, ciertamente desearé lo primero y actuaré según ello. No es que elija desear lo primero o que desee desear lo primero —lo que se llama un metadeseo—, sino simplemente desearé lo primero. De igual manera, si el valor de la honestidad es para mí más fuerte que el de la lealtad, no dudaré en cuestionar el comportamiento de un amigo si lo considero inapropiado. Puedo imaginar, no obstante, que haya personas para quienes el valor de la lealtad prime sobre el de la honestidad —en ellos el primero es más interno que el segundo, en sus respectivos sistemas de estados mentales—, de manera que considerarán inaceptable que alguien atestigüe contra un amigo ante un jurado.

Ahora bien, este tipo de significado del que estamos hablando —lo que una expresión significa al interior del sistema de estados mentales de un individuo— es lo que ese hablante cree que la palabra significa. Por ejemplo «centro del universo», si es geocentrista, o «planeta», si es heliocentrista. Y podríamos llamar a estos significados «primarios» o «literales» —*first meaning* o *literal meaning*— siguiendo la distinción que hace Davidson (2005a [1986]). Por otro lado, el «significado intencional» o «significado del hablante» —*intended meaning* o *speaker's meaning*— será la manera en que un hablante en particular desea usar el significado primario, en una ocasión en particular, para comunicarse con una intérprete específica, es decir, para producir ciertos efectos deseados en ella.

Finalmente, el «significado convencional» —*conventional meaning*— es lo que aparece registrado en los diccionarios, esto es, el uso regular en las prácticas comunicativas de una comunidad de hablantes. Puede decirse que el significado convencional de una expresión, en una comunidad de hablantes dada, es la intersección formada por la superposición de los diversos significados primarios de los hablantes promedio de esa comunidad.

Entonces, aunque el significado que alguien atribuye a una expresión en una ocasión dada —significado intencional— puede variar según las características de un encuentro comunicativo particular, los significados primario y convencional tienen cierta autonomía respecto de las intenciones particulares de uso. Este es el principio que Davidson denomina de la «autonomía del significado» (2005a [1986]), pues sostiene que ese significado —ese uso regular de la expresión en una comunidad de hablantes—, se mantiene más o menos estable, aunque varíen las intenciones de uso particulares de los hablantes. No obstante, es claro que si las intenciones de uso y las creencias de los hablantes particulares varían lo suficiente, también variará el significado primario, y si los significados primarios del promedio de hablantes cambian, cambiará también el significado convencional.

Así comprender las proferencias de un hablante requiere, primero, comprender el significado primario de sus expresiones, para después imaginar cuál es la intención

de uso del hablante en una circunstancia particular y concreta, y en relación con una intérprete específica. Es más, aunque el significado primario de una oración está determinado por su ubicación en un sistema, como este se adquiere en un proceso de socialización y aprendizaje del lenguaje en que lo privado termina constituyéndose a partir de lo público, los sistemas no pueden ser demasiado diferentes unos de otros, de manera que hay importantes sectores de ellos que son comunes. En consecuencia, los significados primarios de hablantes que pertenecen a la misma comunidad no pueden ser demasiado diferentes. Por otra parte, aunque en algunos aspectos los sistemas de estados mentales de culturas diferentes podrían estar muy distantes entre sí, tendrán también muchos sectores comunes que proceden de la adaptación natural al medio, las condiciones de supervivencia, las funciones biológicas, etcétera. En todo caso, es claro que un sistema de estados mentales está constituido por una multiplicidad de subsistemas, los que guardan múltiples relaciones entre sí de intersección, inclusión y exclusión. Más aún, en un sistema no solo hay estados mentales en primer grado sino también en varios grados de intencionalidad.

Se da una dialéctica entre creencias internas y externas, lo que hace que las internas fijen el contenido de las externas y que sean el criterio de admisión o eliminación de creencias al sistema. En general, el criterio es la coherencia con las creencias ya admitidas y, especialmente, con las internas. Sin embargo, un número alto de creencias externas puede eliminar o incorporar un número reducido de creencias internas al sistema. Como en el caso de Galileo, lo que en un momento es una anomalía puede convertirse, con suficiente insistencia, en una creencia externa o incluso interna. Esto nos recuerda la dialéctica entre teoría y observación que ha mostrado la filosofía de la ciencia reciente, tal como vimos en el capítulo anterior cuando analizamos a Kuhn.

Todo sistema está conformado por subsistemas, algunos de los cuales tienen relaciones de incompatibilidad, lo que da lugar a estados de división o partición de la mente. En esto consistirá la irracionalidad, como veremos más adelante. Cuando un subsistema comienza a guiar el comportamiento del agente, los otros pueden pasar a estado de latencia o, incluso, de ocultamiento. Pero es importante recordar que como una creencia es una representación del mundo y una disposición para actuar, un sistema de estados mentales incluye una red interconectada de representaciones y de disposiciones para actuar, esto es, de prácticas sociales, formas de vida o mundos de la vida. Es la posesión de un sistema de estados mentales lo que posibilita la acción, y las tensiones entre subsistemas al interior del sistema mayor se manifiestan y resuelven en última instancia en la conducta del individuo.

Muchas de nuestras creencias son conscientes, es decir, podemos dar cuenta de ellas porque tenemos otras creencias y estados mentales acerca de ellas, pero también las hay inconscientes y no conscientes. Podríamos incluso ver las creencias como

conformado un continuo, tal que en un extremo tenemos a aquellas de las que tenemos conciencia y en el otro a disposiciones no conscientes que virtualmente se identifican con funciones biológicas.

Retomando el tema de la indeterminación del significado, la tesis de Quine es que las experiencias sensibles no están asociadas a oraciones individuales sino, en todo caso, a redes de oraciones. Piensa él que conduce a confusión hablar del contenido empírico de una oración en particular (1963, 1970), pues cualquier oración puede ser sintética o analítica con solo hacer cambios suficientemente drásticos en el sistema. Todo esto está asociado a la tesis de que la experiencia siempre subdetermina la teoría, es decir, que los mismos datos empíricos pueden ser interpretados de múltiples formas y, entonces, pueden corroborar teorías incompatibles entre sí. De igual manera, cualquier oración puede ser observacional o teórica con solo hacer los cambios necesarios en el sistema de creencias.

En definitiva, la tesis de Quine es que no hay un hecho —*a fact of the matter*— de cuál es el significado «real» de una oración o de las creencias expresadas por ella, pues esto depende del tipo de traducción que hagamos, lo que no significa que las expresiones carezcan de significado, pues varias hipótesis de interpretación pueden asignar simultáneamente interpretaciones correctas, aunque diferentes, de una proferencia verbal o de una acción intencional.

En este punto hay algo importante que debe ser señalado acerca de la ontología de los sistemas de estados mentales. Uno podría suponer que estos sistemas son realidades que existen en los cerebros de las personas y que la tarea de las intérpretes es revelarlos y explicitarlos, poniéndolos en palabras. Hay algo de cierto en ello, pero con algunas cualificaciones. En primer lugar, los cerebros están conformados por neuronas y procesos electroquímicos que conforman redes de mucha complejidad. En cierto sentido, en los cerebros no hay creencias, deseos ni emociones. Los conceptos de estados mentales son construcciones culturales, transmitidos por nuestras lenguas, que nos permiten dar sentido a los hablantes. Esto implica que en sentido literal no hay sistemas de estados mentales en los cerebros de las personas sino que estos sistemas son construcciones que hacemos para comprendernos mutuamente. Es altamente probable que, así como el cerebro humano ha evolucionado para asumir que el universo es regular y no aleatorio, también ha evolucionado para asumir que las otras personas tienen estados mentales que conforman sistemas, de manera que aplicamos este principio para entendernos. Así, es factible que diferentes intérpretes atribuyan distintos sistemas de estados mentales al mismo individuo estando ambas acertadas, pues cada una de ellas podría capturar aspectos diferentes del agente. Naturalmente podría también ocurrir que ambas atribuyan sistemas contradictorios entre sí, en cuyo caso habría que ver cuál de ellos resulta más explicativo, es decir, cuál cumple

con más virtudes epistémicas como para que resulte preferible al otro. Todo esto es parte del principio de la indeterminación de la interpretación según el cual la misma evidencia podría dar lugar a la construcción de distintos sistemas de estados mentales, los cuales podrían ser simultáneamente correctos.

Pero para que la intérprete pueda siquiera comenzar a interpretar al hablante, debe asumir el principio de caridad. Volveremos sobre este principio en varias ocasiones, de manera que ahora lo introduzco rápidamente. Aunque Quine lo explicita (1960), este se encuentra por vez primera en la obra de Wilson (1959, 1970), pero es Davidson (1980b, pp. 222, 290; 1984c, pp. 27, 101, 136) quien lo desarrolla y extrae muchas de sus consecuencias. Este principio sostiene que, para que la intérprete pueda comenzar a interpretar al hablante, ella debe asumir que comparte un gran número de creencias y significados con él, es decir, que, en líneas generales, ambos asignan los mismos significados a las palabras y tienen más o menos los mismos objetivos ante las mismas circunstancias. Davidson lo pone de esta manera: la intérprete debe asumir que el hablante es «consistente, un creyente de verdades y un amante del bien. Todo bajo nuestros propios criterios, es innecesario decirlo» (1980b, p. 222).

Algunas personas han interpretado el principio de caridad como si afirmara que, para que haya comunicación, es necesario que hablante e intérprete compartan previamente un numeroso conjunto de creencias. Eso es básicamente correcto, pero debe ser cualificado. La tesis es que la intérprete debe asumir que la mayor parte de creencias del hablante son semejantes a las suyas propias pero, además, dado que ambos comparten la misma realidad objetiva, nosotros en tanto filósofos que queremos explicar el fenómeno de la comprensión debemos asumir que comparten un número muy grande de creencias entre sí y con nosotros también, las cuales deben ser verdaderas. Es decir, debemos asumir que compartimos un conjunto numeroso de creencias verdaderas y no podríamos no hacerlo. Más adelante mostraré que hay un argumento trascendental que justifica esa afirmación.

Davidson piensa que el fenómeno del aprendizaje del lenguaje, sobre la base de una realidad común, garantizaría que hay un conjunto de creencias que debemos compartir. Pero esta afirmación no debe verse como planteada desde un punto de vista privilegiado *sub specie aeternitatis*, sino como presupuesta en los puntos de vista de los diversos interlocutores, incluyéndonos a nosotros mismos. Esa es la tendencia de la filosofía de Davidson: plantear las cosas desde la perspectiva de los interlocutores en las diversas situaciones comunicativas, más que de los espectadores no comprometidos, privilegiados y solitarios. Pero, aunque ambos interlocutores compartan creencias previas, lo imprescindible para que pueda haber comunicación entre ellos es que en la interacción puedan desarrollar creencias compartidas.

Esta diferencia de matiz es sutil pero en extremo relevante. Recuérdese que en «Interpretación Radical» Davidson (1984e [1973], pp. 125-6) dice que lo importante para una teoría del significado no es lo que de hecho sabemos que nos permite comprender al hablante, sino lo que tendríamos que llegar a saber para entenderlo y, además, cómo es que llegamos a saberlo.

Veamos ahora el análisis que hace Davidson del proceso de interpretación una vez que la intérprete lo ha iniciado. Este autor distingue entre «teorías previas» (*prior theories*) y «teorías al paso» (*passing theories*)[3], en un artículo titulado «A Nice Derangement of Epitaphs», publicado originalmente en 1986 (2005a). La teoría previa de cada uno de los interlocutores está constituida por las creencias que él o ella tiene acerca del otro, esto es, acerca de lo que es y lo que cree, antes de comenzar la interacción comunicativa. Al comenzar a interactuar, descubrirá que algo o gran parte de su teoría previa es falso, es decir, que no permite entender mejor al otro o que lo desfigura; entonces se verá obligado a reformular su teoría previa. El producto de esta reformulación es una teoría al paso, para un específico interlocutor.

Una teoría previa es un pequeño sistema de estados mentales construido para dar sentido a un hablante o a una intérprete en particular. Cuando interpretamos a varias personas al mismo tiempo —que es lo que hacemos cotidianamente— construimos simultáneamente varias teorías previas, esto es, pequeños sistemas de estados mentales, que aplicamos diferenciadamente a cada uno o una de nuestros interlocutores. Supongamos que la intérprete está conversando con otras tres personas al mismo tiempo. Ella construirá pequeñas teorías previas para cada uno de ellos. En cada una de esas teorías habrá estados mentales que ella atribuye a los hablantes acerca del mundo, acerca de ella, acerca de los otros hablantes, acerca de los estados mentales que la intérprete y los hablantes tienen unos de otros y así en adelante, hasta en cuatro, cinco o seis niveles de intencionalidad, que es nuestro límite cognitivo.

Una teoría al paso, por su parte, es el producto de la reformulación, corrección o afinamiento de una teoría previa para dar sentido al nuevo comportamiento del interlocutor o a la nueva evidencia que tenemos de él o ella. La teoría al paso modificará el sistema de estados mentales de su portador, es decir su visión del otro, pero también su visión de él o de ella misma, preparándolo o preparándola para producir teorías al paso más creativas en otras interacciones comunicativas. En el siguiente encuentro, la teoría al paso del encuentro anterior oficiará de teoría previa para dar lugar a una nueva teoría al paso y así el proceso comunicativo continuará

[3] Desde hace muchos años traduzco *passing theories* por «teorías al paso» (Quintanilla, 1997), con lo que sugiero que se trata de teorías elaboradas rápidamente y sobre la marcha, como cuando un peón toma otra pieza «al paso» en una partida de ajedrez. Posteriormente otros traductores han usado «en paso», pero me parece que el castellano soporta mal esa expresión.

ininterrumpido e inacabable. Hay una teoría previa y una teoría al paso de la intérprete, como también una teoría previa y una teoría al paso del hablante. En los cuatro casos, las teorías están constituidas por estados mentales de primer grado, segundo grado o n grados.

En el caso de las teorías previas de la intérprete, las creencias de primer grado son las que ella tiene acerca del hablante. Las creencias de segundo grado son las que ella tiene acerca de los estados mentales de él. Las creencias de tercer grado serían creencias de ella acerca de los estados mentales que él tiene acerca de ella, y así hasta los cuatro o cinco grados de intencionalidad. La teoría al paso de la intérprete es la teoría previa modificada para dar sentido a nueva evidencia o, expresado de otra forma, es la manera como ella usa su teoría previa para entender a un hablante en particular.

La teoría previa del hablante también está constituida por varios grados de intencionalidad e incorpora lo que él cree acerca de la intérprete; lo que él cree acerca de los estados mentales de ella; lo que él cree que son los estados mentales de ella acerca de él, y así en adelante. Es en virtud de estos estados mentales, en sus diversos niveles de intencionalidad, que el hablante construirá su mensaje dirigido a la intérprete. De esta manera, el hablante acomodará su discurso de acuerdo a lo que él cree son los estados mentales de la intérprete, en sus diversos grados. Pero al interactuar con ella, él descubrirá que su teoría previa no siempre es correcta ni la apropiada para lograr los objetivos comunicacionales que él tiene respecto de ella, de manera que se verá obligado a modificarla y esa modificación será su teoría al paso.

La teoría al paso del hablante será el producto de la reformulación de su teoría previa para una intérprete específica. En particular, la teoría al paso del hablante será la teoría al paso que él desea que ella desarrolle para poder entenderlo, es decir, está constituida por el tipo de estados mentales que él quisiera que ella llegue a tener para poder dar sentido a sus palabras y acciones.

Es importante notar que la palabra «teoría» puede confundir. No se trata de sistemas estructurados de manera totalmente consciente, sino de conjuntos de habilidades o estrategias de interpretación que están siempre modificándose. Sin embargo, el uso de la palabra «teoría» también alude a la reconstrucción conceptual que un filósofo hace de lo que intérprete y hablante necesitan saber y hacer para comunicarse, así como de lo que ocurre en la mente de la intérprete y del hablante para poder entender el fenómeno de la comprensión.

Lo importante es que para que haya comunicación no es necesario que hablante e intérprete compartan sus teorías previas, sino que sus teorías al paso puedan rozarse o tocarse en algún punto. Esto implica que para que haya comunicación y no malentendido, tanto intérprete como hablante deben tener la habilidad para construir teorías al paso creativas y flexibles para poder adaptarse a los diferentes interlocutores.

La producción de teorías al paso es la generación de pequeños y provisionales acuerdos acerca de cómo utilizar las palabras y, como he señalado, la comunicación se da cuando estas teorías al paso logran aproximarse. Pero, ¿cómo hay que entender que las teorías al paso se aproximen en algún punto? La idea es que los significados que A y B atribuyen a sus palabras y a las ajenas puedan llegar a coincidir o converger, aunque sea transitoriamente. No es necesario que ambos atribuyan los mismos significados a las mismas palabras, por pocas que ellas sean, sino que ambos sepan cuáles son los significados que cada uno de ellos está atribuyendo a las palabras, incluso si no son los mismos. Pero aquí surgen dos problemas importantes:

(1) ¿Quién decide si han coincidido o no tales significados? Es decir, ¿cómo se puede saber si «realmente» cada uno de ellos sabe qué significados atribuye a las palabras o si simplemente ellos creen que han coincidido y todo ha sido un malentendido sistemático en el que ambos creen que se están comunicando, pero «realmente» no ocurre eso? ¿Y qué significa «realmente» aquí? Incluso si hubiera una tercera persona que desde una posición privilegiada observase la escena, esa persona sería un tercer intérprete y tendría que emplear el principio de caridad y teorías al paso —aplicados a cada uno de los dos interlocutores—, con lo que simplemente habríamos añadido un personaje más al posible acuerdo o malentendido. Entonces, decir que A y B han coincidido significa simplemente decir que A cree que han coincidido y que B cree que han coincidido, y que tienen una interacción regular y sistemática cuya fluidez sugiere que hay comprensión entre ambos. Si «realmente» se están malentendiendo sistemáticamente como si fuera una comedia de equivocaciones solo podría decirlo una tercera intérprete, donde, en este caso, «realmente» significa «desde el punto de vista de C» y así sucesivamente. Pero como C también tendría que utilizar el principio de caridad y elaborar teorías previas y al paso, C no podría suponer que está malentendiendo sistemáticamente a A y a B. Si yo soy C no puedo decir que si las teorías al paso de A y B confluyen, desde mi punto de vista, se están malentendiendo sistemáticamente. En el mejor de los casos, C podría suponer que ella está entendiendo a A y a B mejor de lo que ellos mismos se entienden mutuamente.

(2) Por otra parte, si A y B coinciden en los significados que atribuyen a las palabras o, lo que es lo mismo, en la manera como usan las palabras, entonces A y B coinciden en sus creencias acerca de los objetos descritos por esas palabras y en sus disposiciones para comportarse en relación con ellos. Lo que tenemos aquí, entonces, es un tipo de objetividad constituida intersubjetivamente. Si además hay una comunidad de hablantes en la que hay coincidencias

> sistemáticas en el significado atribuido a las palabras y, por tanto, también en las creencias acerca de los objetos referidos por ellas —gracias al principio de inseparabilidad entre significado y creencia— y además observamos coincidencias en las disposiciones para actuar acerca de los objetos, lo que tenemos es amplios sectores donde coinciden teorías al paso que darán lugar a teorías previas convergentes. Esos sectores de coincidencia constituyen el significado literal de las palabras, pero también la objetividad, la cual es previa e independiente de la voluntad individual de los hablantes. Aunque, por supuesto, esa objetividad está sometida al cambio en la medida en que pueden modificarse, y de hecho lo hacen, las voluntades individuales de los hablantes. Así es como emerge el significado en situaciones comunicativas regulares y sistemáticas.

Como ya señalé, para que haya comunicación no se necesita teorías previas compartidas sino teorías al paso que coincidan, aunque sea transitoriamente. El que normalmente compartamos teorías previas no implica que estas sean condición necesaria para la comunicación, sino solamente que así la comunicación empieza de manera más fluida. En principio, la comunicación no requiere de ninguna comunidad previa ni identidad cultural o lingüística compartida, sino una que se construye en tanto la interacción es exitosa. Tradicionalmente se suponía que la comunicación es la fase final de un proceso que tiene los siguientes momentos:

① Convención→ ② Lenguaje→ ③ Comunicación

Pero Davidson sugiere un proceso inverso (1984h [1978], 1984b [1982]):

① Comunicación→ ② Lenguaje→ ③ Convención

No es que para que haya comunicación sea necesario compartir un lenguaje y para que haya lenguaje sea necesario compartir una convención, como sostenía Lewis en su famoso libro *Convention* (2002). Es más bien que, para que haya convención es necesario compartir un lenguaje y este ya presupone la existencia de la comunicación.

Aunque el fenómeno de la comunicación es extraordinariamente complejo y requeriría de un análisis mucho más detallado, puede ser descrito de manera sucinta como el proceso por el cual un individuo produce mediante sus acciones efectos deseados en otro y genera, eventualmente, significados compartidos acerca de tales acciones. Pero el que se compartan significados no implica que estos sean previos a la situación comunicativa. Imaginemos, por ejemplo, que Robinson Crusoe llega a la isla y descubre a un nativo a quien posteriormente llamará «Viernes». Como Robinson está sediento gesticula con las manos como si rociara un líquido en su boca, esperando que Viernes lo entienda y le traiga agua. Si el nativo le trae agua, podemos suponer que ambos habrán comprendido lo que el otro cree y desea, así como los significados

de los gestos realizados por Robinson. A eso solemos llamar «comunicación». Quizá alguien podrá suponer que Robinson deseaba saciar su sed, mientras que Viernes entendió que lo que él quería era participar de un ritual de camaradería mediante el que se bebe agua de un cuenco sagrado. Siempre quedará, por tanto, un grado de indeterminación que es parte consustancial a la comunicación. Pero es esencial notar que la indeterminación no procede de que nunca sabremos si «realmente» nos comunicamos o no, sino de que siempre habrá varias maneras diferentes de describir correctamente la situación comunicativa.

Por eso, no es relevante preguntar si «realmente» Viernes entendió las «verdaderas» intenciones de Robinson o si lo que ocurrió fue una mera coincidencia, porque no existe un ideal de comunicación —llegar a captar los «verdaderos» significados, deseos y creencias de los otros— que nunca llegaremos a capturar plenamente. No existen significados y creencias «en sí mismas» que debamos captar. En el contexto de la interpretación, los estados mentales y significados son hipótesis interpretativas que forman parte de una teoría que construimos y atribuimos a los demás, para hacer inteligible su comportamiento. Naturalmente, esto no niega que las personas tengan experiencias fenoménicas reales que son parte de lo que llamamos «estados mentales» y que existan los significados como prácticas sociales regulares de uso de las expresiones, en las comunidades de hablantes. En otras palabras, no estoy negando la realidad de los estados mentales y los significados.

Así, entonces, la comunicación es el proceso mutuo y coordinado de producir efectos deseados en el otro en relación con un mundo objetivo, mientras se construyen estados mentales y significados compartidos. La versión ontogenética más temprana de este proceso es el fenómeno de la atención conjunta, que emerge en el bebe, la madre y un mundo que ambos habitan en la primera infancia. Sobre esto hablamos en el apartado 2.3 y vimos que es un complejo proceso que tiene raíces no conscientes e instintivas, muchas de las cuales compartimos con otras especies animales.

Cuando los efectos deseados son logrados frecuentemente, tiende a producirse una práctica social regular. Así emergen los significados y el lenguaje. Supongamos, entonces, que cada vez que Robinson gesticula con las manos de una manera específica Viernes le trae agua. Supongamos también que este proceso se repite suficientes veces. Podremos decir correctamente que esos gestos significan «agua» para Robinson y Viernes, es decir, que ellos han comenzado a desarrollar un incipiente lenguaje común en el que tiene sentido decir que comparten significados, más allá de si Robinson está pensando solo en saciar su sed y Viernes en un ritual iniciático de camaradería. Por ello, sería igualmente correcto decir que los gestos de Robinson también significan «Por favor, tráeme agua», «Tengo sed y deseo agua» o «Verdaderamente me caería bien un cuenco con agua».

En ese sentido, el lenguaje presupone la comunicación, aunque no la comunicación el lenguaje. El lenguaje será, por tanto, ese conjunto de prácticas sociales compartidas —regularidades sociales— que gobiernan el uso de ciertas acciones con la finalidad de producir determinados efectos previstos en otros miembros de nuestra comunidad. Finalmente, si este lenguaje rudimentario se llega a hacer más frecuente y compartido, podremos decir que nos hallamos frente a una convención social.

Volvamos ahora al holismo de la interpretación para integrar el significado con la referencia. Dice Davidson:

> Sería un error suponer que de alguna manera podríamos primero determinar lo que un hablante cree, desea, espera, intenta, teme, y luego pasar a una respuesta definida a la pregunta de a qué refieren sus palabras. Pues la evidencia sobre la que todos estos temas dependen no nos permite separar las contribuciones del pensamiento, acción, deseo y significado una por una. Teorías totales es lo que debemos construir, y muchas teorías lo harán igualmente bien. (1984f [1979], pp. 240-241).

Si hay indeterminación [de la interpretación] es porque, una vez que se presenta toda la evidencia, permanecen abiertas maneras alternativas de determinar los hechos (1984 [1974a], p. 154).

Necesitamos ahondar un poco más en este tema. Hay un problema que se plantea en torno a la autonomía de la referencia que lo podemos plantear de manera análoga a la autonomía del significado. El principio de la autonomía del significado sostiene que el significado de una expresión —las prácticas sociales que gobiernan su uso regular y el significado que cada interlocutor atribuye a sus expresiones— se mantiene estable aunque cambien las intenciones al proferirla. Esta es la distinción entre significado y uso particular, y es necesaria para evitar que el concepto mismo de significado se desvanezca. Pero a pesar de esta relativa autonomía del significado —que es relativa porque si cambian mucho las intenciones de uso de los hablantes, cambiará también el significado de la expresión— el significado sigue siendo un concepto atribuido al agente al interior de un conjunto mayor y holista de conceptos interconectados.

Ahora el problema es cómo hay que entender la referencia. La posición tradicional, inspirada en Frege, asume una rígida autonomía de la referencia. Según esta posición no importa cuánto cambien nuestras creencias acerca del objeto —o el significado de las palabras que refieren al objeto—, el objeto mismo —es decir, el referente— se mantiene estable. Esto es necesario para poder construir el significado a partir de la referencia y es una posición deudora de la concepción referencialista del significado que revisamos en el apartado anterior. Una concepción ideacional como la de Locke diría que el significado es una propiedad monádica de las expresiones.

Y una concepción referencialista sostendría que el significado es una propiedad relacional diádica, que se constituye a partir de la relación de referir a la realidad. Como vimos, la posición que defiendo es que el significado es una propiedad relacional triádica entre intérprete, hablante y realidad.

Pero ahora una pregunta es hasta qué punto el cambio de creencias —o el cambio de significados— puede alterar la referencia misma. Abordamos este problema en el tercer capítulo, en el contexto de la filosofía de la ciencia. Así por ejemplo, ¿hasta qué punto el cambio de creencias que Galileo tiene respecto de Tolomeo acerca del significado de la palabra «Sol», hace que cuando Galileo y Tolomeo dicen «Sol» refieran a objetos distintos? Un defensor clásico de la autonomía de la referencia diría que las palabras de uno refieren a un objeto de manera independiente de las creencias que uno tenga respecto de él y de los significados que uno atribuya a las palabras que usa para nombrarlo. Galileo y Tolomeo referirían al mismo objeto, aunque lo interpreten de diferente manera y aunque la palabra «Sol» tenga significados diferentes en boca de cada uno de ellos (McGinn, 1977; Harré, 1986).

En el otro extremo estaría aquella posición que sostuviera, como en un tiempo lo hizo Kuhn (1971), que al cambiar las creencias acerca del objeto no solo cambia el significado sino también el referente mismo, es decir, que el objeto es distinto. De esta manera, Galileo y Tolomeo referirían a objetos diferentes. Así sería porque los objetos no están constituidos de manera previa e independiente de nuestras creencias acerca de ellos y porque la referencia está parcialmente determinada por la intención de uso.

Una posición holista como la que defiendo tendría que admitir la autonomía de la referencia y la autonomía del significado, pero con algunas cualificaciones. Por ejemplo, en la medida en que Galileo y Tolomeo compartan un número de creencias acerca del objeto al cual están apuntando con el dedo —que es amarillo, circular, caliente, que está a mucha distancia de la Tierra, etcétera— es posible decir que refieren a lo mismo, aun si atribuyen significados parcialmente diferentes a las palabras que emplean para «Sol». El holismo excluye la posibilidad de una total inconmensurabilidad y también la total inescrutabilidad de la referencia. Así como basta tener parcial conmensurabilidad para tener conmensurabilidad, basta con tener parcial referencia para tener referencia. Lo mismo se puede decir de la indeterminación de la interpretación.

En el caso que las creencias de los hablantes cambiasen radicalmente —uno lo ve amarillo y el otro azul, uno lo ve esférico y el otro triangular, uno lo percibe caliente y el otro frío— tendría sentido decir que no están refiriéndose al mismo objeto por la sencilla razón de que no están hablando de lo mismo. Así, ambos interlocutores se atribuirían mutuamente alucinaciones o patologías de la percepción —que no es otra cosa que atribuir errores generales, pero al interior de un sector del sistema de creencias—, con lo cual no se podría decir que hablan acerca del mismo objeto.

Entonces, si tiene sentido decir que dos personas hablan de lo mismo debemos aceptar que comparten más creencias acerca del objeto de las que no comparten. Lo importante, en todo caso, es que la referencia no funda el significado, ni tampoco el significado funda la referencia, sino ambos conceptos se construyen simultáneamente en la interacción interpretativa. Si es así, tampoco puede haber una total independencia entre significado y referencia, así como tampoco la hay entre creencia y referencia.

Pero uno debería preguntarse cómo comienza exactamente el proceso de la interpretación radical. Según Davidson, la intérprete radical comienza asumiendo que el hablante emite proferencias que él considera verdaderas, pero este punto ha sido disputado por varios autores. Según Malpas (1992), asumir que el hablante profiere oraciones que él cree que son verdaderas es ya atribuirle creencias. Así, el punto de partida de la interpretación no sería el asumir que el hablante profiere oraciones que él cree que son verdaderas, sino más bien asumir que el hablante tiene, *grosso modo*, las mismas creencias y deseos que la intérprete. Es decir, para Malpas —y en esto también Grandy (1973)— el punto de partida es el principio de caridad. Uno podría ir más lejos aún que ambos y afirmar que el punto de partida es la atribución que hace la intérprete de sus propios estados mentales al hablante. Es decir, ella asume en un primer momento que el hablante es un otro semejante. Solo ahí puede comenzar la interpretación. Por tanto, comprender a alguien será, por lo menos en un nivel explícito y consciente, construir para esa persona una teoría en la que podamos interconectar sus acciones, significados y estados mentales pertinentes. Pero esa teoría tendrá que emerger de nuestro propio sistema de actitudes proposicionales, es decir, de la manera en que nosotros nos entendemos a nosotros mismos y tendrá que estar interconectada con nuestros propios estados mentales y acciones. En otras palabras, construir un sistema de atribuciones para un hablante es tejer una red que conecte nuestros estados mentales con los de él. Sobre ese tema volveremos en detalle y en varias ocasiones, pero ahora debemos abordar de una manera algo más técnica el problema del significado al interior de una teoría de la interpretación.

Capítulo cinco
Significado, verdad e interpretación

5.1. El principio de verificación

Una idea que estoy defendiendo en este libro es que el significado de una expresión es una propiedad relacional triádica emergente, que surge a partir de una sucesión regular de situaciones comunicativas exitosas. También afirmo que el concepto de significado solo tiene sentido al interior de una teoría holista de la interpretación. Pero es necesario desarrollar algunos detalles técnicos de esa propuesta. En este capítulo sostendré que dos oraciones tienen el mismo significado si tienen las mismas condiciones de verdad, es decir, si ante las mismas circunstancias del mundo los hablantes estarían dispuestos a aseverarla como verdadera.

Esa intuición es conocida como una concepción veritativo-condicional del significado y está emparentada con el principio de verificación que se puede rastrear hasta el primer Wittgenstein (1975)[1], algunos positivistas lógicos y Alfred Ayer (1952). Pienso que, aunque el principio de verificación en la versión clásica de Ayer es demasiado estrecho para ser parte de una adecuada teoría semántica, hay todavía un sentido en el que el significado de una oración está estrechamente relacionado con sus condiciones de verificación. En este capítulo quisiera explorar en qué sentido el principio de verificación es demasiado estrecho y en qué sentido es todavía adecuado.

Mi tesis central será que una doctrina verificacionista del significado puede funcionar únicamente al interior de una semántica y una teoría de la interpretación holistas; por el contrario, una doctrina verificacionista asociada a una semántica atomista no puede sino fracasar. Así pues, verificacionismo y holismo pueden ser una buena alianza para explicar el significado, mejor que otras teorías disponibles.

[1] «Conocer una proposición significa conocer qué es el caso si esta proposición es verdadera» (Wittgenstein, 1975, parágrafo 4.04).

Comenzaré discutiendo las versiones clásicas del principio de verificación de Ayer. Trataré de mostrar por qué este principio es una doctrina paradójica e incompleta para después sugerir, sobre la base de la interpretación davidsoniana de Tarski, cómo puede ser reformulado, con la finalidad de integrarlo a una doctrina holista de la interpretación. Al final del capítulo intentaré señalar cómo se puede construir una semántica que pueda integrar las diversas —y en ocasiones consideradas excluyentes— intuiciones que tenemos acerca del significado. Esto me permitirá mostrar algunas conexiones entre el holismo semántico y algunas tesis de Wittgenstein y Heidegger.

Una adecuada teoría del significado lingüístico debería estar en condiciones de proporcionar respuestas a, por lo menos, estas tres preguntas:

(1) Qué hace que una oración, o una palabra dentro de una oración, tenga significado en un lenguaje dado.

(2) Cómo es posible determinar el significado de una oración, o una palabra dentro de una oración, en un lenguaje dado.

(3) Qué hace posible que un hablante «conozca» un lenguaje sobre la base de cierta información disponible (Dummett, 1975; Evans & McDowell, 1976; Davidson, 1984c).

El principio de verificación, en la versión clásica de Ayer, pretende responder a (1) y solo parcialmente a (2), de ahí que no sea propiamente una teoría del significado sino una herramienta diseñada para demarcar entre oraciones significativas y oraciones que no lo son. Ahora voy a discutir su criterio de demarcación.

Solo para subrayar lo obvio, creo que sería justo decir que un importante punto logrado por los positivistas lógicos fue mostrar la profunda conexión entre el significado de una oración y sus condiciones de verdad. Ellos nos mostraron que el significado de una oración solo puede ser explicado en relación a la aseveración de su verdad por los hablantes. De esta suerte, para un hablante de un lenguaje L, conocer el significado de una oración en L es conocer sus condiciones de verdad, esto es, las circunstancias del mundo en que la oración sería verdadera o falsa.

Sin embargo, muchos positivistas lógicos, como Ayer, escribieron como si el significado y la verdad de una oración en L fuesen independientes de los significados y los valores de verdad del resto de oraciones de L, y como si el método de verificación de una oración fuese en principio independiente de los significados y valores de verdad de las otras oraciones del lenguaje. En otras palabras, Ayer parece estar asumiendo que las observaciones empíricas que hacen que una oración sea verdadera son lógicamente independientes de los significados de las oraciones que previamente han sido asumidas como verdaderas. Este es un presupuesto que procede del atomismo

lógico de Russell y del primer Wittgenstein y es, a mi juicio, el principal problema del principio de verificación sostenido por Ayer. Él dice que podemos comprender una oración si sabemos qué sería que esta oración fuese verdadera, es decir, qué clase de observaciones empíricas justificarían esa oración. Esa intuición me parece fundamentalmente correcta. Empero Ayer también dice que el tipo de conocimiento que requerimos para saber si una oración en L es verificable, o no, es, en principio, independiente de nuestro conocimiento de las otras oraciones de L, y ese presupuesto me parece básicamente errado. Ayer es consciente de que la experiencia siempre verifica o desacredita sistemas de oraciones y no oraciones de manera aislada (1952, p. 94), pero también cree que no necesitamos conocer los significados ni los valores de verdad de sistemas de oraciones para saber si alguna observación en particular verifica o desacredita una oración en particular.

En la nueva introducción a *Language, Truth and Logic*, escrita en 1946 y reimpresa en todas las siguientes ediciones, Ayer (1952, p. 10) se desplaza hacia una posición todavía más extrema, en la que sostiene que las oraciones observacionales pueden ser conclusivamente verificadas, de suerte que se dirige hacia un atomismo aún más radical acerca de la experiencia.

A diferencia de Ayer, pienso que podemos entender qué sería que una oración fuese verdadera precisamente porque podemos comprender qué sería que muchas otras oraciones fuesen verdaderas, es decir, porque podemos entender un lenguaje. Esto, por otra parte, no es diferente de estar en condiciones de interpretar ciertos datos observacionales como si se justificara la verdad de ciertas oraciones. Así pues, saber o creer que ciertas oraciones son verdaderas es indesligable de tener una interpretación de nuestra experiencia de la realidad. Para poder conocer el significado de una oración es preciso conocer el significado de muchas otras y para saber que una oración es verdadera es necesario saber que muchas otras también lo son.

Pero esto no significa necesariamente que debamos abandonar el principio de verificación, tal vez significa que deberíamos ampliar su contexto. Frege solía decir que una palabra tiene significado solo en el contexto de una oración; por tanto, deberíamos decir, con Quine y Davidson, que una oración tiene significado solo en el contexto de un lenguaje. La expansión del principio del contexto, de la oración a la totalidad del lenguaje, es la manera de salvar el principio de verificación sin sus inconvenientes.

Uno de los problemas centrales del principio de verificación es que no resulta fácil formularlo. Así pues, intentaré describir su versión más temprana y después discutiré versiones más sofisticadas. Ayer (1952, p. 35) dice que uno puede entender una oración si y solo si tiene una idea de lo que tendría que hacer para verificar la proposición que la oración expresa, y si uno tiene una idea de qué clase de observaciones la justificarían. Ya que él está comprometido con el empirismo, su idea

de «verificación» está enraizada en la existencia de condiciones observacionales que justificarían la proferencia de la oración. Pero en este punto es conveniente recordar la distinción que hace este autor entre verificabilidad práctica y verificabilidad en principio. Una oración puede ser verificable en la práctica si, en efecto, tenemos los medios para verificarla. Este es el caso, por ejemplo, de la oración «No hay oxígeno en la Luna». Sabemos qué hacer para verificar la oración y contamos con los medios para hacerlo. De otro lado, una oración puede ser verificable en principio, aunque no en la práctica, si sabemos qué clase de observación la verificaría, aún si no tenemos los medios para hacerlo. Un ejemplo de esto sería la oración «Hay oxígeno en Alfa Centauro». Ayer abandona la verificación práctica y se queda con la verificación en principio. También distingue entre verificación fuerte y débil. Una oración es verificable en el sentido fuerte si y solo si su verdad puede ser conclusivamente establecida en la experiencia, mientras que es verificable en el sentido débil si la experiencia la hace probable. En sus trabajos tempranos Ayer abandonó el sentido fuerte y se quedó con el débil, ya que el sentido fuerte implicaría un estándar imposible de cumplir. Por ejemplo, proposiciones cuantificadas universalmente del tipo «Todos los x son y» serían inverificables en el sentido fuerte y, por tanto, asignificativas. Esto tendría como una consecuencia absurda que todas las leyes de las ciencias empíricas resultaran asignificativas, porque todas ellas tienen la forma de proposiciones universales. Ya que todas las leyes empíricas son hipótesis acerca de la experiencia futura la verificación fuerte es inútil, siendo solo viable la verificación débil. Sin embargo, como hemos visto, en la nueva introducción Ayer sostiene que hay algunas oraciones que pueden ser verificables conclusivamente: las oraciones observacionales. Estas oraciones describen experiencias individuales y Ayer afirma que no pueden errar porque no relacionan esta experiencia con nada más, y, por tanto, no hay espacio para el error. Quisiera subrayar nuevamente que esta movida fue desafortunada, porque radicalizó su atomismo semántico y epistemológico.

Veamos ahora con más detalle la nueva formulación que hace del principio de verificación en la nueva introducción a *Language, Truth and Logic* de 1946. Para poder preguntar si una proposición es o no verificable debemos previamente saber que es una proposición, es decir, debemos saber que tiene significado, de otra manera ni siquiera sería una proposición. Para afrontar esta paradoja Ayer propone algunas distinciones terminológicas. Sugiere definir una oración (*sentence*) como cualquier cadena gramatical, sea significativa o no (1952, p. 8). Así, «Ideas verdes incoloras duermen furiosamente» sería tan buena oración como «Aristóteles escribió la *Metafísica*». Ayer continúa diciendo que cualquier oración en modo indicativo, ya sea significativa o no, expresa una afirmación (*statement*) (1952, p. 8). Por otra parte, también sostiene que dos oraciones mutuamente traducibles expresan la misma afirmación.

Aquí surge, entonces, un problema importante. Como mostró Quine en «Two Dogmas of Empiricism» (1963), el concepto mismo de traducibilidad presupone el concepto de identidad de significado. Decimos que dos oraciones son mutuamente traducibles si tienen el mismo significado. Así, la traducibilidad no puede ser criterio de identidad de significado y la tesis de Ayer es circular. Finalmente, sostiene que las oraciones que son significativas expresan una proposición. Dice que las proposiciones son una subclase de las afirmaciones (1952, p. 8), y que la diferencia es que las proposiciones son afirmaciones significativas, mientras que no toda afirmación es significativa y, por tanto, no toda afirmación expresa una proposición. Pero ya que antes había dicho que dos oraciones traducibles entre sí expresan la misma proposición y dado que el concepto de traducibilidad presupone el de significado, dos oraciones son traducibles si tienen significado, y si tienen significado ya expresan una proposición. Por otra parte, si una afirmación carece de significado entonces no es el caso que alguna oración pueda expresarla. Así, el concepto de afirmación (*statement*) definido por Ayer carece de utilidad, porque o bien significa lo mismo que «oración» si no necesariamente posee significado o que «proposición» sí debe poseer significado.

El punto de Ayer, en su nueva formulación, es que el principio de verificación se aplica a afirmaciones más que a oraciones, lo que le permite decir que una afirmación es literalmente significativa si y solo si es o bien analítica o empíricamente verificable (1952, p. 9). Y una afirmación es empíricamente verificable, y entonces significativa, si «algunas oraciones observacionales pueden ser deducidas de ella en conjunción con ciertas otras premisas, sin que puedan ser deducibles solo de esas premisas» (1952, p. 11).

El problema con esta formulación es que hace que cualquier oración sea significativa y esto es algo de lo que Ayer posteriormente se dio cuenta. Tomemos, por ejemplo, las siguientes oraciones:

(1) Ideas verdes incoloras duermen furiosamente.

(2) Si ideas verdes incoloras duermen furiosamente entonces el césped es verde.

Es claro que de (1) y (2) juntos podemos deducir la oración observacional «El césped es verde», que no se sigue separadamente de (1) ni de (2). Esto mostraría que «Ideas verdes incoloras duermen furiosamente» es una oración significativa y, por supuesto, esto es una reducción al absurdo de todo el principio.

Ayer llegó a reconocer que el principio de verificación no funciona, bien porque es demasiado restringido o porque es demasiado amplio, y, ya que él pensó que el punto de vista expresado por este principio es substancialmente correcto (1952, p. 5), se trataría simplemente de encontrar el punto adecuado entre estos dos extremos.

Discrepo con esa tesis. Pienso que lo que debe ser transformado no es el principio mismo, en su versión temprana, sino su contexto.

Hay muchas otras objeciones que se han hecho y que podrían hacerse contra el principio de verificación. Se ha dicho, por ejemplo, que el principio se refuta a sí mismo, porque cuando uno lo expresa en una oración, la oración misma no es verificable, y, por tanto, es asignificativa. Ayer ha contestado a esta objeción que el principio no es una oración sintética sino un criterio estipulativo (1952, p. 16).

No voy a detenerme en las objeciones que señalan autorreferencialidad en el principio, pues estas han sido ampliamente discutidas. Quisiera formular, más bien, una objeción diferente. El principio de verificación de Ayer asume que podemos verificar una proposición de manera aislada, independientemente de las otras proposiciones en el lenguaje o la teoría a la cual esa oración pertenece. Pero ese presupuesto está errado. Ya que una proposición está interconectada con muchas otras proposiciones en un lenguaje o un sistema de creencias, para poder verificar una proposición cualquiera debemos considerar muchas otras proposiciones relevantes del lenguaje o sistema de creencias. Ayer asume la tesis atomista de que la verdad de una oración es independiente de la verdad de las otras oraciones del lenguaje. De esta manera, quisiera discutir hasta qué punto el principio de verificación está comprometido con el atomismo y si es posible arreglar un matrimonio entre verificación y holismo.

Cuando Ayer analiza las oraciones observacionales simplifica el problema, porque no es posible comprender ni verificar una oración empírica independientemente de una multitud de otras oraciones. Consideremos, por ejemplo, la oración «Veo un vaso de vino tino». El comprender y verificar esta oración requiere comprender y conocer la verdad de muchas otras oraciones relacionadas, como, por ejemplo:

(1) Ver un objeto x es tenerlo frente a los ojos y reconocer su forma y su color.

(2) No puedo ver nada a menos que tenga los ojos abiertos.

(3) El vino es una bebida alcohólica.

Estos son solo algunos ejemplos de oraciones que, en condiciones normales, un hablante tendría que entender y considerar verdaderas para poder comprender la oración en cuestión. Tendríamos que preguntarnos si tales oraciones son verificables antes de preguntar si la oración «Veo un vaso de vino tinto» es verificable. Pero para que un hablante pueda comprender y conocer el método de verificación de esas otras oraciones, ciertamente tendría que entender y conocer los métodos de verificación de muchas otras oraciones y así sucesivamente. Esto nos conduciría a un regreso al infinito, lo cual mostraría que no hay oraciones, ni siquiera las oraciones observacionales, que sean conclusivamente verificables. De otro lado, también mostraría que

uno solo puede conocer los métodos de verificación de sistemas de oraciones y no de oraciones aisladas.

Me parece que desde los trabajos de Quine y Davidson este punto debe quedar claro: conocer el significado de una palabra requiere tener una multitud de creencias acerca del referente o del concepto aludido por la palabra. Asimismo, conocer el significado de una oración requiere comprender una multitud de oraciones que incluyen las palabras que aparecen en la oración dada. Ahora quisiera relacionar este tema con el rechazo de Ayer a las oraciones metafísicas.

Sostiene Ayer que la mayor parte de las afirmaciones de la metafísica tradicional no son verificables en principio, porque no sabemos qué clase de observaciones constituirían una buena justificación de tales oraciones. La oración que él elije como ejemplo para demostrar su argumento es una tomada al azar de un libro de Bradley: «El Absoluto entra en, pero no es capaz de, la evolución y el progreso»[2] (1952, p. 36). La objeción obvia sería que Ayer es incapaz de verificar tal oración precisamente porque ha sido tomada al azar de un libro, sin considerar el sistema de creencias que harían significativa la oración. Tomemos un ejemplo más simple: «La materia es energía condensada y la energía es materia enrarecida». Esta oración ha sido tomada al azar de la teoría de la relatividad de Einstein. ¿Sabemos cómo verificar esa oración? ¿Sabemos qué clase de observaciones la justificarían? Ciertamente no, a menos que conozcamos lo suficiente de la teoría de la relatividad de Einstein como para dar sentido a la oración. Esto es así, en primer lugar, porque si no estamos familiarizados con la teoría de la relatividad ni siquiera sabríamos qué significan las palabras «energía» y «materia» en tal teoría. Pero, además, conocer el sistema de creencias al que la oración pertenece sería ya el primer paso para conocer el significado de la oración.

Tomemos como otro ejemplo la célebre idea hegeliana de que «El curso de la historia es el autoconocimiento y autoproducción del Espíritu». Uno sabría cómo verificar esa oración si uno supiera los significados de las palabras «historia», «Espíritu», «autoconocimiento» y «autoproducción» en el lenguaje de Hegel. Si uno conociera el sistema hegeliano estaría en condiciones de precisar el sistema de creencias al cual la oración pertenece y así uno podría «verificar» tal oración al interior del sistema hegeliano, es decir, uno podría saber lo que esa oración significa para Hegel y por qué es verdadera para él. Uno llegaría incluso a saber qué clase de observaciones empíricas podrían justificar tal oración.

[2] «*The Absolute enters into, but is itself incapable of, evolution and progress*».

Esta es entonces mi tesis: una oración es verificable solo al interior de un sistema de creencias. Nótese que no he dicho que una oración es verdadera al interior de un sistema de creencias. Conocer el método de verificación de una oración dada es conocer el sistema de creencias al cual esa oración pertenece y cómo esa oración se interconecta con el resto de oraciones del sistema[3]. Por supuesto podría ocurrir que el sistema de creencias en su totalidad fuese falso, pero las oraciones que a él pertenecen han de ser significativas para que puedan serlo. Por otra parte, un sistema de creencias en su totalidad solo puede ser falso si es contrastado con otro sistema mayor que es asumido como verdadero y que brinda los criterios de verdad del subsistema considerado falso.

Ayer no estaría de acuerdo con esto. Él diría que independientemente de cuánto conozca uno sobre Hegel no hay ninguna observación posible que justifique la oración «El curso de la historia es el autoconocimiento y autoproducción del Espíritu». Veamos. Si sé que la tesis de Hegel es que el curso de la historia está determinado por la toma de conciencia de los pueblos y también que la noción de «espíritu» significa lo que un pueblo sabe de sí, es decir, la autodescripción de un pueblo en cierto momento de su desarrollo, sabré que hay muchas observaciones empíricas que pueden verificar o falsar tal oración.

Así pues, pienso que es posible reformular el principio de verificación solo si proveemos un principio de contextualización suficientemente amplio como para hacer el significado de una oración relativo a un lenguaje o sistema de creencias. Me parece, también, que en esta dirección va —o por lo menos debería ir— el programa holista de Davidson y su reinterpretación de Tarski. Considero que después de esta reformulación, aunque todavía podamos proveer el significado de una oración a través de sus condiciones de verdad, no sería posible distinguir, fuera de un contexto dado, entre oraciones significativas y oraciones asignificativas, que era precisamente el objetivo del principio verificacionista de Ayer. Voy a basarme en el programa de Davidson para mostrar cómo podrían integrarse verificacionismo y holismo. Sin embargo, me alejaré con frecuencia de sus tesis explícitas y extraeré consecuencias o sugeriré tesis independientes con el objetivo de iluminar la noción de significado.

[3] Considérese el siguiente texto de Wittgenstein: «Toda verificación, toda confirmación e información de una hipótesis tiene lugar dentro de un sistema. Y este sistema no es un punto de partida más o menos arbitrario y dudoso para todos nuestros argumentos: no, pertenece a la esencia de lo que llamamos un argumento. El sistema no es tanto el punto de partida, como el elemento donde los argumentos tienen vida» (1972, aforismo 105).

5.2. Interpretación y condiciones de verdad

Es materia de debate si las intuiciones semánticas que desarrolla Davidson en sus primeros escritos se abandonan en su época tardía o se reformulan de diversas maneras. El propio autor pensó lo segundo, aunque mucha gente piensa lo primero. Mi impresión es que las particularidades técnicas de la propuesta davidsoniana han conducido a que muchos filósofos se concentren en discutir los detalles y pierdan de vista las intuiciones que habitan el conjunto. En este libro me interesa iluminar esas intuiciones y me dirijo hacia los detalles técnicos solo si veo que su análisis puede iluminar el objetivo último. Una de esas intuiciones centrales es que el fenómeno de la interpretación no debe verse únicamente desde una perspectiva de tercera ni de primera persona, sino de manera integrada y tomando también en consideración la segunda persona. Este modelo triangular permite ver la interpretación de una manera involucrada e intersubjetiva, y no de manera desvinculada. Esto que vale para la interpretación en general, se aplica también para sus componentes, como el significado y la referencia.

La propuesta de Davidson es invertir la definición que Tarski (1944, 1956) da de la verdad, con la finalidad de construir teorías empíricas de interpretación para lenguas naturales. Su idea es que es posible aplicar los procedimientos de Tarski a un lenguaje cuya interpretación no es asumida, pero para el cual la verdad es asumida como un primitivo. Esta definición proporciona condiciones necesarias y suficientes para que una oración sea verdadera, y proporcionar tales condiciones es una manera de especificar el significado de una oración mediante otra oración que tenga las mismas condiciones de verdad.

La idea de Tarski es que no deberíamos ver el valor de verdad de una oración como una función de la referencia de sus términos componentes. Por ello sustituye la problemática noción de referencia por una más clara noción de satisfacción, como la relación entre las oraciones abiertas de un lenguaje y las varias secuencias de objetos que pueden hacerlas verdaderas. Así la verdad de las oraciones cerradas de un lenguaje es entendida como una función de la satisfacción de aquellas oraciones abiertas. La importancia de esta definición de verdad es que no involucra ninguna noción semántica para la definición, pues ni siquiera el concepto de satisfacción requiere de una noción semántica previa. La definición es materialmente adecuada y formalmente correcta. Lo primero significa que la definición expresa la noción intuitiva que tenemos de la verdad; lo segundo, que esto se produce de manera no ambigua gracias al empleo de términos que han sido explícitamente especificados.

Mientras Tarski asume el significado para dar una teoría de la verdad, Davidson asume la verdad para dar una teoría del significado. Según esto, si uno conociera

las condiciones de verdad de p y también supiera que q tiene las mismas condiciones de verdad, uno estaría en condiciones de saber que q es una buena traducción de p, es decir, que p y q significan lo mismo para una particular intérprete en un momento dado. De esta manera, mientras Ayer quería usar el concepto de condiciones de verdad para saber si una oración tiene significado o no, Davidson usa las condiciones de verdad para proveer una traducción apropiada para una oración dada. Esta traducción no nos dirá si la oración tiene o no significado —como si esta fuese una propiedad monádica—, solo nos dirá para quién tiene significado, es decir, en qué clase de manual de traducción o sistema de creencias resulta inteligible. Si tenemos una oración s, p será una buena traducción de s, si y solo si s y p tienen las mismas condiciones de verdad para una intérprete en particular. Las equivalencias así construidas toman la forma de teoremas en una teoría del significado para una lengua dada. Estos teoremas son llamados «oraciones-T» y su estructura formal es la siguiente:

(T) La oración «s» es verdadera en L si y solo si p.

Donde «s» es la oración del lenguaje objeto, es decir el lenguaje a ser interpretado; las expresiones «la oración» y «es verdadera si y solo si» pertenecen al metalenguaje, es decir, al lenguaje usado por la intérprete; y, finalmente, «p» es el lenguaje sujeto, esto es, el producto de la interpretación. Por ejemplo:

(T) La oración «*Grass is green*» es verdadera si y solo si el césped es verde.

Por supuesto el lenguaje sujeto, el lenguaje objeto y el metalenguaje podrían estar expresados en la misma lengua, con lo cual tendríamos:

(T) La oración «El césped es verde» es verdadera si y solo si el césped es verde.

Como parte del objetivo de una teoría del significado es proporcionar el significado de cualquier oración en un lenguaje dado, la aplicación del principio de Tarski en la versión de Davidson sería exitosa si, para cualquier oración s de L, la teoría pudiera proporcionar una oración p que sea una adecuada traducción de s. Una descripción semántica de un lenguaje sería completa solo si pudiéramos encontrar una oración-T para cualquier oración construible en el lenguaje. Así, nuestra comprensión de oraciones individuales depende de nuestro conocimiento de sus condiciones de verdad, y nuestra comprensión de las palabras individuales que constituyen las oraciones depende de nuestra comprensión de oraciones completas, específicamente de oraciones consideradas verdaderas por nosotros. La tesis es, entonces, que la comprensión del significado de una palabra depende de nuestra

comprensión de las condiciones de verdad de la oración que la contiene. Además comprender el significado de una oración dada requiere comprender el significado de muchas otras, es decir, requiere poseer un sistema de creencias y esto, naturalmente, implica pertenecer a una comunidad de hablantes y habitar una forma de vida. El acceso a los valores de verdad de las oraciones es, en principio, anterior al acceso a los significados y a los referentes de las palabras componentes pero, naturalmente, uno solo puede comprender la verdad de una oración si previamente comprende los significados de sus términos componentes. Tenemos pues un modelo holista en el que el todo se comprende por las partes y las partes por el todo; estamos frente a una versión técnica del principio del círculo hermenéutico desarrollado, entre otros, por Heidegger y Gadamer.

Hay aquí un triángulo formado por un hablante, una intérprete y el mundo que comparten. Los hechos del mundo proporcionarán fundamentos para que la intérprete encuentre traducciones adecuadas de las oraciones del hablante. Resultará, entonces, claro que el significado de una oración no puede estar aislado de los significados de las otras oraciones generables en el lenguaje de la intérprete o, lo que es lo mismo, de las otras creencias en el sistema de creencias de la intérprete. Las condiciones de verdad se asignan por medio de observaciones empíricas, tal como en el caso de Ayer, pero no de oraciones aisladas sino de sistemas integrados de oraciones. Esto permitiría la construcción de una teoría del significado que sea al mismo tiempo verificacionista y holista.

El objetivo de Davidson es mostrar que entidades no lingüísticas como objetos físicos, hechos u observaciones empíricas no pueden fundar el significado —como suelen creer los partidarios de las teorías referencialistas del significado y las formas clásicas del principio de verificación— si no es con la participación de un sistema de creencias que sirva de contexto. De esta manera, nos alejamos de una interpretación atomista y referencialista del principio de verificación para acercarnos a una interpretación holista del mismo. En una oración-T, p es una adecuada hipótesis interpretativa para s, si y solo si s y p tienen las mismas condiciones de verdad, según los criterios de la intérprete que, en gran medida, serán los criterios de su comunidad.

Según este modelo, el concepto de verdad es mucho más básico y fácil de entender que el concepto de significado, de ahí que se pueda asumir que entendemos el concepto de verdad para poder iluminar el de significado. ¿Cuál es este concepto de verdad? Simplemente el concepto mismo de creencia, en el que «p es verdadero» equivale simplemente a p. Como se sabe, esta es una teoría «descomillada» de la verdad —*disquotational theory of truth*— que es una teoría de la redundancia y también una concepción deflacionista de la verdad (Davidson, 2005b, capítulos 1-6).

Así, una adecuada teoría de la interpretación para un lenguaje L será aquella que esté en condiciones de asignar una oración-T para cualquier oración construible en L. La teoría de la interpretación no dirá si una oración de L es verdadera o falsa, significativa o asignificativa, solo dirá cuáles son las condiciones de verdad de la oración según los criterios del lenguaje-sujeto y del metalenguaje, es decir, según los criterios de la intérprete misma y de la comunidad a la cual ella pertenezca. Ahora bien, señalar las condiciones de verdad de una oración es lo mismo que mostrar las interconexiones que guardan las diversas oraciones generables en ese lenguaje y sus relaciones con los hechos asumidos del mundo o, lo que es lo mismo, señalar las condiciones de verdad de una creencia es señalar las interconexiones que tiene con otras creencias del sistema de creencias. Por ejemplo, supongamos que yo creo con Aristóteles que «Todos los hombres tienden al conocimiento por naturaleza». Las condiciones de verdad de esa oración están dadas por lo que en el lenguaje-objeto significan las expresiones «todos», «los hombres», «tienden», «al conocimiento», «por naturaleza», según los criterios de interpretación que se tiene desde el metalenguaje y el lenguaje-sujeto. Otra forma de decir lo mismo es: las condiciones de verdad de esa creencia están dadas por las creencias que en un sistema se tiene acerca de lo que significan las palabras «todos», «los hombres», «tienden», «al conocimiento», «por naturaleza», desde otro sistema de creencias que interpreta al sistema de creencias-objeto. Yo no podría creer que «Todos los hombres tienden al conocimiento por naturaleza» si no tuviera muchas otras creencias acerca de «todos», «los hombres», etcétera. De igual manera, las condiciones de verdad de «Todos los hombres tienden al conocimiento por naturaleza» están dadas por las condiciones de verdad de las diversas oraciones que en un lenguaje L se pueden construir con las palabras «todos», «los hombres», «conocimiento», etcétera.

De esta manera, construir una teoría de la interpretación para L es construir un conjunto de hipótesis que toman la forma de oraciones-T. A su vez, esto es mostrar las interconexiones que hay entre las diversas oraciones generables en L, es decir, las diversas creencias en un sistema de creencias dado. Si nuestro proyecto incluye también interconectar creencias, significados, deseos y acciones, estamos en camino a construir un modelo verdaderamente holista del significado y la interpretación.

Pero, además, como las oraciones-T conectan oraciones de dos lenguajes o sistemas de creencias diferentes a partir de la asunción que tienen las mismas condiciones de verdad —según los criterios de la intérprete— las oraciones-T muestran las interconexiones que también hay entre dos lenguajes o sistemas diferentes. Son estas interconexiones, asumidas por la intérprete, las que permiten la interpretación y la conmensurabilidad entre sistemas de creencias que incluso pueden ser muy diferentes entre sí.

5.3. Significado y comprensión

Ahora podemos preguntarnos de una manera más puntual qué es el significado de una expresión y qué hace que una expresión tenga significado. El planteamiento en términos de oraciones-T podría hacernos suponer que se trata de un retorno a la doctrina verificacionista clásica. Sin embargo, esto no es exacto. El modelo holista que estoy defendiendo, basado en la teoría davidsoniana de la interpretación radical, es más bien una versión depurada y técnica de la doctrina acerca del significado del segundo Wittgenstein en términos de usos sociales regulares y compartidos[4]. Intentaré ahora mostrar cómo me parece esto posible.

Definimos el significado en términos de la traducción de oraciones asumidas como verdaderas por distintos hablantes. El asumir la verdad de una oración es manifestar un conjunto integrado de creencias. Si entendemos una creencia como la incorporación de una disposición para actuar, entonces mantener un conjunto de creencias será tener un sistema integrado de disposiciones para actuar. Aunque el significado se dice de muchas maneras, podríamos describir lo que significa que una expresión tenga significado de varias maneras distintas, pero plenamente compatibles.

(1) Bajo cierta descripción, el significado de una expresión estaría dado por el conjunto de creencias que, acerca del objeto referido por esa expresión, tiene una comunidad de hablantes. De esta manera, el significado de x en L sería el conjunto integrado de creencias que el promedio de los hablantes de L tiene acerca del objeto referido por x. Si x no tiene referente, entonces por el concepto, evento, representación mental, etcétera, mentado o aludido por x. Como una creencia es una disposición para actuar, entonces el significado de x sería el conjunto de disposiciones para actuar, en relación con el objeto referido por x o el concepto aludido por x, que los miembros de una comunidad de hablantes comparten en cierto momento de la historia de la comunidad.

(2) Como una disposición para actuar es un hábito de comportamiento, podemos también definir el significado de x como aquellas prácticas sociales de los miembros de una comunidad que gobiernan el uso de x. Esta no es sino la concepción

[4] Este modelo es también compatible con el que sugiere Habermas: «En el campo de la teoría del significado defiendo la idea de que comprendemos un acto de habla significado literalmente cuando conocemos las condiciones bajo las que podría aceptarse como válido por el oyente. Esta versión pragmáticamente extendida de la semántica condicional de la verdad está apoyada por el hecho de que conectamos la ejecución de los actos de habla con varios requisitos de validez: los requisitos de la verdad de las proposiciones (o de las presuposiciones existenciales de los contenidos proposicionales), los requisitos de exactitud de una elocución (con respecto a los contextos normativos existentes) y los requisitos de la veracidad de una intención expresada» (1995, p. 323).

wittgensteiniana del significado entendido como uso regular[5]. Así pues, el significado de x sería también el sistema integrado de acciones que, en una comunidad o forma de vida, gobierna el uso de x y, en general, la actitud y la relación frente al objeto o concepto referido por x. El «ver como» de Wittgenstein, a diferencia del «ver que», permite situar a un objeto en un contexto. Es hacerlo familiar mostrando las relaciones que tiene con otros objetos conocidos, es decir, interpretarlo. Aquí hay una conexión evidente entre el holismo de Wittgenstein y el de Heidegger. Entender el lenguaje como constituido por, y como inseparable de, las prácticas sociales que realizamos y la forma de vida que habitamos, nos permite entenderlo como el lugar donde está depositada, de manera altamente condensada y densificada, nuestra precomprensión del mundo (Heidegger, 1986, pp. 31-34; Gadamer, 1977b, p. 9). Estas prácticas sociales que constituyen la precomprensión, en tanto disposiciones para actuar no conscientes, actúan a manera de estructura previa frente a la cual se constituirán las nuevas creencias y las nuevas formas de interrelación con el mundo. En los casos más básicos es incluso nuestro criterio de verdad y falsedad. Considérese lo que dice Wittgenstein:

> Pero yo no me procuré mi figura del mundo porque me cerciorara de su corrección; ni lo asumo porque esté convencido respecto de su corrección. No: es el telón de fondo heredado contra el cual distingo entre lo verdadero y lo falso (1972, aforismo 94).

(3) ¿Es la precomprensión también un sistema de creencias? Sí, en tanto es un sistema de disposiciones para actuar no consciente, es decir, del que no siempre nos percatamos ni podemos dar razón. Un sistema de creencias es un continuo que va desde una mayor conciencia y capacidad para dar razón hasta una menor conciencia y menor capacidad para dar razón[6]. Por otra parte, la intuición heideggeriana de que la palabra «otorga ser» y «da sentido» a los objetos puede entenderse como que las prácticas sociales que constituyen un lenguaje y una forma de vida configuran nuestra relación con los objetos, y los convierten en obstáculos o posibilidades, instrumentos o barreras, es decir,

[5] «Para una *gran* clase de casos de utilización de la palabra «significado» —aunque no para *todos* los casos de su utilización— puede explicarse esta palabra así: el significado de una palabra es su uso en el lenguaje» (1988, p. 43).

[6] La idea de que una creencia es una disposición para actuar se encuentra no solo en la tradición pragmatista sino también insistentemente en Wittgenstein, 1972, aforismos 144, 204, 229, 254, 284, 360, 395, 414, 427, 431.

les dan sentido para nosotros y, por tanto, nos dan sentido a nosotros también en el mundo. Dice Heidegger:

> Hablar es articular «significativamente» la comprensibilidad del «ser en el mundo», al que es inherente el «ser con» y que se mantiene en cada caso en un modo determinado del «ser un con otro» «curándose de» (1986, parágrafo 34, p. 180).
>
> Es solo la palabra la que otorga la venida en presencia, es decir, el ser, aquello en que algo puede aparecer como ente (1987, p. 204).
>
> Decir y ser, palabra y cosa, se pertenecen mutuamente la una a la otra de una manera velada aún, escasamente meditada e imposible de abarcar por ningún pensamiento (1987, p. 213).
>
> Y la idea que las expresiones verbales emergen de las prácticas sociales compartidas —los significados— y no al revés, puede verse también en:
>
> A las significaciones les brotan palabras, lejos de que a esas cosas que se llaman palabras se las provea de significaciones (1986, parágrafo 34, p. 180).

(4) Charles Guignon está en la misma línea de interpretación:

> La fenomenología heideggeriana de la agencia humana comienza con una descripción de la vida como un «acontecimiento» capturado en «interacciones» con instrumentos en contextos ordinarios. En nuestras actividades prerreflexivas, él sugiere, nos encontramos absortos en realizar acciones y en lidiar con situaciones que encontramos «significativas» en el sentido en que las cosas nos *importan* o *cuentan* para nosotros en maneras específicas. Lo que se nos muestra en tales contextos no es una colección de objetos brutos para ser representados, sino una totalidad de instrumentos organizados por nuestros proyectos en una red de medios/fines (1990, p. 654).

(5) Lo que Wittgenstein sugiere, sin embargo, es que nosotros solo podemos aprender las palabras (y, en consecuencia, captar lo que los objetos son) si ya tenemos una comprensión del mundo y una comprensión enraizada en una previa capacidad lingüística (Guignon, 1990, p. 358)[7].

(6) Ya que el significado es inseparable de las creencias, también es posible describir el significado de x como cierto estado mental en que se encuentran

[7] Hoy es ampliamente reconocida la influencia del pragmatismo estadounidense en Husserl y Heidegger, tanto de manera directa como indirecta (Quintanilla, 2014a, p. 39). También se puede rastrear la mutua influencia entre Henri Bergson y William James (Quintanilla, Escajadillo & Orozco, 2009, capítulos 1 y 2) y la manera en que ambos influyeron en la filosofía peruana a través de Pedro Zulen y otros autores (Quiroz, Quintanilla & Rojas, 2015).

los hablantes al proferir o interpretar x. Esta descripción es correcta siempre que se entienda que los estados mentales son inseparables de acciones, en tanto disposiciones para actuar, con lo cual no retornamos a una concepción ideacional o mentalista del significado.

(7) Para que en una comunidad se constituya el significado de x es necesario que se cree una regularidad en el uso de x. Es decir, es necesario que se produzcan ciertos efectos regulares —estados mentales tales como creencias, emociones, sentimientos, deseos, etcétera— relativamente predecibles en quienes profieren e interpretan x. A esos efectos relativamente predecibles y a las condiciones regulares que los producen también es posible llamarlos el significado de x.

En principio, el significado literal de una expresión, que es una abstracción o idealización a partir de ciertos casos concretos de uso, es independiente de las condiciones particulares de uso de la expresión. Como hemos visto en el capítulo anterior, a esto se suele llamar el «principio de la autonomía del significado» y alude a la distinción entre significado y uso, siendo esta una distinción de grado, como veremos en breve. El significado de x es la regularidad social que gobierna el uso de x en una comunidad de hablantes. El uso, por otra parte, es la aplicación que se puede hacer de x en una circunstancia concreta. Así, por ejemplo, yo puedo usar x para describir con verdad un hecho, para ser irónico, para hacer una metáfora, para orar, para pedir un favor, para mentir, para dar una orden, etcétera. En estos diversos actos de habla el significado de x será el mismo, es el uso de x el que variará. Sin embargo, si el uso cambia sistemáticamente el significado también cambiará. Puede imaginarse, entonces, un continuo en el que tenemos, en un lado del espectro, una mayor regularidad que gobierna el uso social de x; llamaremos entonces a esa regularidad el significado convencional de x y eso es lo que aparecerá en los diccionarios. Esta regularidad del uso de x produce ciertos estados mentales como efectos predecibles en los hablantes y las intérpretes de x. Como las creencias son disposiciones para actuar, la regularidad en el uso puede describirse también como un sistema de creencias compartidas por los hablantes acerca de x y acerca de lo aludido por x. En este lado del espectro tenemos una mayor determinación del significado de x. Para emplear la distinción de Wittgenstein, podríamos describir este lado como el ámbito del «decir». En el otro extremo del continuo, lo que tenemos es el uso irregular, que puede ser cierta práctica social que se está constituyendo acerca del uso de x, aunque no esté todavía plenamente constituida, o puede ser también un uso auténticamente original y novedoso de x. En este extremo del continuo, tenemos efectos —estados mentales— impredecibles y creencias no necesariamente compartidas por la comunidad. Es en este extremo del espectro que se producen los casos de cambio conceptual y creación de

significado, pues la imaginación y la creatividad aparecen, precisamente, en circunstancias en las que no hay una regularidad constituida acerca del uso de una expresión. Cuando un hablante, por azar o por genialidad, inaugura un uso exótico o extravagante de una expresión, produce efectos no previstos y creencias inusuales. Esta es la provincia de la metáfora y, en general, de los usos lingüísticos no literales que, a través del significado literal —de los usos regulares—, en circunstancias y contextos apropiados producen efectos inusuales que pueden producir cambio conceptual y creación de significado. Crear una metáfora es proferir una oración con significado literal en una circunstancia cuyo uso produce efectos impredecibles, curiosos y eventualmente iluminadores. Así, la distinción entre el uso literal y el uso metafórico de una oración es un asunto de pragmática y no de semántica; más aún, podría decirse que la semántica es parte de la pragmática. Asimismo, la muerte de la metáfora, es decir, el caso en que un uso extravagante se hace regular, es la creación de significado y el crecimiento del lenguaje. En el extremo del continuo que estamos describiendo encontramos mayor indeterminación del significado lo que correspondería, según la distinción wittgensteiniana, no al ámbito del decir sino al ámbito del mostrar.

Además, cuando una intérprete asigna condiciones de verdad a las proferencias verbales de un hablante, está utilizando como criterio para hacer tal asignación sus propias creencias, es decir, sus propias disposiciones para actuar. Para que la intérprete pueda encontrar condiciones de verdad en una proferencia ajena es necesario que, desde su punto de vista, asuma que comparte un número muy grande de creencias con el hablante. Esto significa que la intérprete encuentra en la proferencia ajena creencias familiares, disposiciones para actuar compartidas y, por tanto, regularidades sociales comunes. Para que ella pueda considerar que ciertos eventos naturales son proferencias verbales a las cuales puede atribuírseles condiciones de verdad, es necesario que ella pueda reconocer allí prácticas sociales compartidas. Esto demostraría que, en tanto uno pueda reconocer comportamiento en el otro y no meramente eventos naturales, uno estará ya interpretando al otro y, en consecuencia, estará integrando su propio sistema de creencias —las prácticas sociales y la forma de vida a la que pertenece— con el sistema de creencias del otro. Pero si bien en este proceso siempre hay cierto nivel de indeterminación del significado, no tendría sentido hablar de inconmensurabilidad.

A lo largo de estos dos últimos capítulos nos hemos concentrado en el significado que emerge a partir del uso regular, es decir, en uno de los extremos del continuo. Ahora debemos abordar el otro extremo, con lo cual nos preguntaremos acerca de la interpretación de las oraciones usadas metafóricamente. Al analizar la metáfora revisaremos nuevamente la naturaleza del significado, tanto a la luz del proyecto davidsoniano como más allá de este.

Capítulo seis
Significado, metáfora y cambio conceptual

6.1. Creencias, significados y formas de vida

Interpretamos el comportamiento de las personas, pero también sus proferencias verbales y sus textos escritos, pues una oración es una acción convertida en palabras. Algunas de las más complejas oraciones son las metáforas, porque ellas confrontan nuestros presupuestos y retan nuestra creatividad.

La metáfora se encuentra en el corazón del mito tanto como del logos y el crecimiento mismo del lenguaje depende de ella. El fenómeno metafórico es interesante en sí mismo como el caso paradigmático de la creatividad lingüística, pero es también interesante porque el análisis de la metáfora ilumina el problema del significado. Desarrollaré esa idea en este capítulo, a partir de los trabajos sobre significado e interpretación de Davidson en torno a los límites del discurso literal, es decir, en torno a la metáfora y a la posibilidad de que el discurso figurativo sea un instrumento capaz de expresar contenidos que se encuentran más allá de las posibilidades del discurso literal.

Comenzaré con la descripción de lo que Davidson entiende por significado e interpretación, luego mostraré cómo él aplica estas ideas al problema de la interpretación de una metáfora. Finalmente, me alejaré de sus tesis explícitas e intentaré mostrar las consecuencias que se desprenden de este análisis y cómo la reflexión sobre el fenómeno de lo metafórico ilumina la naturaleza de la interpretación y del significado.

En 1984, Davidson publicó, bajo el título de *Inquiries into Truth and Interpretation*, un grupo de artículos sobre temas que estaba trabajando desde fines de la década de 1950. Estos textos giran en torno a cuatro temas principales: las relaciones entre significado y verdad, el problema de la interpretación radical, las relaciones entre lenguaje y realidad, y los límites del discurso literal. Desde entonces, Davidson ha desarrollado una teoría acerca del significado, la naturaleza de la interpretación y el problema de la comunicación; es decir, una revolucionaria filosofía del lenguaje

que enriquece sus trabajos sobre filosofía de la mente y teoría de la acción que publicó en 1980 bajo el título de *Essays on Actions and Events*.

El proyecto de Davidson vincula, y tal vez supera, dos tradiciones coexistentes en la filosofía contemporánea del lenguaje. De un lado, aquella influida por Frege, Tarski, Carnap y Quine, que consiste en la aplicación de la semántica formal a los lenguajes naturales y que utiliza como instrumento principal el análisis de las condiciones de verdad con la finalidad de proporcionar el significado de una oración. De otro lado, aquella tradición vinculada a las relaciones entre hablantes e intérpretes, y los actos de habla que ellos pueden realizar. Esta tradición, que dirige su atención principalmente a las nociones de comunicación e intención fue desarrollada inicialmente por el segundo Wittgenstein, Austin y Grice.

Davidson une elementos de estas dos fuentes con la finalidad de proporcionar una teoría del significado que, al ir más allá de los objetivos tradicionales de la filosofía del lenguaje, pretende proponer una concepción de la comprensión humana. En tanto el propósito de su empresa es suficientemente amplio como para incluir una concepción de lo mental y una teoría de la acción, es justo decir que su proyecto es en espíritu hermenéutico, en el sentido de que se interesa por la pregunta general acerca de qué es interpretar y comprender a un ser humano y a sus producciones.

Como vimos en los capítulos anteriores, en sus trabajos tempranos Davidson abordó el problema de la interpretación radical —cómo es posible la comunicación entre individuos que carecen de todo lenguaje, tradición o contexto común—, con la finalidad de iluminar qué es lo que determina el significado de las expresiones lingüísticas. Él vio su proyecto como la construcción de una teoría del significado para lenguajes naturales individuales. Pronto descubrió, sin embargo, que el problema de dar el significado de una expresión requiere de fijar simultáneamente las creencias y deseos que son inseparables de los significados en las proferencias de hablantes específicos. El proyecto ya era en sí mismo holista pero todavía tenía una fuerte influencia logicista.

En «A Nice Derangement of Epitaphs» (2005a [1986]), Davidson amplió el holismo y postuló una teoría de la comunicación en la que no se presupone que para que haya comunicación entre hablantes estos tengan necesariamente que compartir un lenguaje, un código previo o un conjunto de convenciones[1]. Lo que Davidson sugiere es la necesidad de revisar radicalmente los conceptos tradicionales de comunicación y lenguaje, y entender la comunicación como el proceso en el que los hablantes interpretan a sus interlocutores al atribuirles sus propios sistemas de estados mentales, para luego hacer modificaciones en las atribuciones.

[1] Begby y Ramberg han editado un número completo de *Inquiry* de 2016 dedicado a discutir este influyente artículo.

En ese artículo Davidson distingue, en el proceso de la interpretación, entre teorías previas y teorías al paso. Esto ya lo vimos en el cuarto capítulo, pero por su importancia lo volveremos a abordar de una manera diferente. Para la intérprete, la teoría previa es cualquier cosa que ella sabe —es decir, es una amalgama de creencias acerca del lenguaje y creencias acerca del mundo— y que le permite entender las proferencias de un hablante particular. Esto hace que, en principio, ella esté mejor preparada para entender a ciertos hablantes en vez de otros. Su teoría al paso expresará la manera en que ella interpreta las proferencias del hablante. En cierto sentido, su teoría al paso será la aplicación de su teoría previa a ciertos específicos hablantes, situaciones y proferencias. De otro lado, para el hablante la teoría previa es todo lo que él cree acerca de la teoría previa de la intérprete, es decir, la manera en que el hablante se dirige a la intérprete en su esfuerzo por ser comprendido. Su teoría al paso, por otra parte, expresará la teoría que el hablante intenta o desea que la intérprete utilice al entender sus proferencias. Es decir, la teoría al paso del hablante mostrará la clase de ajustes o cambios que él desea que ella haga en su propia teoría previa con la finalidad de poder comprenderlo a él.

Un importante tema de investigación, aunque alejado de nuestro propósito inmediato, es el de mostrar las semejanzas entre este análisis y la manera en que Gadamer (1977b, 1992) concibe la comprensión de un texto. No vamos a detenernos en ello, pero sí me gustaría sugerir un par de relaciones. Mientras para Davidson el paradigma de la interpretación es la comunicación entre hablantes, para Gadamer lo es la lectura de un texto. Para este último, al leer establecemos un diálogo en el que nuestro interlocutor es el texto mismo, no su autor, y entablamos una conversación en la cual su sentido último no está dado únicamente por el supuesto sentido literal —y unilateral— del texto, sino por lo que el texto tiene que decirnos y nuestra interpretación de ello. En esto Gadamer se aleja de la hermenéutica intencionalista de Schleiermacher (1986, 1996) y Hirsch (1967, 1978).

La comprensión tendría como resultado que la lectura nos permite ver en el texto y en nosotros mismos aspectos que no habíamos percibido anteriormente. Así como en Davidson, para Gadamer la intérprete tiene en mente un modelo de lectura —una anticipación de sentido y un conjunto de prejuicios, es decir, una teoría previa— que no es arbitrario, sino que está determinado por su pertenencia a una comunidad y a una tradición. En la medida en que la intérprete aplica el modelo al texto, este se ve corroborado o contradicho por aquel, y el modelo se modifica, pero al modificarse altera su comprensión del texto y, además, su comprensión de ella misma al iluminar sus propias anticipaciones de sentido. Tanto para Davidson como para Gadamer no existe tal cosa como un significado real del texto o de las palabras del hablante en sí mismos, pues el significado debe ser completado por la situación histórica de la intérprete.

Esta comprensión del sentido del texto o de las palabras del hablante es, en realidad, la tensión suscitada por nuestra cercanía y nuestra distancia respecto del sujeto de nuestra interpretación. Hay, pues, entre intérprete y autor una cooperación interpretativa de la que emergerá el significado (Eco, 1981).

Adquirimos una teoría previa al habitar una forma de vida o al estar inmersos en un horizonte. Pero el punto es que, en sentido estricto, el significado no pertenece a la provincia de las teorías previas. En efecto, la noción de teoría previa es solo una idealización de lo que da continuidad a las diferentes teorías al paso que podemos desarrollar. El significado emerge en la confrontación entre teorías al paso —en la fusión de horizontes, diría Gadamer—. Así, para que la comunicación sea exitosa no se requiere que los hablantes compartan la misma teoría previa. Lo necesario es que ambas teorías al paso puedan tocarse en algún punto, pues finalmente esa va a ser la comunicación, el efímero roce de teorías al paso.

En un primer momento es importante notar que la interpretación que la intérprete hace de los significados de las palabras del hablante, comienza antes que ella empiece a comprender la manera en que él está utilizando las palabras en una situación específica. Esta es la distinción entre significado y uso, a la cual Davidson ha denominado el «principio de la autonomía del significado» que, aunque importante, es una distinción de grado. Debemos afirmar que tanto el significado primario o literal —el que el hablante tiene en cuenta durante la comunicación— como el significado convencional —el que se encuentra en los diccionarios— se mantienen estables incluso si el hablante cambia su propósito al usar las palabras o si las usa en diferentes contextos o teniendo en cuenta diferentes efectos deseados en la intérprete, pues de otra manera la noción de significado se vería trivializada y perdería todo contenido.

La intérprete atribuye significados y creencias a las proferencias del hablante haciendo hipótesis que toman la forma de acuerdos tácitos acerca de cómo usar las palabras. Estas son las oraciones-T que hemos discutido previamente, es decir, oraciones que tienen la siguiente estructura lógica:

(T) «S» es verdadera en L si y solo si p.

Donde p es una oración nombrada por s, si el lenguaje objeto es parte del metalenguaje.

La intérprete también atribuye deseos al hablante haciendo hipótesis acerca de lo que él quisiera que fuese el caso, es decir, acerca de la clase de oraciones que él quisiera que fuesen verdaderas. Estas hipótesis y acuerdos se hacen mediante una triangulación entre el hablante, la intérprete y el mundo que comparten.

La interpretación del comportamiento verbal tiene dos momentos. En primer lugar la intérprete debe preguntarse qué significan en su sentido habitual las palabras usadas por el hablante y, en segundo lugar, qué es lo que él desea hacer al utilizarlas. Primero, ella busca los significados literales y solo después se pregunta por la acción que el hablante deseó realizar o por el significado que él intentó expresar en esta situación y contexto específicos, si es que ella tiene razones para suponer que en esta ocasión el significado literal y el significado intencional divergen. Esto ocurre en cualquier caso de interpretación lingüística, lo que incluye la interpretación de metáforas, solo que con las metáforas este doble proceso es más obvio.

Davidson (2005a [1986], pp. 92-93) afirma que el significado literal o primario es proferido por un hablante específico en una ocasión particular, pero si el emisor y la intérprete son hablantes promedio, el significado primario será lo que encontramos en un típico diccionario, es decir, el significado convencional. Lo que se está sosteniendo, entonces, es que los significados primarios dependen de la relación entre hablante e intérprete. Si son hablantes promedio, el significado literal o primario emergerá en la confrontación entre sus teorías al paso como la apropiada interpretación que la oyente hará de las palabras del hablante. Esta interpretación estará muy cerca de sus teorías previas, es decir, de lo que diría otro hablante promedio o un diccionario típico. Pero si nuestros personajes no son hablantes promedio, el significado literal emergerá de la confrontación entre la teoría al paso de la intérprete y la intención del hablante de ser interpretado de un modo y en una situación particular. En este caso la intersección entre ambas teorías al paso se encontrará muy alejada de sus teorías previas. Sin embargo, la comunicación podrá ser exitosa siempre que se llegue a producir la confluencia entre teorías al paso.

Los hablantes también son intérpretes de sí mismos. Si estoy leyendo un texto que escribí hace muchos años puedo necesitar desarrollar una teoría al paso para entender significados, creencias y deseos que ya no tengo. Pero si interpreto las palabras que estoy profiriendo en este mismo instante, mi teoría al paso coincidirá con mi teoría previa. Aunque Davidson no acepta que tenga sentido decir que uno se autointerpreta, no tengo inconvenientes en afirmar que todo hablante es un intérprete de sí mismo y, como tal, puede errar en la manera en que entiende sus propias palabras y creencias, comparadas con la interpretación promedio o como él mismo podría hacerlo en otro momento de su vida. El autointérprete tiene, sin embargo, lo que Davidson llama «la autoridad de la primera persona», es decir, la interpretación que el hablante hace de sus propias palabras es sincrónicamente inmejorable, mientras que la interpretación que uno hace de las palabras ajenas es siempre perfectible (Davidson, 2001a [1984]).

Esto hace que en el caso de autointerpretación sincrónica la indeterminación del significado sea mínima y pueda, incluso, parecer inexistente, aunque nunca lo sea. La indeterminación se hace mayor cuando hay más distancia, en términos de significados, creencias y deseos, entre las teorías previas de los interlocutores. Aunque la indeterminación del significado nunca puede ser total siempre puede ser mayor.

Cuando el hablante profiere oraciones que la intérprete tiene dificultades en comprender —ya sea porque el hablante atribuye significados distintos a las palabras o, lo que es lo mismo, porque tiene creencias diferentes acerca de los objetos referidos por ellas— la intérprete tiene que hacer un esfuerzo por producir toda suerte de hipótesis —acerca de las creencias del hablante y el significado que da a sus palabras— para poder conferir algún sentido a su conducta y sus proferencias. En este proceso una buena intérprete deberá desarrollar toda su creatividad e imaginación para acceder a creencias y significados que le pudieran resultan ajenos o eventualmente absurdos, mientras que una intérprete limitada entenderá solo aquellas proferencias que contengan creencias y significados semejantes a los suyos, con lo cual este tipo de intérprete perderá el contenido del mensaje del hablante. Después de este proceso en que la habilidad de la intérprete en producir hipótesis creativas ha sido desafiada, ella incorporará algunos de los resultados a su sistema de creencias y a la teoría previa construida para este específico hablante, es decir, habrá adquirido nuevas creencias acerca del mundo y, entonces, algunos de los significados que ella solía atribuir a las palabras habrán cambiado. Su habilidad para desarrollar nuevas teorías al paso se hará más sutil para nuevos y más difíciles casos. Aunque Davidson no aborda este tema, su concepción de la producción de teorías al paso puede ser un buen modelo explicativo del fenómeno del cambio conceptual, como lo veremos pronto. En condiciones normales cambiamos nuestras creencias —y, por tanto, los significados que atribuimos a las palabras— cuando encontramos un interlocutor —que puede ser una persona, un libro, una película o un poema— que cuestiona y desafía nuestros presupuestos y prejuicios con sus propias creencias y comportamiento.

Algo así acontece cuando estamos expuestos a un tipo de discurso que cuestiona nuestras creencias y supuestos previos. Cuando un científico newtoniano lee a Einstein, un geómetra euclidiano lee a Riemann o un filósofo encuentra a otro de concepciones diferentes. Un caso paradigmático en que nuestras creencias y el significado que atribuimos a las palabras se ven cuestionados es cuando estamos expuestos a una metáfora que nos permite ver algún objeto o problema bajo una luz diferente, es decir, que nos sugiere ver algo como algo, o que nos permite reparar en aspectos del mundo o de nosotros mismos de los que no habíamos sido conscientes. Esta situación produce cambio conceptual, es decir, variación en creencias y significados y, en ocasiones, también creación de nuevos significados y creencias, esto es, creencias

que nadie tuvo anteriormente. Desde un punto de vista cognitivo, la cualidad de una metáfora se encuentra en directa proporción a la posibilidad de cambio conceptual e iluminación que pueda provocar en las intérpretes. Esto nos permite ver al mundo, a nuestro interlocutor o vernos a nosotros mismos bajo otra luz, de una manera diferente, lo que nos produce la sensación de estar viendo un nuevo ángulo de las cosas o de estar descubriendo algo que ahora consideramos importante, valioso o profundo. Por supuesto, desde este punto de vista la cualidad de una metáfora será relativa a las creencias y otros estados mentales de la intérprete, pero tales estados mentales no pueden ser demasiado diferentes de una intérprete a otra si comparten de manera general los significados regulares de las palabras.

Una de las ideas que estoy desarrollando es que el significado no es algo que provenga únicamente del hablante. Para que haya significado necesitamos una intérprete. Así pues, diferentes intérpretes pueden asignar diferentes significados a las mismas proferencias y, aunque puede haber mejores y peores interpretaciones, no existe tal cosa como la única interpretación correcta, porque no hay tal cosa como el verdadero significado de una oración o una palabra. Este es el principio de la indeterminación de la interpretación, que está en el corazón mismo del fenómeno hermenéutico. A su vez esto nos conduce a la provincia de la metáfora, pues las metáforas son el caso paradigmático de oraciones que pueden ser ampliamente interpretables sin que exista un criterio fijo de interpretación. En este sentido, ellas representan un caso de severa indeterminación del significado. Por ello deberemos preguntarnos ahora cuál es la relación entre metáfora y significado.

6.2. ¿Qué significan las metáforas?

Una buena interpretación de una metáfora debe ser original y penetrante, pues no se trata simplemente de intentar reproducir las intenciones del hablante sino de completar el proceso de la creación de significado. Una buena intérprete de metáforas debe ser tan original y creativa como el mismo autor de la metáfora, tal como Davidson dice que ocurre en cualquier proceso de comunicación en que la teoría al paso de la intérprete debe ser suficientemente imaginativa para captar las proferencias del hablante. No hay reglas o principios que puedan guiar a la intérprete de metáforas, de la misma manera que no hay reglas para la producción de metáforas o teorías al paso. La interpretación de una metáfora es un acto de creatividad y la creatividad no consiste en seguir reglas sino en proponerlas. La construcción de teorías al paso es precisamente la confección de reglas de atribución de significado y no hay ninguna regla que explique cómo inventamos reglas de interpretación. Por ello, la interpretación de una metáfora muestra cómo deberíamos interpretar

toda oración, incluso aquellas no metafóricas. De otro lado, es como si la interpretación de una metáfora fuese un caso en miniatura de interpretación radical. Voy a intentar desarrollar estas intuiciones.

La concepción tradicional de la metáfora, aquella que procede de Aristóteles (1968, 2002), sostiene que una metáfora es una oración respecto de la que no estamos concernidos por su significado literal —es decir, por la interpretación ordinaria que de ella haría una intérprete promedio—, sino por un significado peculiar, inusual y tal vez extravagante que, sin embargo, nos proporciona algún tipo de iluminación o información —sea acerca del hablante, acerca de nosotros como intérpretes o acerca del mundo— que la oración interpretada en su sentido literal no puede producir. La idea central en esta concepción aristotélica es que, al lado del significado literal de una metáfora hay un segundo significado metafórico. Este emergería como (1) una «transferencia de significado» de un concepto a otro, en la que atribuimos a un sujeto características que literalmente no le corresponden con la finalidad de llamar la atención a una similitud; o como (2) una «extensión de significado», en la que tomamos una expresión con una extensión determinada y la aplicamos a una nueva e inusual extensión, también con la finalidad de llamar la atención a ciertas similitudes.

El ejemplo de Aristóteles es la metáfora homérica «Aquiles es un león», en la que el poeta transfiere ciertos rasgos que pertenecen naturalmente a los leones —braveza, ferocidad, valor, etcétera — a Aquiles, para hacer que el lector advierta que Aquiles es como un león en poseer tales cualidades. La transferencia de significado está acompañada de extensión de significado, porque estamos ampliando el sentido ordinario de «león» para predicar algo acerca de un hombre. En ocasiones el significado extendido se convierte en usual y entonces la metáfora muere para convertirse en una oración literal. Esta concepción de la metáfora, que proviene directamente de Aristóteles, se encuentra presente en autores como Richards (1936), Black (1977, 1979, 1981) y Ricœur (1981, 1986), entre otros.

Davidson coincide en que en el fenómeno metafórico se produce un desplazamiento de características que pertenecen ordinariamente a un contexto hacia otro, pero se opone a la idea de que las metáforas se constituyan como un segundo significado. Piensa que «las metáforas significan lo que las palabras significan en su más literal interpretación y nada más» (1984h [1978], p. 245). Afirma que cuando proferimos una metáfora significamos las palabras en su sentido literal, pero de una manera que produce o sugiere en la intérprete, y por supuesto también en el hablante, sensaciones, emociones o pensamientos inusuales e inesperados. Para él es importante mostrar que los efectos que la metáfora produce no pueden ser su significado porque, en tanto tales efectos son, en principio, varios e inagotables, la misma noción de significado se vería trivializada. Este es nuevamente el principio de la autonomía del significado.

La tesis davidsoniana depende de la distinción entre lo que las expresiones lingüísticas significan para los hablantes ordinarios y cómo los hablantes las usan con la finalidad de producir ciertos efectos en intérpretes específicas, es decir, la distinción entre semántica y pragmática. La distinción entre significado y uso es crucial porque solo podemos asignar condiciones de verdad —oraciones-T— a oraciones en abstracto, esto es, independientemente de sus contextos de uso o de las razones de los hablantes para proferirlas. Si asignásemos significado a las oraciones considerando sus contextos particulares de uso, ya que estos son, en principio, infinitos en variedad, la noción de significado perdería todo contenido y no significaría nada en absoluto. Pero esto no implica que las oraciones tengan significados literales en sí mismas independientemente de sus contextos de uso. Lo que se está afirmando es que las condiciones de verdad de las oraciones literales se asignan a casos ordinarios de uso y no a situaciones especiales.

Hay, sin embargo, una continuidad entre significado y uso. El significado literal de una expresión es lo que podemos hacer con esa expresión —la manera en que podemos usarla— en una comunidad de hablantes independientemente de situaciones particulares. Así pues, en principio, los significados literales son previos a cualquier contexto particular de comunicación y el significado literal de una expresión es el conjunto —siempre difuso— de regularidades familiares en el uso de esa expresión en los diferentes procesos comunicativos en los que pueda participar.

Pero la distinción entre un caso ordinario de interpretación y uno especial, como el metafórico, no es radical, sino más bien un asunto de grado. Una metáfora se hace familiar cuando se congela para convertirse en una oración literal. Las metáforas pertenecen a los bordes crecientes del lenguaje y el significado no es más que una noción abstracta que empleamos para poder hacer inteligibles ciertos tipos familiares de comportamiento humano; el uso metafórico pertenece a los bordes de esas familiaridades. Rorty lo pone de la siguiente manera:

> Davidson […] piensa que nociones semánticas como «significado» tienen un papel solo al interior de los estrechos (aunque crecientes) límites del regular y predecible comportamiento lingüístico —los límites que marcan (temporalmente) el uso literal del lenguaje. […] Decir, como lo hace Davidson, que «la metáfora pertenece exclusivamente al dominio del uso» es decir simplemente que, ya que las metáforas (mientras todavía vivas) no son parafraseables, caen fuera del área clarificada (1991d, pp. 163-164).

Lo que llamamos «significado literal de una oración» es el uso ordinario del hablante ordinario; sin embargo, lo que realmente está ahí recogido es la intersección de muchas teorías al paso. Al aceptar significados literales no estamos contradiciendo

la tesis ya expuesta de que es solo en el proceso de interpretación en el que el significado emerge y que solo hay significado allí donde hay hablantes e intérpretes. Aunque el significado literal es, en principio, previo a cualquier proceso particular de interpretación y a cualquier sistema de creencias, fue constituido originalmente en algún proceso particular de interpretación que es ahora irrelevante para nuestra intención de uso.

Imaginemos el siguiente escenario: Romeo está caminando por las calles de Verona y se encuentra con Benvolio, quien le pregunta a dónde se dirige. Romeo contesta que va a ver a Julieta. Benvolio pregunta quién es Julieta. Entonces, Romeo contesta: «Julieta es el Sol». Lo primero que Benvolio hará es interpretar la oración en su sentido literal: afirmará que Julieta es un astro incandescente. Pronto se dará cuenta, sin embargo, que Romeo no puede haber querido decir eso de manera literal y se aplicará a buscar posibles asociaciones entre Julieta y el Sol para poder hacer inteligible la situación comunicativa en que se encuentra. Así, y en gran medida esto dependerá de su habilidad como intérprete, Benvolio podrá suponer que Romeo quiso mostrarle que Julieta es fuente de calor y vida —las que, por otra parte, siguen siendo metáforas— o que Julieta es hermosa y dorada. Podrá suponer, también, que Julieta es el centro alrededor del que la vida de Romeo gira, si Benvolio es heliocentrista, porque si es geocentrista pensará, más bien, que Julieta gira alrededor de Romeo, lo que sin duda tendrá consecuencias muy diferentes para su relación sentimental. Pero también podría ser que Benvolio piense que Romeo le está dando a entender que Julieta es una rubia con sobrepeso o que hay ciertas horas del día, especialmente en verano, cuya presencia es intolerable. Todo esto es parte de lo que la metáfora muestra, aunque para que la metáfora muestre apropiadamente se necesita un productor agudo y una intérprete ingeniosa, pues la metáfora solo termina de nacer cuando la intérprete imagina las posibles asociaciones que ella indica.

El caso de «Julieta es el Sol» es hoy bastante simple y hasta quizá algo cursi, aunque ciertamente no lo era en tiempos de Shakespeare[2]. Por ello, piénsese en una metáfora más acorde con nuestros tiempos. Imagínese que Romeo dice: «Julieta es un agujero negro». Hay una infinidad de posibles asociaciones que podríamos hacer entre Julieta y los agujeros negros, y, en gran medida, la capacidad de Benvolio como intérprete dependerá de lo que sepa sobre los agujeros negros, de lo que crea acerca de Julieta y de lo que Benvolio suponga que Romeo cree acerca de Julieta y los agujeros negros. Romeo puede querer decir que Julieta le atrae tanto que nada de él puede escapar a su gravedad. Pero también podría sugerir que Julieta es posesiva y destructiva.

[2] El ejemplo está tomado del acto II, escena 2, líneas 1-4, en el que Romeo dice: «*But, soft! What light through yonder window breaks? It is the east, and Juliet is the sun!*» (Shakespeare, 1993, p. 772).

Aunque también podría insinuar que ella es incansable y desenfrenada. Ahora supongamos que inmediatamente después de decir que Julieta es un agujero negro, Romeo cita a Steve Allen y dice: «La comedia es tragedia más tiempo». Sin duda deberemos interpretar la primera metáfora a la luz de la segunda, pues es como si Romeo hubiera hecho una metametáfora. Probablemente nos inclinaremos, por tanto, a una interpretación más dramática, aunque algo jocosa y seguramente menos agradable para Julieta, de la metáfora «Julieta es un agujero negro». Pero volvamos a nuestra metáfora inicial. Cuando Romeo dice que Julieta es el Sol, lo que nos está pidiendo es precisamente que veamos a Julieta a la luz del Sol.

Así pues, el fenómeno metafórico es importante desde varios puntos de vista: es interesante porque ilumina el ámbito de lo que no puede ser dicho, permite aclarar el fenómeno del cambio conceptual y la creación de nuevo significado, explica cómo el significado se constituye mediante la participación solidaria entre hablante e intérprete y también muestra cómo se constituye el discurso literal.

Ahora bien, Davidson no afirma que los efectos que una oración puede producir en un proceso particular de interpretación, es decir, en cierta intersección de teorías al paso, debería ser denominado «significado». Él diría que las palabras «Julieta es el Sol» proferidas por Romeo solo tienen significado literal y que la particular interpretación que alguien haga de ellas no constituye un segundo significado figurativo. Es importante establecer esto porque si la interpretación de Benvolio constituyese un segundo significado respecto del literal, tendríamos que decir que las maneras en que Capuleto, Montesco, Mercucio, Baltasar y Teobaldo interpretan las palabras de Romeo constituyen significados superpuestos al literal. Más aún, el significado de una oración dada variaría según el estado mental y las creencias y deseos actuales de las posibles intérpretes. Esto ciertamente conduciría a una completa trivialización de la noción de «significado» en la que ella ya no significaría nada en absoluto. El punto es que todos aquellos efectos que una proferencia puede producir en una intérprete son estados mentales, algunos de los cuales pueden ser estados cognitivos como creencias, mientras que otros son emociones o simples sensaciones. Entre otras cosas, solemos llamar significado literal a los estados mentales cognitivos familiares que una expresión produce en condiciones ordinarias en una comunidad de hablantes.

Esta concepción del significado como efectos ordinarios y familiares se aleja de las teorías semánticas clásicas, como la teoría referencialista desarrollada por Frege (1984), el primer Wittgenstein (1975) y Russell (1905) y también de la teoría ideacional sostenida por Locke (1982). Más bien, se acerca a las intuiciones acerca del significado presentes en las *Investigaciones filosóficas* de Wittgenstein (1988) y, curiosamente, hay quienes también han encontrado asociaciones con Derrida (Wheeler, 1992). En todo caso, creo que es posible afirmar que la filosofía del lenguaje de Davidson es una versión

depurada y explícita de muchas intuiciones acerca del significado que se encuentran presentes en el segundo Wittgenstein.

Davidson afirma que si las oraciones usadas metafóricamente tuviesen un significado figurativo además del literal, habría dos posibilidades: o bien (1) aquellos significados metafóricos serían parafraseables en significados literales, con lo cual no tendríamos significado metafórico en absoluto, sino solo significado literal expresado de una forma extravagante, o (2) los significados metafóricos no serían parafraseables en significados literales. Pero, si así fuera, no tendríamos absolutamente ninguna razón para decir que ahí hay significado. Como puede verse, lo que Davidson rechaza es la tesis aristotélica de que el significado metafórico emerja en la transferencia y extensión del significado literal. Su posición es que podemos extender cuanto queramos el significado literal de una expresión y siempre tendremos como resultado significado literal.

Pero, entonces, tenemos que preguntarnos más radicalmente qué es una metáfora. Una metáfora es simplemente una oración literal, a veces falsa, a veces absurda, a veces obviamente verdadera, que tiene la característica de inspirar en una intérprete algunas emociones, sentimientos, sensaciones o cierto grado de conciencia respecto de algunos rasgos del mundo, de su interlocutor o de ella misma, sugiriendo originales e interesantes comparaciones entre los sujetos de la oración. Ya que estas inspiraciones son, en principio, infinitas, no hay reglas ni teoría que puedan explicar cómo funciona una metáfora ni cómo puede producir el tipo de iluminación que produce.

Para que la metáfora produzca los efectos que produce, la intérprete tiene que entender las palabras en su sentido literal. Considérese nuevamente la oración «Aquiles es un león». Los significados ordinarios de las palabras clasifican objetos de cierta manera, por ejemplo Aquiles y leones. Para que la intérprete pueda captar la metáfora debe ser capaz de entender las palabras «Aquiles» y «león» en su sentido ordinario, para después comenzar a buscar relaciones y comparaciones entre ellos. Pero, por supuesto, todo lo que se dice sobre la pragmática de la metáfora es también cierto acerca de la pragmática de las oraciones literales. Esto es reconocido por el propio Davidson:

> Sostengo que el carácter inconcluso de lo que llamamos el parafraseo de la metáfora brota del hecho de que pretende explicitar lo que la metáfora nos hace ver, y en esto no hay ningún término claro. *Diría lo mismo de cualquier otro uso del lenguaje* (1984h [1978], p. 263).

Desde el punto de vista de la intérprete, la diferencia entre interpretar una oración literal y una metáfora es la siguiente: al interpretar una oración familiar y bien contextualizada, la intérprete no tiene ninguna motivación para buscar algún

otro uso diferente del ordinario. Pero si el hablante profiere una oración obviamente falsa o absurda en un contexto razonable que no le hace pensar a la intérprete que el hablante ha perdido temporalmente la razón, la intérprete entenderá la oración en su sentido literal y, después de advertir su extravagancia, asumirá que el hablante tenía una motivación especial para proferirla; ya sea si el hablante quiso ser irónico, agudo o metafórico. Así, la intérprete comenzará la tarea de buscar esa motivación ulterior y empezará por considerar semejanzas interesantes entre los sujetos de la oración. Eventualmente, esta búsqueda producirá en ella una especial agudeza o iluminación acerca del tema de la metáfora. Por supuesto esto también podría ocurrir al escuchar una oración literal, una pieza de jazz o al observar una pintura. La diferencia radicará en que la oración literal, la pieza de jazz, la pintura y la metáfora dirigirán nuestra atención a diferentes cosas y diferentes relaciones posibles. Así pues, en realidad no hay ninguna diferencia de principio entre interpretar una oración usada literalmente o metafóricamente. Mientras una oración literal es aquella en la que las creencias y deseos son fácilmente interpretables a partir de la información que del hablante y de la lengua empleada tenemos en nuestra teoría previa, una oración usada metafóricamente es aquella en que las posibilidades de interpretación son más variadas y nos vemos obligados a desarrollar teorías al paso más creativas para darles sentido. Al decir esto, lejos de quitar valor cognitivo a la metáfora lo que estamos haciendo es mostrar que desde un punto de vista semántico no hay diferencia alguna entre una oración literal y una metáfora, pues toda la diferencia radica en las condiciones de su uso.

Desde luego, toda esta concepción de la metáfora depende de admitir que las metáforas congeladas dejan de ser metáforas para convertirse en oraciones literales porque, en general, las oraciones literales son en muchos casos metáforas muertas. Curiosamente, la suerte de una metáfora exitosa, es decir aquella a la cual el uso convierte en ordinaria, es la extinción como metáfora. Pero si las metáforas son el aspecto más creativo del discurso, su muerte es el crecimiento del lenguaje. En un contexto diferente Quine escribe lo siguiente:

> La metáfora, o algo semejante, gobierna tanto el crecimiento como nuestra adquisición del lenguaje. Lo que se convierte en un refinamiento subsecuente es el discurso cognitivo mismo, en su forma más literal y seca. Las pulcramente trabajadas extensiones de la ciencia son el espacio abierto en la jungla tropical, creados al eliminar los tropos (1978, pp. 161-162).

En la imagen de Quine, las oraciones literales son estas zonas pulcramente trabajadas zonas en las que el hábito y la regularidad terminan por imponer orden en medio de la exuberancia, mientras que las metáforas son precisamente la frontera con la jungla.

Las metáforas se solidifican y se convierten en oraciones literales y, así, el claro crece con la incorporación de nuevas regularidades. Las oraciones literales están conformadas por usos regulares de expresiones que generan efectos convencionales y fáciles de predecir en los intérpretes. Las oraciones metafóricas, por el contrario, son el producto de la subversión o de la trasgresión de las regularidades que gobiernan de manera convencional el uso de las expresiones en una comunidad de hablantes dada. Pero entre una oración usada literalmente y una usada metafóricamente lo que hay es un continuo, no una separación nítida. Las metáforas pertenecen a los bordes crecientes del lenguaje y el significado no es más que una noción abstracta que empleamos para poder hacer inteligibles ciertos tipos familiares de comportamiento humano; el uso metafórico pertenece a los bordes de esas familiaridades.

Confeccionar una buena metáfora es producir una oración que puede alertarnos a considerar algunas cosas bajo una luz diferente. Decir, con Goethe, que «la arquitectura es música congelada» nos insinúa el hecho de que la arquitectura, como la música, es un juego de formas, detalles, equilibrios y contrastes; pero que, a diferencia de la música que se distiende en el tiempo, la arquitectura es un juego que se extiende en el espacio. La música es continuidad en el flujo, la arquitectura es presencia y permanencia. La arquitectura es, pues, música congelada. Pero nada de esto es el significado de esa metáfora, es solo la descripción de algunas de las sensaciones, emociones o intuiciones que la metáfora puede producirnos.

Hasta aquí hemos intentado explicar cómo afecta en la intérprete una metáfora penetrante. Pero, ¿por qué es que proferimos metáforas? Una primera respuesta es que al hacerlo producimos efectos que el discurso literal es incapaz de expresar. Sin embargo, ¿qué clase de habilidad, saber o industria es aquella que nos permite elaborar buenas metáforas? En realidad es imposible explicar cómo es que llegamos a pensar una metáfora original. Es simplemente un asunto de creatividad y la creatividad no puede ser explicada porque entonces, al decir de Rorty, los poetas y los genios serían superfluos. Como dice Davidson de la producción de teorías al paso y de teorías en general, es solo un asunto de «ingenio, suerte y sabiduría» (2005a [1986], p. 107).

Ahora bien, podríamos ver la relación entre oraciones literales y metafóricas en comparación con la distinción entre «decir» y «mostrar», que acuñó el primer Wittgenstein (1975). Para este autor, uno emplea el lenguaje en el ámbito del decir (*sagen*) cuando lo usa para describir la manera como es o como puede ser el mundo, o para comunicar contenidos que pueden tener forma proposicional, es decir, que pueden ser verdaderos o falsos. En cambio, uno emplea el lenguaje en el ámbito del mostrar (*zeigen*) cuando lo usa para producir en el interlocutor estados mentales que no describen estados de cosas y que no pueden ser formulados proposicionalmente

y, por tanto, no son pasibles de verdad o falsedad. Cuando mostramos, evocamos en nuestro interlocutor estados mentales que no podrían ser expresados mediante el uso representacional del lenguaje.

Podríamos así sugerir —cosa que no hizo el propio Wittgenstein— que el ámbito del decir está asociado al discurso literal, mientras el ámbito del mostrar lo está al discurso metafórico. Una metáfora no describe estados de cosas sino muestra algunas de las, en principio, infinitas asociaciones que puede haber entre dos o más conceptos. Al hacer eso, la metáfora exhibe lo que no puede ser dicho; sugiere, evoca o insinúa lo que no puede ser descrito de manera proposicional. Una metáfora nunca puede volcarse en un conjunto, por largo que este sea, de oraciones literales, porque la metáfora alude a muchas posibles relaciones entre los conceptos involucrados, algunas de las cuales estaban en la mente del autor de la metáfora, de manera consciente o inconsciente, pero muchas otras no lo estaban sino solo aparecerán en la mente de quien la interpreta creativamente.

Según un célebre aforismo de Heráclito, el oráculo de Delfos no afirma ni niega, sino *semainein*, expresión que significa señalar o indicar, en el sentido de mostrar o insinuar. Algunos traductores de Heráclito al inglés traducen *semainein* por *to intimate*, intimar, lo que genera la connotación de que el oráculo no solo muestra sino además lo hace con la cercanía de la intimidad, es decir, familiariza. El oráculo de Delfos solía expresarse con metáforas, nunca de manera literal, y lo que hace Heráclito es una metáfora de la metáfora.

Pero, ¿por qué el uso metafórico del lenguaje tiene un poder que el uso literal no tiene? ¿Qué cualidad cognitiva o epistémica tiene la metáfora que va más allá de su valor estético? Como vimos, la respuesta es que una metáfora aguda nos permite ver algún objeto o problema bajo una luz diferente, es decir, nos permite ver algo como algo distinto, nos permite reparar en aspectos del mundo o de nosotros mismos de los que no habíamos sido conscientes. El segundo Wittgenstein llamaba «representación perspicua» o «ver aspectos» (1988) al hecho de que siempre vemos una situación o un objeto a la luz de otros, pues nunca vemos ni pensamos los objetos o fenómenos aislados de otros que también tenemos en cuenta. Por eso Aristóteles en la *Poética* dijo que «lo más importante con mucho es dominar la metáfora. Esto es lo único que no se puede tomar de otro, y es indicio de talento; pues hacer buenas metáforas es percibir la semejanza» (1459a 3-8).

Pero una metáfora aguda no solo nos permite ver relaciones entre objetos o ideas que no habíamos visto antes, también puede generar cambio y creación conceptual, de manera que ese será el tema que abordaremos ahora. Voy a concentrarme en un tema puntual que ha sido abordado por Paul Ricœur y que no ha sido desarrollado por Davidson. Trataré de mostrar cómo el enfoque davidsoniano acerca

del significado y la metáfora puede servir como punto de partida para iluminar los fenómenos del cambio y la creación conceptual. El análisis que Davidson hace del proceso en el que una intérprete desarrolla teorías al paso para hacer inteligible el comportamiento verbal inusual puede ser visto como el proceso en que una persona desarrolla nuevas intuiciones o teorías para hacer inteligibles nuevos aspectos de la realidad. El caso paradigmático de esta situación es la interpretación de una metáfora que, al cuestionar nuestras creencias, nos obliga a producir nuevas y más creativas teorías al paso, con lo cual transforma nuestra visión del mundo y nos transforma a nosotros con ella. Intentaré, entonces, conciliar la afirmación de que no existe un significado metafórico con la tesis de que la metáfora puede ser instrumento de cambio y creación de significado.

En lo que sigue de este capítulo, delimitaré el territorio y me detendré un momento en las relaciones entre creencia, significado y formas de vida. Esto nos conducirá al problema del cambio y la creación de creencias y significados. Finalmente, abordaré el problema de la interpretación de la metáfora y su relación con la creación conceptual. En el caso más radical, la creación conceptual implica un desplazamiento de eventos no proposicionales a eventos proposicionales, es decir, de eventos que solo pueden ser descritos con un vocabulario físico a eventos que también pueden ser descritos con un vocabulario intencional.

6.3. Las metáforas como instrumentos de cambio conceptual

La expresión «X cree que p» puede ser analizada de varias maneras. De un lado, como el estado mental de una persona que se representa el hecho de que p; de otro lado, como una disposición para actuar como si p fuese verdadera. Así, es posible ver un sistema de creencias como un tejido integrado de estados de conciencia o de disposiciones para actuar, es decir, como una visión del mundo y también una forma de vida. La comunidad de creencias integra a los miembros de una forma de vida, pero también los integran los significados que los miembros de esa forma de vida atribuyen a las proferencias verbales o, en general, a las acciones. Una acción —un evento natural que puede ser descrito mediante un lenguaje intencional— es significativo para una comunidad si es posible encontrar en esta prácticas sociales regulares que gobiernen su uso.

De esta manera, uno podría ver el significado de una expresión, a la manera del segundo Wittgenstein, como el uso social regular, pero también podría decirse que el significado de una expresión es el conjunto de creencias, acerca del objeto referido por la expresión, que comparte una comunidad de hablantes. En todo caso, ambas definiciones apuntan a lo mismo, porque si una creencia es una disposición para

actuar, entonces el conjunto de creencias que constituyen el significado de una expresión es, en realidad, un conjunto de disposiciones para actuar respecto del objeto aludido o descrito por la expresión, con lo cual desde el punto de vista de un agente un objeto será algo respecto de lo cual podemos tener una actitud o una disposición conductual. Ya hemos visto todo eso, pero lo estoy recodando para explicar lo que sigue.

Podría decirse que una cosmovisión o una forma de vida es una gran red de sistemas entrecruzados de creencias y prácticas sociales, mayoritariamente compartidos. La pertenencia a la comunidad permite el desarrollo del sistema de creencias del individuo, lo que incluye creencias acerca del mundo, acerca de sí mismo en relación con el mundo y acerca del lenguaje compartido, en varios niveles de intencionalidad. Estas reflexiones nos obligan a preguntarnos por la manera en que cambia un sistema de creencias.

Por «cambio conceptual» entenderé el proceso de alteración de los valores de verdad de las creencias y, por tanto, también de los significados atribuidos a las expresiones, al interior de un sistema de creencias. Entenderé por «creación conceptual» el proceso en el que nuevas creencias y significados se insertan en un sistema de creencias en el que antes no estaban presentes. Así, el caso más radical de creación conceptual acontece cuando alguien tiene creencias que nunca nadie tuvo anteriormente y, entonces, genera significados nunca presentes en comunidad lingüística alguna. En este sentido la creación de creencias y significados implica un desplazamiento de eventos no proposicionales a eventos proposicionales, es decir, de eventos naturales que solo pueden ser descritos utilizando un vocabulario físico, a eventos que también pueden ser descritos utilizando un vocabulario intencional.

Intentaré sugerir que, aunque las oraciones metafóricas no transmiten significados ni creencias, sí pueden producir significados y creencias y, entonces, pueden ser instrumentos de cambio y creación conceptual. En casos de cambio conceptual, la intérprete deberá atribuir creencias que, no habiendo sido transmitidas por las proferencias verbales del hablante, sí han sido causadas por estas[3].

De maneras diferentes, filósofos de distintas procedencias como, entre otros, Ricœur (1981, 1986) y Black (1977, 1979, 1981), han desarrollado doctrinas acerca de la metáfora en la que esta es considerada como instrumento principal para la creación de significado. Se suele contraponer la teoría davidsoniana de la metáfora a la de Ricœur o la de Black, porque estas asumen el punto de vista aristotélico, según el cual

[3] Aunque la teoría de la metáfora de Davidson ha sido empleada para la reflexión sobre la teoría literaria no abordaré ese tema aquí; me voy a concentrar solo en los aspectos semánticos y psicológicos del problema. Para una aplicación del modelo davidsoniano a la literatura véase Dasenbrok, 1993.

las metáforas tienen un segundo significado figurativo, al lado del literal, mientras que aquella niega la existencia de uno metafórico. Quisiera mostrar, sin embargo, que rechazar la posibilidad de un significado metafórico no implica negar que la metáfora pueda ser causa de cambio conceptual y creación de significado.

Ricœur afirma que producimos metáforas con el objetivo de colmar vacíos semánticos. Cuando no tenemos disponible una palabra para expresar cierto estado mental o cierto aspecto del mundo, tenemos que forzar la lengua trayendo una palabra de otro contexto con el fin de hacer posible nuestra descripción. El punto es, según este autor, que en esos casos estamos creando un significado que no existía anteriormente. Ricœur suscribe la posición desarrollada por Aristóteles en la *Poética* (21, 1457b-1459b) y la *Retórica* (III, 2, 1405a y ss.), según la cual se crea una metáfora cuando transferimos un atributo que en condiciones normales pertenece a un sujeto a otro sujeto a quien literalmente no le corresponde, con el fin de mostrar una semejanza entre ambos que en ocasiones normales nos pasaría desapercibida. Esta transferencia de significado suele ir acompañada de una extensión de significado. La aplicación extravagante del atributo hace que se amplíe el significado original con la inclusión de nuevos elementos en su extensión. El producto de esta ampliación de significado es, pues, el significado metafórico. Cuando la palabra transferida es ubicada en el nuevo e inusual contexto, extiende su significado y adquiere un sentido ligeramente distinto. Así, para utilizar un ejemplo aristotélico, después de algunas ocasiones en que llamamos a una copa de vino «el escudo de Dionisio», al interior de cierta comunidad de hablantes, el significado de «escudo de Dionisio» se extiende para incluir «copa de vino». En principio es imaginable que en alguna comunidad de hablantes, por efecto del uso regular, sea perfectamente literal llamar a una copa de vino «el escudo de Dionisio», con lo cual notaremos que la metáfora ha perdido originalidad y se ha congelado. Así, entonces, para una metáfora es la transgresión de un orden convencional con el objetivo de producir cierto significado que las palabras usadas en su sentido ordinario no pueden producir. La tesis de Ricœur es que el proceso mismo de la transgresión de un orden establecido es el de la producción de nuevo significado, es decir, un nuevo orden. Este nuevo significado describe la realidad y nuestros estados mentales de una manera original y, eventualmente, iluminadora.

Max Black también estaría de acuerdo en que la producción de una metáfora acarrea la creación de significado. Sostiene que cuando proferimos una metáfora quebramos ciertas reglas convencionales del lenguaje para expresar lo que aquellas reglas no nos permiten expresar.

A primera vista se podría creer que las posiciones de Aristóteles, Ricœur y Black entran en conflicto con la de Davidson pues, para que la metáfora pueda producir

creación de significado, sería necesario que exista un significado metafórico al lado del significado literal. Una tesis como la de Davidson, que rechaza el significado metafórico, parecería incapaz de ofrecer una adecuada teoría del cambio conceptual. Algo así sostiene Charles Taylor (1995b y 1995c). En lo que sigue intentaré conciliar una teoría del significado que rechaza la noción de significado metafórico con la posibilidad de que la metáfora sea causa de creación conceptual.

Las metáforas no transmiten un segundo significado al lado del literal y, ya que los significados son inseparables de las creencias y como las creencias conllevan contenido cognitivo, las metáforas no transmiten un conjunto de creencias o un contenido cognitivo especial, al lado del literal. Pero las metáforas pueden producir en la intérprete efectos de diferentes tipos, algunos de los cuales pueden ser nuevas creencias y, entonces, nuevos significados y nuevo contenido cognitivo.

Esta idea asume una distinción entre transmitir y causar estados mentales. La distinción es la siguiente: las proferencias de un hablante «transmiten» estados mentales —sean estos cognitivos como creencias o no cognitivos como sentimientos y emociones— cuando causan en la intérprete un contenido similar al que fue expresado por la proferencia en una interpretación estándar. Esto es, el significado literal transmitido por una oración es lo que un miembro típico de una comunidad de hablantes interpretaría en el proceso de producir oraciones-T. Por otra parte, las proferencias de un hablante «producen» estados mentales cuando causan en la intérprete ciertos estados que no están vinculados con la oración proferida ni con una interpretación estándar de ella. Por ejemplo, si Kurt dice con tono angustiado «Está lloviendo» y yo comprendo que él cree que cae agua de las nubes y que probablemente desea que busquemos refugio, podemos decir que, vía su proferencia, él me transmitió sus creencias y deseos. Sin embargo, si Kurt dice «Está lloviendo» y yo recuerdo los versos de Antonio Cisneros: Llueve entre los duraznos y las peras, / las cáscaras brillantes bajo el río / como cascos romanos en sus jabas (1992, p. 191), diría que él no transmitió esos contenidos, sino que los produjo, ya sea intencionalmente o no.

La distinción entre transmitir y producir es comparable a aquella entre razones y causas. Una oración literal puede ser una razón para la justificación de una creencia, mientras que una sensación puede ser una causa —nunca una razón— de una creencia. En este sentido las metáforas son como sensaciones, no justifican sino causan creencias. Naturalmente una oración literal, que es una razón para sostener determinada creencia, puede también ser su causa, pero las metáforas —como las sensaciones— solo pueden ser causas. A eso apunta la tesis davidsoniana del carácter no cognitivo de las metáforas. Pero el punto es que la frontera entre lo que tomamos como una sensación y lo que tomamos como una creencia puede ser flexible, porque depende del estado mental que la intérprete atribuya al agente. En principio, distintas intérpretes

podrían atribuir diferentes estados mentales al mismo agente sin alterar demasiado la interpretación general; esta es una consecuencia de la indeterminación de la interpretación aplicada a estados mentales.

La frontera entre lo proposicional y lo no proposicional también es flexible, y la creación de significado es precisamente el desplazamiento de lo no proposicional a lo proposicional como, por ejemplo, el caso de una metáfora —o una sensación— que causa una creencia que no teníamos anteriormente. Podríamos ponerlo de esta manera: el cambio conceptual es un proceso cognitivo que acontece en la variación de creencias y significados. La creación conceptual, por el contrario, no es un proceso cognitivo. Es, más bien, el proceso en que algo que no es cognitivo —no proposicional— se hace cognitivo —proposicional—. La creación de significado implica que ciertos eventos que solo pueden ser descritos con un vocabulario físico pueden empezar a ser descritos con un vocabulario intencional, como creencias y deseos.

Si cambio los valores de verdad de un subsistema de mis creencias y si este cambio es suficientemente radical como para modificar mi comprensión de los significados de algunas palabras —por ejemplo el significado de la palabra «Sol» antes y después de la revolución copernicana—, podemos decir que ha habido cambio conceptual al interior de un sistema de creencias. Si un investigador científico observa un nuevo fenómeno al punto que este produce en el investigador creencias que no tuvo anteriormente, y que quizá nadie tuvo anteriormente, podemos decir que se ha producido creación de creencias y significado.

Pero hay algunos detalles que me gustaría aclarar. En primer lugar, es obvio que la misma oración puede transmitir ciertas creencias y producir otras. La diferencia entre lo que es transmitido y lo que es producido radica en que lo primero tiene una relación cercana con los significados literales de la oración proferida, mientras que lo segundo está solo arbitrariamente conectado con los significados literales y depende prioritariamente de los estados mentales actuales de la intérprete, es decir, de sus actuales creencias, emociones, deseos, temperamento y experiencias. En el caso de la producción de estados mentales hay una relación más bien floja entre la proferencia y los efectos.

Por otra parte, resultará claro que la distinción entre transmitir y producir no es radical sino sumamente flexible, porque una oración literal puede transmitir un significado literal y producir muchas otras creencias desconectadas en cierto tipo de intérprete o en una intérprete estándar que se encuentre bajo condiciones excepcionales. En condiciones normales, y dado el conocimiento de la lengua, una oración transmitirá un contenido similar en diferentes intérpretes y producirá efectos muy distintos, tal vez incluso no relacionados unos con otros, en el mismo grupo de intérpretes.

En términos generales, lo que nos interesa de una oración literal es el tipo de efectos que transmite, mientras que lo relevante en una metáfora es el tipo de efectos que produce. Esta distinción también es de grado, porque ninguna interpretación calza perfectamente con la interpretación estándar —que, por otra parte, solo es una idealización— y, para que una proferencia produzca ciertos efectos, la intérprete debe tener la capacidad de captar la interpretación estándar. Pero, por supuesto, hay muchos estados fronterizos entre lo que una oración transmite y lo que produce. Ahora quisiera discutir más específicamente el tipo de efectos que una oración usada metafóricamente puede producir.

Los efectos que una metáfora puede producir en una intérprete no son necesariamente diferentes de los efectos que cualquier otra circunstancia del mundo puede producir. La diferencia, que es ciertamente importante, es que la metáfora dirige nuestra atención a los sujetos de la oración sugiriéndonos buscar asociaciones relevantes entre ellos. Veamos ahora la naturaleza de tales efectos. Davidson dice que:

> [La] metáfora y el símil son dos, entre infinitos instrumentos, que sirven para alertar nuestra atención a ciertos aspectos del mundo, invitándonos a hacer comparaciones (1984h [1978], p. 256).
>
> No es solo que no podemos ofrecer un catálogo exhaustivo de lo que ha sido atendido cuando somos conducidos a ver algo bajo una nueva luz; la dificultad es más fundamental. Lo que notamos o vemos no es, en general, de tipo proposicional (p. 263).

Pero el problema que aparece es qué exactamente debe entenderse por «no proposicional» y en qué sentido una comparación entre dos cosas puede ser no proposicional. Intentaré sugerir respuestas a ambas preguntas. El punto es que los efectos que una metáfora puede producir no son proposicionales en el sentido en que las actitudes proposicionales y los demás estados mentales involucrados permanecen ampliamente indeterminados. Cuando escuchamos una metáfora que nos impresiona nos encontramos confusos en una jungla de emociones. Podemos gustar o no de la metáfora y tal vez tengamos algunas creencias producidas por ella, pero todo esto se encuentra en un estado nebuloso en el que difícilmente podríamos distinguir entre lo que creemos y lo que sentimos respecto de nuestro objeto de atención. No podemos separar entre el contenido cognitivo de nuestras creencias y las emociones involucradas. En ocasiones incluso podríamos estar indecisos acerca de si hay alguna creencia o emoción en absoluto. También podemos estar inseguros sobre hasta qué punto se trata de creencias más que de emociones o viceversa.

Después de comenzar el proceso de interpretación podemos o no llegar a determinar, definir y distinguir algunas de las creencias, deseos y emociones que la metáfora

produce en nosotros. Si nos las arreglamos para hacerlo, estos efectos pueden llegar a convertirse en proposicionales, si no, permanecerán no proposicionales y, ciertamente, habrá también situaciones fronterizas. Pero en cualquier caso, la metáfora no llega a transmitir esos efectos, la metáfora solo transmite su significado literal. Los efectos producidos en una intérprete pueden ser diferentes de aquellos producidos en otro y no hay criterios de corrección ni para la producción ni para la interpretación de los efectos producidos en la intérprete, ya que tampoco el hablante está del todo seguro acerca de las creencias y emociones que causaron o acompañaron la metáfora. Algunos de los efectos producidos pueden ser creencias, aún si en algunos casos quedan ampliamente indeterminados.

Cuando comparamos dos cosas o conceptos, diferentes tipos de efectos son generados. Tenemos creencias acerca de la manera en que los objetos son comparables y también hay emociones que la comparación nos produce. Donde hay creencias y emociones también hay significados y deseos, porque las creencias son inseparables de los significados, y las emociones incluyen actitudes de deseo o falta de él. Hay relaciones estrechas entre metáfora y creencia, porque los efectos que una metáfora puede producir en una intérprete varían según las creencias actuales de la intérprete acerca de los sujetos de la metáfora, así como de las creencias del hablante acerca de los sujetos de la metáfora. Así, por ejemplo, cuando Romeo dice «Julieta es el Sol», los efectos en la intérprete dependerán de:

(1) Lo que la intérprete crea que es el Sol. Por ejemplo, la interpretación variará si ella es una heliocentrista o una geocentrista.

(2) Lo que ella crea que el hablante cree acerca del Sol. Por ejemplo, si ella cree que él es heliocentrista o geocentrista.

(3) Lo que ella crea que es el referente de «Julieta». Por ejemplo, si cree que es una mujer, un barco o un caballo.

(4) Lo que ella crea que el hablante cree acerca del referente de «Julieta».

(1) y (3) son creencias de primer grado de intencionalidad, mientras que (2) y (4) lo son de segundo grado. En principio, podríamos imaginar creencias de n-grados, el límite únicamente son las limitaciones del cerebro humano.

Aunque los efectos en la intérprete son condicionados por sus creencias, tales efectos pueden o no ser creencias; y si los efectos son creencias serán probablemente de un tipo muy general. Alguien, por ejemplo, podría llegar a creer que Romeo está sobreimpresionado por Julieta o que Julieta es una persona que tiene que llegar a conocer algún día.

Las metáforas suelen producir también emociones. Tales emociones están condicionadas por los cuatro tipos de creencias mencionadas y suelen ir acompañadas de deseos. Además, así como las creencias influyen en las emociones, estas influyen en las creencias. El punto es que cuando escuchamos una metáfora que nos impresiona, podemos reconocernos incapaces para distinguir entre las creencias/significados y las emociones/deseos producidos por esta. Por ejemplo, podríamos no estar seguros de si la actitud que tenemos respecto de Julieta es cognitiva o emocional. Esta situación sería llamada no proposicional.

Los efectos producidos por la metáfora son constituidos por relaciones originales e inusuales entre los conceptos. Así, la originalidad de la metáfora puede producir un estado mental no proposicional que es nuestra incapacidad para determinar los diferentes efectos o actitudes que la metáfora produce en nosotros. En este respecto, el hablante no está en una mejor situación que la intérprete, así que consideremos ahora al hablante y el tipo de estados mentales que él quiere expresar, estados mentales que son la causa de la proferencia de la metáfora.

Un punto importante en la teoría davidsoniana de la interpretación es que los significados, creencias y deseos están interrelacionados. Ya que Davidson sostiene que el significado de una metáfora no es diferente de su significado literal, podemos preguntar si los significados y deseos expresados por la metáfora son los mismos transmitidos por su significado literal. Podríamos formular la pregunta de esta manera: ¿Qué cree Romeo acerca de Julieta, y qué quiere expresar, cuando dice que ella es el Sol? ¿Cree él y quiere expresar que ella es una esfera incandescente? Pienso que hay cuatro posibles respuestas. Voy a rechazar las primeras tres, luego sostendré la cuarta.

(1) Él cree que ella es esférica e incandescente.

(2) Él no tiene creencias conectadas con su proferencia. Según esto, Romeo tiene muchas creencias acerca de Julieta, pero cuando él profiere la metáfora ninguna de ellas es expresada. Él produce solamente una exclamación emocional que no es diferente de un suspiro.

(3) Romeo cree que Julieta es como el Sol porque ella es el centro de su mundo y le da calor, vida, luz, etcétera, y su proferencia expresa esas creencias.

La hipótesis (1) es evidentemente absurda porque sería inconsistente con las otras creencias que Romeo tiene acerca de Julieta y que le permiten interactuar con ella. Las hipótesis (2) y (3) comparten un mismo problema en tanto rechazan el principio de la inseparabilidad entre significado y creencia. La relación entre el significado literal de las palabras y las creencias expresadas por ellas se convierte en algo innecesariamente misterioso.

(4) Veamos una cuarta posibilidad regresando a la noción de indeterminación. La única diferencia entre las proferencias de Romeo «Julieta es rubia» y «Julieta es el Sol» radica en que, en el primer caso, los significados, creencias y deseos involucrados en la oración pueden ser encontrados con poca indeterminación y al emplear básicamente una teoría previa. En el segundo caso, el proceso de encontrar los significados, creencias y deseos requiere de un trabajo más exigente que, sin embargo, exhibirá una mayor indeterminación y requerirá lo mejor de nuestra habilidad para producir una adecuada teoría al paso. Los problemas comenzarán cuando la intérprete quiera relacionar la oración con el comportamiento general del hablante. Aquí emergerá la indeterminación.

Cuando Romeo dice «Julieta es rubia» entendemos que él cree que ella es rubia y que desea transmitir su creencia a nosotros. Cuando dice «Julieta es el Sol», podemos imaginar el tipo de creencias que tiene —ella es el centro de... etcétera— y podemos también imaginar algunos de sus deseos, pero habrá poca uniformidad entre las distintas intérpretes posibles. Las diferencias entre sus interpretaciones dependerán de sus teorías previas y su habilidad para emplear buenas y apropiadas teorías al paso. En otras palabras, cuando Romeo dice «Julieta es el Sol», él tiene ciertas creencias relacionadas con su proferencia, el problema es que para una intérprete promedio no es fácil encontrar la relación. En este caso Romeo solo está ligeramente menos confundido que la oyente. Él también es un intérprete de sus propias palabras y le es más fácil interpretar una oración literal antes que una metáfora, incluso si es él quien las profiere. Él está en mejor situación que la intérprete en tanto posee la autoridad de la primera persona acerca de los significados que emplea y acerca de sus creencias previas sobre el sujeto de la metáfora. Su problema es relacionar esos significados con esas creencias. Sigue siendo correcto que la metáfora es una invitación para hacer comparaciones imaginativas e iluminadoras, es decir, para producir nuevas creencias acerca de diferentes cosas, pero aunque las comparaciones expresadas en esas creencias no pueden ser explicadas en términos de regularidades, no son enteramente arbitrarias y son producidas por los significados literales de las oraciones, es decir, por regularidades. Así, el proceso de comprender una metáfora requiere de una interpretación amplia, esto es, de una interpretación que incluya toda la información disponible acerca del hablante y su entorno, más que solo lo que creemos son los significados de sus palabras. Esto no es diferente de lo que ocurre en la interpretación de cualquier oración literal ordinaria, pero en el caso de la metáfora se ve con más radicalidad. Intentaré expresar esto con más claridad al comparar las metáforas con los símiles.

En tanto la metáfora es principalmente un asunto de pragmática antes que de semántica, las metáforas son semejantes a los símiles. Los efectos pragmáticos de la oración «Julieta es el Sol» son iguales a los de la oración «Julieta es como el Sol». El símil no produce un contenido determinado porque Julieta es como el Sol en innumerables sentidos y el símil no nos dice en qué sentido ella es como el Sol. Pero, ¿en qué sentido lo es? Depende de los correspondientes conjuntos de creencias de primer grado acerca de Julieta y el Sol, tanto de la intérprete como del hablante. Lo mismo ocurre en relación con las creencias de la intérprete y el hablante acerca de Julieta siendo como el Sol. Esto sugiere que no hay un hecho por el que dichas creencias sean las verdaderas creencias expresadas por la metáfora; solo hay un espectro de diferentes posibilidades, tanto en la mente de la intérprete como del hablante.

Este es el momento de ver en qué sentido puede una metáfora producir, tanto en el hablante como en la intérprete, cambio conceptual y creación de significado. Hemos visto que una oración transmite significados literales, siendo aquellos lo que un miembro promedio de la comunidad de hablantes interpretaría al producir oraciones-T. También hemos visto que una oración produce efectos cuando causa en la oyente estados mentales que no están necesariamente relacionados con el significado literal de la oración. Sugeriré ahora ver la relación entre significados transmitidos y efectos producidos de la manera siguiente: cuando hay regularidades estrictas entre los efectos que una oración puede producir en los miembros promedio de una comunidad de hablantes, estamos en condiciones de llamar a tales efectos regulares «significados». Así, podemos decir que la oración transmitió esos significados como los efectos producidos por la mayor parte de oraciones literales en circunstancias usuales. En otras palabras, los significados son solo efectos ordinarios que, por su frecuencia y regularidad, son suficientemente familiares como para ser asociados a ciertas acciones y estados mentales, y no a otros. El cambio de significado ocurre cuando ciertos efectos poco frecuentes producidos por una palabra en una oración se hacen más familiares, con lo cual decimos que la palabra ha adquirido un nuevo significado. La creación de significado ocurre cuando algunos efectos se hacen suficientemente regulares como para que merezcan un nombre y un lugar especial en nuestro sistema de creencias y nuestro lenguaje.

En tanto los efectos producidos por metáforas vivas son poco regulares y no es fácil determinar las creencias involucradas con ellos, no podemos llamar a tales efectos «significados». Cuando cierta acción —la proferencia de una oración, un gesto o cierto movimiento de las manos— es usada de manera suficientemente regular entre un número considerable de personas como para constituir en esa comunidad las nociones de «corrección», «regla» y «práctica social», es que tenemos significados. El significado está atado a usos familiares del lenguaje. Pero ciertamente

no es posible trazar una frontera clara entre este tipo especial de efectos predecibles, al que llamamos «significado», y todos los otros posibles.

Los significados literales están fijados por condiciones de verdad que toman la forma de oraciones-T. Las oraciones-T son acuerdos, basados en creencias y deseos, acerca de cómo usar las palabras. Decir que las metáforas no tienen significado adicional al significado literal equivale a decir que las metáforas no tienen un lugar establecido en nuestro sistema de creencias y que, entonces, las metáforas están más allá de nuestros actuales acuerdos. Así, al escuchar a Romeo decir que Julieta es el Sol, la intérprete es invitada a realizar toda suerte de comparaciones entre aquellos dos objetos. En un primer momento, los efectos pueden ser nebulosos y no proposicionales. Más adelante, tales efectos pueden convertirse en nuevas e inesperadas creencias acerca de Julieta y el Sol. Lo que en un primer momento pueden ser solo efectos no proposicionales para una intérprete creativa pueden convertirse en nuevas creencias y significados. Algunas veces se pueden constituir nuevos acuerdos y, entonces, las metáforas pueden perecer para convertirse en oraciones literales.

He intentado mostrar que la muerte de una metáfora puede ser vista como una imagen de lo que normalmente llamamos cambio conceptual: lo que en un primer momento suena obviamente falso o absurdo es, en un segundo momento, considerado iluminador y sugerente. Finalmente, si la metáfora es exitosa y muere, se convertirá en un pequeño sistema de creencias considerado por algunos como verdadero. Esta puede incluso crear significado si origina en nosotros creencias acerca de algo para lo cual antes no teníamos una opinión.

Este proceso no solo amplía nuestra visión del mundo y de los otros sino también de nosotros mismos. Por supuesto la comprensión de una metáfora es solo un ejemplo de una actitud humana mucho más general, que también se ve caracterizada en las relaciones intersubjetivas o interculturales. Es el esfuerzo por entender el comportamiento diferente, las creencias distintas y los deseos extraños lo que termina por constituirnos en nuestra relación con los demás. Por ello, la comprensión no requiere de acuerdos previos sino solamente de nuestra capacidad para constituirlos. El caso paradigmático de esto es lo que ocurre cuando nos confrontamos por primera vez con una metáfora penetrante. Al comienzo es solo una oración falsa, incomprensible o absurda; pero es una oración que nos cuestiona lo suficiente como para que, después de unos momentos, encontremos que allí hay algo iluminador y profundo.

En la segunda parte de este libro he discutido la naturaleza del significado porque muchas formas de comprensión, aunque no todas, requieren que las intérpretes atribuyan significados a las proferencias verbales de los hablantes. En la tercera parte abordaré directamente la naturaleza de la comprensión.

Tercera parte
La naturaleza de la comprensión

Capítulo siete
La comprensión como la creación de un espacio compartido

7.1. El principio de caridad

El principio de caridad constituye uno de los desarrollos del pensamiento de Davidson que más interés ha concitado en la filosofía reciente. No solo es relevante para temas específicos en filosofía del lenguaje, epistemología y filosofía de la mente, sino que también tiene importantes consecuencias para la reflexión teórica en disciplinas como la psicología y las ciencias sociales. Sin embargo, para percibir esos alcances es necesario ver a este principio como un conjunto de tesis que va mucho más allá de la simple metodología de la interpretación.

La comprensión puede ser unidireccional, bidireccional o multidireccional. En el primer caso, la intérprete trata de comprender a un agente que no interactúa con ella o que no sabe que está siendo interpretado. En el segundo caso, ambos interactúan e intentan comprenderse mutuamente. Si este tipo de comprensión es exitosa, podremos hablar de comunicación. En el tercer caso, un grupo de individuos trata de comprenderse unos a otros. En principio me referiré a las formas más exigentes y completas de comprensión bidireccional, entendiendo que la unidireccionalidad es un aspecto de aquella y que la multidireccionalidad es una complejización de varias formas de comprensión bidireccional superpuestas, pues la comprensión de grupos tiene características propias que van más allá de la comprensión de parejas individuales. Sin embargo, analizar lo que ocurre en las formas más complejas de comprensión bidireccional nos facilitará entender lo que ocurre en las otras formas de comprensión.

En este capítulo me propongo sugerir una formulación del principio de caridad a partir de la cual es posible extraer consecuencias acerca de la naturaleza de la comprensión. Deseo mostrar que esta formulación hace explícito el abandono de la hermenéutica intencionalista y de la concepción internista de la mente, aquellas

según las cuales comprender al otro es conocer sus estados mentales internos, y más bien permite entender la comprensión como la creación de un espacio compartido entre intérprete y agente en torno a una realidad común. Así, entender al otro no sería solo reconstruir su vida mental ni conocer los estados mentales que causaron sus acciones, sino generar una forma de vida común, una intersección de horizontes que conlleva a una invitación al cambio, la tolerancia y la adaptación. Las alusiones a Wittgenstein y a Gadamer muestran que hay muchas conexiones que es posible hacer entre estos tres autores. Mostraré algunas de ellas en los próximos capítulos.

Hay muchas formulaciones diferentes del principio de caridad, pero la formulación clásica de Davidson aparece en el siguiente párrafo, que por su importancia citaré en inglés y luego la traduciré al castellano:

> *In our need to make him make sense, we will try for a theory that finds him consistent, a believer of truths and a lover of the good (all by our own lights, it goes without saying). Life being what it is, there will be no simple theory that fully meets this demands. Many theories will effect a more or less acceptable compromise, and between these theories there will be no objective ground for choice* (1980b, p. 222)[1].
>
> En nuestra necesidad de hacerlo inteligible, elaboraremos una teoría que lo encuentre consistente, un creyente de verdades y un amante del bien (todo según nuestros propios criterios, es innecesario decirlo). Siendo la vida como es, no habrá una teoría simple que cumpla con estas exigencias. Muchas teorías lograrán un arreglo más o menos aceptable, y entre esas teorías no habrá bases objetivas para la elección.

El problema con las formulaciones davidsonianas del principio de caridad es que tienden a ser imprecisas y no explicitan suficientemente la complejidad de sus detalles y presupuestos. Esto ha originado cuestionamientos de fondo. En algunos lugares Davidson dice que el objetivo del principio de caridad es optimizar el acuerdo (1984c, p. 197) o minimizar el desacuerdo (p. XVII). Más adelante aclara esas formulaciones y sostiene que:

> El objetivo de la interpretación no es el acuerdo sino la comprensión. Mi tesis siempre ha sido que solo se puede asegurar la comprensión si se interpreta de una manera que permita lograr un acuerdo apropiado. La noción de «apropiado»,

[1] Se puede encontrar otras formulaciones de este principio en las páginas 221 y 290 del mismo libro, así como en Davidson, 1984c, pp. 137, 153, 168, 169, 196, 197. Todas ellas tienden a ser una elaboración del *dictum* quineano: «La necedad de un interlocutor, más allá de cierto punto, es menos probable que una mala traducción» (Quine, 1960, p. 59). Como ya señalé, el principio de caridad fue propuesto, por primera vez, por Wilson, 1959, p. 432; 1970, p. 300, como parte de una teoría de la referencia. Posteriormente fue formulado de una manera más amplia por Quine, 1960, pp. 57-58, en su análisis de la traducción radical.

> sin embargo, no es más fácil de especificar que decir qué constituye una buena razón para sostener una creencia en particular (p. XVII)[2].

En este capítulo intentaré proponer una formulación más precisa del principio de caridad que sea, al mismo tiempo, más hábil para defenderse de los diversos cuestionamientos sugeridos. Para ello haré uso de la vieja noción de empatía, que en la filosofía reciente ha sido reformulada como capacidad de simulación. También trabajaré sobre la noción de triangulación, que Davidson y Cavell utilizan (Davidson, 2001b; Cavell, 1993, 1998, pp. 449-467).

La concepción de empatía, desarrollada por Schleiermacher y Dilthey, se puede rastrear hasta el concepto de *sympathy* en Hume (1975), quien pensaba que en el origen de la moral se encuentra la empatía en tanto nos permite alejarnos de nuestra perspectiva e interés para comprender la perspectiva e interés de otra persona. Este autor sostenía que la empatía está presente en la comprensión de los demás tanto como en los juicios estéticos y éticos, y que tenemos una tendencia natural a «proyectar» nuestros sentimientos —morales y estéticos— sobre las personas y las cosas.

La empatía alude a la capacidad de ponerse en el lugar del otro para padecer con él sus venturas o desdichas. En tanto teoría de la comprensión, este concepto ha sido cuestionado por estar comprometido con una hermenéutica intencionalista y con un modelo transposicional de la comunicación,[3] según el cual comprender al otro es ser capaz de conocer sus estados mentales, o su escenario interior para ponerlo en términos de Ryle (1967), mediante la «transposición» de esos estados desde su mente a la nuestra. Ryle denominó «el mito del fantasma en la máquina» a la idea de que la mente es una suerte de escenario interior al que solo uno tiene acceso, de manera que la historia de una persona es en realidad dos historias paralelas, la de su cuerpo, que es pública, y la de su mente, que es privada. Para este modelo comprender a otra persona sería lograr reproducir en nuestro escenario interior lo que creemos que ocurre en el suyo.

Aunque Locke es el defensor paradigmático del modelo transposicional (1982, libro III, capítulo II, sección I), también podemos encontrarlo en la teoría de la empatía de Dilthey. Parece, pues, que las formulaciones clásicas de la empatía están comprometidas con una concepción internista y cartesiana de la mente, en la que comprender al otro es reproducir en nuestro escenario interior —por transposición— lo que creemos que está ocurriendo en su propio escenario interior.

---

[2] «*The aim of interpretation is not agreement but understanding. My point has always been that understanding can be secured only by interpreting in a way that makes for the right sort of agreement. The «right sort», however, is no easier to specify than to say what constitutes a good reason for holding a particular belief*».

[3] Algunos, como Bennett, 1971, lo denominan el modelo de la «traducción», pero yo prefiero denominarlo «transposicional» para no confundirlo con los modelos de traducción radical de Quine y Davidson. Brandom, 1994, también lo ha llamado «transportacional».

Quizá por esos motivos, en la filosofía de la mente reciente y en la psicología experimental se ha acuñado el concepto de «simulación», que se entiende como la capacidad para imaginar ser el otro en condiciones contrafácticas. En el segundo capítulo he presentado la teoría de la simulación, que se propone explicar la manera como el niño desarrolla una «teoría de la mente» y así puede imaginar ser el otro mediante el uso de la capacidad de la metarrepresentación. Esto ha sido extensamente estudiado en la bibliografía actual (Davies & Stone, 1995a, 1995b; Whiten, 1991; Astington, Harris & Olson, 1988).

Otra opción es mantener el concepto de empatía en un sentido no intencionalista. No voy a detenerme ahora en las diferencias que pudiera haber entre empatía y simulación, tema sobre el que volveré más adelante, solo subrayaré que me propongo emplear nociones de empatía y simulación no intencionalistas ni internistas. Más adelante entraré en el debate sobre si uno simula «ser» el otro bajo condiciones contrafácticas o simula «tener» los estados mentales del otro bajo condiciones contrafácticas, prefiriendo la primera opción con algunas modificaciones.

Deseo sostener que el objetivo del principio de caridad es minimizar la divergencia inexplicable al mostrar que el criterio para la explicación de la divergencia debería ser la noción de simulación. La idea será que atribuimos al agente los estados mentales que nosotros creemos que tendríamos si estuviéramos en las circunstancias en que creemos que él está, a menos que tengamos razones para creer que él no puede tener esos estados mentales. Así debemos tener un criterio para determinar cuáles son sus estados mentales idiosincrásicos y por qué él los tiene mientras que nosotros no; o por qué nosotros tenemos estados mentales que él no tiene o no podría tener. Como he señalado, nuestro criterio será la simulación. Imaginamos cómo es ser él en una situación en particular y le atribuimos los estados mentales que creemos que nosotros tendríamos si fuéramos él en esa circunstancia. En lo que sigue expondré las nociones de triangulación y simulación, y mostraré cómo se pueden integrar con una reformulación del principio de caridad.

Como vimos en el segundo capítulo, la triangulación es un fenómeno que ocurre cuando dos interlocutores reaccionan al mismo tiempo al mundo que comparten, reaccionando a su vez cada uno a las reacciones del otro. La triangulación madura es un desarrollo ontogenético de la atención compartida que se desarrolla en los infantes. En el caso de la interpretación radical, la intérprete es consciente de que ciertas circunstancias del entorno causan al agente a producir cierta proferencia verbal p. Sin embargo, la intérprete no sabe cuál es el significado de p para el agente ni cuáles creencias en él han sido causadas por esas circunstancias. Lo que la intérprete sabe es cuáles creencias han sido causadas en ella en las mismas condiciones. De esta manera, ella atribuirá al agente las creencias que esas condiciones han causado en ella.

En otras palabras, ella atribuirá al agente algunas de sus creencias y asignará, como el significado de p en el lenguaje del agente, una oración de su propio lenguaje con las mismas condiciones de verdad. Como hemos visto anteriormente, esto se formula de manera técnica como la producción de oraciones-T.

Davidson rechaza la idea de que la intérprete «proyecte» sus estados mentales en el otro, por considerar que esto tiene una connotación antirrealista (1997). Discrepo de él en lo relativo a las connotaciones antirrealistas. La intérprete puede proyectar sus propios estados mentales precisamente con la finalidad metodológica de encontrar los del otro, los cuales tienen realidad independiente de la intérprete aunque son identificados en circunstancias intersubjetivas, pues la evidencia siempre subdeterminará los distintos manuales de interpretación que se puede hacer de ellos. En otras palabras, los estados mentales del otro son reales pero abiertos a un ámbito de indeterminación. Sin embargo, es verdad que el uso del término «proyección» puede sugerir una concepción demasiado egocéntrica de la interpretación. Volveré sobre este punto más adelante.

En general, la intérprete asumirá que el agente reaccionará frente a los estímulos externos básicamente de la misma manera como ella reaccionaría. Considérese a ambos interlocutores reaccionando frente a una manzana, por ejemplo. La manzana causará percepciones en ambos interlocutores y cada uno de ellos observará las reacciones del otro. Así, cada uno correlacionará sus propias percepciones y creencias, en relación con la manzana, con los del otro. Lo importante es que cada interlocutor reaccionará frente al mundo y también frente a las reacciones que el otro tiene respecto del mundo. Cada interlocutor deberá ver las reacciones ajenas como una perspectiva diferente del mismo mundo. De esta manera, cada uno asociará el concepto de «manzana» con sus propias reacciones y con las reacciones del otro. Al hacerlo, cada uno aprenderá el concepto «manzana» de manera intersubjetiva, asumiendo que hay un uso objetivo que este concepto tiene, y atribuirá al otro creencias y deseos relacionados con el objeto manzana y con el concepto «manzana». Más aún, cada uno de ellos aprenderá a asociar las reacciones del otro con sus propias reacciones y así también aprenderá a simular ser el otro bajo circunstancias similares; es decir, aprenderá a empatizar con el otro en relación a un mundo compartido.

7.2. El inicio de la interpretación

Ahora bien, la intérprete podrá comenzar a interpretar al agente solo si ella tiene dos conjuntos de supuestos iniciales acerca de él. De un lado, un conjunto de creencias de primer grado en torno del agente y del mundo que comparten; de otro lado, un conjunto de creencias de segundo y más grados acerca de las creencias del agente,

así como acerca de sus deseos y propósitos al realizar sus acciones. La intérprete formará estas hipótesis interpretativas al considerar el comportamiento del agente tanto como su entorno y se formará creencias acerca de los estados mentales del agente al atribuirle parte de su propio sistema de estados mentales. Pero, en este punto, la intérprete necesitará un criterio para determinar cuáles estados mentales deberán ser atribuidos a él. Sostengo que ese criterio es la simulación. Al simular ser él, ella intentará ver el mundo desde la perspectiva que ella cree que él tiene. Así, si la interpretación comienza cuando ella le atribuye a él un aspecto de su propio sistema de estados mentales, ella tendrá que asumir que, en líneas generales, él cree lo que es verdadero y desea lo que es deseable, todo según los criterios de ella misma. Este, por supuesto, es el principio de caridad.

Habrá que recordar que el principio de caridad no está haciendo una sugerencia metodológica para la interpretación; este principio sostiene que es condición de posibilidad de la interpretación que la intérprete asuma que en líneas generales:

(1) El comportamiento del agente es consistente.

(2) El agente mantiene los mismos supuestos lógicos que la intérprete.

(3) Las creencias del agente son básicamente verdaderas.

(4) El agente se comporta siguiendo su mejor juicio. Esto es, el agente persigue los mismos fines que la intérprete al realizar sus acciones bajo las mismas condiciones generales.

La justificación principal del principio de caridad toma la forma de un argumento trascendental, en tanto muestra que ciertas prácticas de tipo A —la atribución de estados mentales a agentes intencionales— presupone necesariamente ciertas otras prácticas de tipo B —la aplicación de las condiciones del principio de caridad—. Ya que no podemos evitar estar comprometidos con A, entonces también debemos estar comprometidos con B[4].

Al justificarse en un argumento trascendental, el principio de caridad se muestra como una condición necesaria para la interpretación y para la atribución de estados mentales, no es una afirmación empírica acerca del comportamiento de los intérpretes. Si fuese una afirmación empírica, podría falsarse mediante contraejemplos empíricos. Por ello, ningún descubrimiento antropológico podría mostrar que es falso, solo un análisis más fino de los conceptos de interpretación. El punto es que, de no asumir el principio de caridad, no podríamos reconocer comportamiento

[4] Hay una copiosa bibliografía acerca de la validez de los argumentos trascendentales. No entraré en el debate ahora, aunque en líneas generales asumiré su validez si el argumento está bien formulado. Para el debate, véase Hintikka, 1972; Bieri, Kruger y Horstmann, 1979; y Malpas, 1990. Para una aplicación de este debate al caso de Davidson véase Carpenter, 2002.

intencional en las personas, por tanto, no podríamos reconocerlos como agentes y no podríamos atribuirles ningún tipo de creencias ni deseos.

Como dice Lewis:

> Una persona podría no tener creencias, deseos o significados en absoluto, pero es analítico que, si los tiene, entonces están constituidos en líneas generales de acuerdo con los principios que definen la creencia, el deseo y el significado (1983, p. 12).

Analizaré ahora cada una de las condiciones del principio de caridad que han sido señaladas líneas arriba. La condición (1) sostiene que debemos asumir consistencia general en el comportamiento del agente si esperamos reconocerlo como comportamiento intencional. La consistencia debe ser entendida tanto en un sentido lógico como en un sentido pragmático. En un sentido lógico, la atribución de demasiadas contradicciones cuestiona la posibilidad de adscribir un sistema de estados mentales interconectados, así como la posibilidad misma de dar contenido a cada uno de ellos. La intérprete solo podrá determinar el contenido de cada estado mental al interconectarlo con los otros estados mentales del sistema atribuido. Así es también la manera como aprendemos nuevas creencias, no lo hacemos una por una sino mediante pequeños subsistemas. Considérese la siguiente afirmación de Wittgenstein:

> Cuando comenzamos a creer algo, no creemos una proposición simple sino un sistema total de proposiciones. (La luz ilumina gradualmente la totalidad.) (1972, aforismo 141).

Pero en un sentido pragmático, el comportamiento de un agente es consistente si sus estados mentales y acciones no son incompatibles según la perspectiva de la intérprete. Serían incompatibles si se excluyeran mutuamente.

La condición (2) sostiene que, ya que la intérprete atribuye al agente parte de su propio sistema de estados mentales, no podrá evitar reconocer el comportamiento del agente si no es desde su propia estructura lógica. La condición (3) establece que la intérprete debe asumir que las creencias del agente son básicamente verdaderas, es decir, que son semejantes a las de ella. Esto se sigue de la necesidad de atribuir, mediante el proceso de la triangulación, un sistema integrado de creencias, en el que las interacciones entre estas fijan el contenido de cada una de ellas. En este sistema, la falsedad debe ser excepcional[5]. Un creyente sistemático de falsedades no podría ser triangulado con el entorno. Tampoco podría determinarse cuáles son sus creencias falsas, porque la intérprete solo precisará la falsedad de una creencia del agente

[5] «Mientras más creencias correctas tenga el creyente, más notorios serán sus errores. Demasiadas equivocaciones simplemente distorsionan el enfoque» (Davidson, 1984c, p. 168).

contra el telón de fondo de creencias verdaderas que le permiten reconocer su significado y ella solo podrá reconocer creencias verdaderas si estas creencias también son creencias de ella.

Así pues, una intérprete no podría creer que su interlocutor sea un creyente sistemático de falsedades, a costa de que se evaporen los conceptos mismos de «creencia» y «falsedad». Pero no solo eso, mediante el uso de un argumento trascendental semejante al que justifica el principio de caridad, podría afirmarse de manera más general que sería imposible que alguien fuese un creyente sistemático de falsedades. El argumento tomaría más o menos esta forma: Puesto que todo sistema de creencias se construye en una situación interpretativa en que la intérprete lo atribuye al agente para hacerlo inteligible, y como sería imposible que alguna intérprete construya para su interlocutor un sistema constituido principalmente por creencias falsas, entonces sería imposible que existiese un sistema constituido principalmente por creencias falsas. No podría ocurrir que alguien tenga un sistema de creencias principalmente falsas independientemente de toda interpretación, porque al afirmar esa posibilidad se estaría negando la tesis central de que los contenidos de los estados mentales se atribuyen de manera intersubjetiva y no son propiedades monádicas de mentes individuales. De esta manera, el principio de caridad tiene como consecuencia afirmar una propiedad de los sistemas de creencias como tales: pertenece al concepto mismo de creencia que la mayor parte de creencias de un sistema deben ser verdaderas, es decir, que cualquier intérprete que construya un sistema de creencias para comprender a alguien tendrá que asumir que la mayor parte de creencias son verdaderas.

Pero ahora el problema es cómo llegar a formular el principio de caridad de una manera rigurosa y precisa. Podríamos hablar de maximización de acuerdo y racionalidad, y minimización de falsedad e irracionalidad, que es la formulación habitual que hace Davidson, pero en este caso habría que aclarar qué clase de acuerdo debe asumir la intérprete, esto es, acuerdo acerca de qué tipo de creencias. Además, habría que establecer el criterio que ella tendría que emplear para explicar las divergencias en creencias entre ella y el agente. Davidson no ofrece respuestas a estas preguntas y tampoco tiene un criterio para la explicación de la divergencia. Por ello, me parece preferible decir que el objetivo del principio de caridad es minimizar las divergencias inexplicables de estados mentales entre la intérprete y el agente usando como criterio la noción de simulación.

Hemos visto que la intérprete comenzará la interpretación atribuyendo al agente sus propios estados mentales, pero ella necesitará un criterio para determinar cuáles de sus estados mentales deberá atribuirle. Como hemos visto, ella le atribuirá aquellos estados mentales que ella cree que tendría si fuera el agente en las circunstancias en que él se encuentra. Al hacer eso, intentará ver el mundo desde el punto de vista

del otro. De esta manera, no solo le atribuirá muchas de sus propias creencias sino también creencias que ella no tiene o que considera que son falsas. Las creencias sobre las cuales la intérprete debe asumir acuerdo tendrán que ser aquellas que la intérprete considere que debe atribuir al agente para poder hacerlo inteligible. Y, por supuesto, no hay manera de saber por adelantado cuáles serán esas creencias.

Al asumir que las creencias del agente son básicamente verdaderas, la intérprete supone también que él acierta *grosso modo* en la manera en que el mundo es y, por tanto, que comparten un mundo que es previo e independiente de ellos. Aquí también tenemos un argumento trascendental que prueba que sería imposible no creer en la objetividad e independencia de la realidad. El argumento no prueba que exista un mundo previo e independiente de nosotros, lo que prueba es que es condición necesaria para la interpretación que asumamos eso. Ahora bien, ya que sería imposible creer lo contrario, hemos probado que es necesario que creamos que el mundo que compartimos es previo e independiente de nosotros. Esto es todo lo que se necesita probar en relación a la existencia de la realidad. Así satisfacemos nuestras intuiciones realistas, que no podríamos dejar de tener, y también dejamos satisfecho a quien considera que conocemos el mundo desde cierta interpretación que ya tenemos de él, de suerte que es imposible siquiera concebir el mundo si no es interpretado de alguna forma. Esto resulta importante porque justifica una tesis realista, mediante un argumento trascendental, en la naturaleza misma de la interpretación y no en una tesis metafísica. Esta posición podría ser denominada, como lo hace Dreyfus (1996) en otro contexto, «realismo hermenéutico». Al triangular entre sus propios estados mentales, los del otro y los objetos del mundo que comparten, los interlocutores asumirán que su acuerdo procede de la objetividad misma del mundo que habitan. Pero también podría ser llamado «realismo directo», porque no asume ningún intermediario epistémico entre nosotros y la realidad, sino un contacto directo.

Grandy (1973) ha ofrecido una versión diferente del principio de caridad, a la que ha denominado «principio de humanidad». En ese texto el autor comenta la versión de Quine (1960), porque la de Davidson todavía no había aparecido. Según él, este principio no tiene como objetivo maximizar el acuerdo sino minimizar el desacuerdo ininteligible o, lo que es lo mismo, maximizar la inteligibilidad, aunque no necesariamente el acuerdo. Hay quienes creen que el principio de humanidad es un avance respecto del de caridad, al implicar que puede haber inteligibilidad sin acuerdo (Papineau, 1978). Macdonald también cree que deberíamos sustituir humanidad por caridad:

> Más que tomar la caridad al extremo de minimizar todo desacuerdo, parecería una estrategia interpretativa más sensata el estar concernido únicamente con minimizar el desacuerdo ininteligible (Macdonald & Pettit, 1981, p. 37).

A primera vista estas objeciones son sorprendentes, porque parecen implicar que puede haber inteligibilidad sin un fondo de acuerdo presupuesto o asumido. Precisamente el punto del principio de caridad es intentar probar que solo puede haber inteligibilidad y desacuerdo contra un fondo de acuerdo compartido, incluso si es tan pequeño y transitorio como el encuentro fugaz de teorías al paso. Pero lo que Grandy en realidad está sugiriendo es que, aunque debe haber un fondo de acuerdo, algunas veces podemos hacer más inteligible al agente si le atribuimos un subsistema de creencias falsas, siempre que podamos explicar cómo es que él las adquirió. Así, él objetaría una formulación del principio de caridad que sostuviera, como lo hace Quine, que la mejor traducción es aquella que maximiza el acuerdo. El problema con la formulación de Quine es que no explica cómo en ocasiones la atribución de falsas creencias puede ayudar a hacer más inteligible al agente. La posición de Quine intentará minimizar todo lo posible la atribución de creencias falsas, mientras que Grandy sugeriría minimizarlas solo cuando no podamos explicar cómo fueron adquiridas estas creencias.

> Al defender la maximización del acuerdo en la interpretación [el principio de caridad] aconseja el abandono de consideraciones sobre si es posible que los interpretados hayan obtenido conocimiento de las verdades de las que ellos mismos son interpretados como habiendo alcanzado acuerdo (1973, p. 29).

Coincido con Grandy en este punto. El principio de caridad debería apuntar a minimizar el error y la irracionalidad solo cuando sean inexplicables. La pregunta es cuál será nuestro criterio para la atribución de creencias específicas al agente. Este autor sugiere tomarnos a nosotros mismos como modelos, pero no desarrolla una noción de simulación y, aunque se acerca a ella, lo hace al revés.

> Así pues, tenemos como una restricción pragmática de la traducción la condición de que el patrón de atribuciones sobre las relaciones entre creencias, deseos y el mundo, deba ser tan similar al nuestro como sea posible. […] El que nuestra simulación de la otra persona sea exitosa, dependerá fuertemente de la similitud entre su red de creencias y deseos y la nuestra (p. 43).

Ahora bien, hacer al agente lo más semejante a uno podrá ser proyección, pero no puede ser simulación, precisamente porque la idea de la simulación es el intentar alejarnos de nuestros propios estados mentales para imaginar ser el agente, dejando espacio en nosotros mismos para albergar al otro en su particularidad. Por eso, la versión que necesitamos del principio de caridad es una que permita a la intérprete suficiente flexibilidad para poder desplazarse más allá de sus propios estados mentales, imaginando los ajenos. Ella no solo deberá proyectar una parte de sus

propios estados mentales al agente, sino también deberá imaginar otros que ella no tiene y que quizá jamás podría tener, incluso inusuales o exóticos. Pero para poder hacer eso, ella deberá ser capaz de atribuirle a él aquellas creencias que ella piensa que estaría en condiciones de adquirir si fuera él. Posteriormente, y según el comportamiento del agente, ella tendrá que hacer los ajustes necesarios para atribuir nuevos estados mentales. Intentaré aclarar este punto al contestar una objeción de Devitt.

Devitt nos invita a imaginar a un intérprete no religioso tratando de comprender a un nativo involucrado en algunos ritos religiosos. La aplicación del principio de caridad, piensa Devitt, no permitiría hacer inteligible al nativo. Dice:

> ¿Qué haré con las partes no verbales de su comportamiento religioso? Parecen inexplicables. Interpretémoslo no caritativamente, sin embargo, y tal comportamiento se ordenará adecuadamente (1981, p. 116).

Si solo proyectamos en el agente lo que consideramos verdadero utilizándonos a nosotros mismos como modelo, jamás lo comprenderemos adecuadamente. En efecto, ese modelo es demasiado egocentrista para poder ser correcto: convierte al agente en un desdoblamiento de la intérprete. Precisamente por eso es que la simulación se hace necesaria.

Veamos ahora cómo ocurriría toda la situación interpretativa. En primer lugar, la intérprete tiene que atribuir al agente aquellas creencias que ella considera obviamente verdaderas acerca de su mundo compartido. Luego de observar el comportamiento del agente y las nuevas circunstancias, ella deberá atribuirle aquellas creencias que ella piensa que tendría si fuera él. De esta manera, si ella observa al agente realizando un tipo de comportamiento que parece presuponer creencias falsas —imaginemos, por ejemplo, que él está adorando al Sol—, ella deberá imaginar cómo sería tener creencias verdaderas acerca de las características evidentes del entorno físico, pero no creencias científicas especializadas acerca de las características del Sol, mientras también se tiene creencias verdaderas acerca de la importancia del Sol para la vida humana. De aquí podemos encontrar razonable, y hasta predecible, que el agente crea que el Sol es una suerte de divinidad que debe ser adorada. La intérprete atribuirá al agente algunas creencias falsas acerca del Sol, lo que le permitirá encontrar la racionalidad de su comportamiento permitiéndole a ella comprenderlo mejor, pero contra el fondo de muchas otras creencias que ella considera verdaderas, es decir, que comparte con el agente.

7.3. Compartiendo formas de vida

Hay un sentido en que un concepto es un conglomerado de creencias, de manera que compartir un sistema de creencias es compartir también un sistema integrado de conceptos. No solo atribuimos a los demás nuestras creencias sino también nuestro concepto de creencia, y al interpretar al otro le atribuimos nuestra convicción de que el concepto de creencia involucra necesariamente la suposición de que las creencias pueden ser verdaderas o falsas, al punto de que creer que p entraña creer que p podría ser falsa. Pertenece al concepto de creencia que uno asuma que sus creencias son verdaderas, pero también pertenece al concepto de creencia el creer que esas mismas creencias podrían ser falsas. Este es un rasgo fundamental del falibilismo y es evidencia de flexibilidad cognitiva, que es la capacidad de adaptarse a situaciones, opiniones y puntos de vista diferentes e inesperados, sin perder la propia perspectiva pero poniéndola en cuestión y modificándola de ser necesario. La flexibilidad cognitiva se superpone con la empatía cognitiva, entendida como la capacidad para entender a aquellos cuyas opiniones no compartimos y reconocer que esas opiniones podrían permitirnos ver aspectos de la realidad que no habíamos reconocido.

Asumir esto es poseer los conceptos de subjetivo y objetivo: esto es, que hay una manera como las cosas son que es independiente de nuestra voluntad. Por ello, la intérprete deberá atribuir al agente, al mismo tiempo, un sistema de creencias verdaderas y el concepto de creencia, así como la capacidad para distinguir entre lo subjetivo y lo objetivo, es decir, una ontología realista y la creencia en que hay una verdad objetiva que no depende de las voluntades individuales.

Pero recuérdese que no pretendo probar que existe una verdad objetiva que sea independiente de las voluntades individuales; lo que sostengo es que es necesario que lo presupongamos para que la interpretación, y todo lo que viene con ella, sea posible. Esto también puede funcionar como un argumento trascendental, porque si es necesario para la interpretación que asumamos que existe una verdad objetiva, y como no podríamos evitar ser intérpretes, no podemos evitar creer que existe una verdad objetiva. Por otra parte, no puedo evitar comportarme como si existiese una verdad objetiva. Luego, si creo que existe una verdad objetiva, estoy lógicamente obligado a sostener que existe una verdad objetiva. Estas conclusiones objetivistas y realistas, procedentes de un argumento trascendental, son consecuencias inevitables del principio de caridad. Pero todo esto es posible solo si la intérprete puede triangular entre ella, el agente y el mundo que ambos comparten y que consideran que es independiente y previo a ellos.

Como he sostenido, una intérprete podrá atribuir a un agente una creencia solo al interior de un sistema mucho mayor de creencias y otros estados mentales

interconectados con esta. Ya que la definición de creencia que estoy empleando incluye, aunque no se reduce a, una disposición para actuar, entonces un sistema de creencias compartido por intérprete y agente puede ser visto como un sistema compartido de disposiciones para comportarse o como un sistema de prácticas sociales que constituye una forma de vida. Con frecuencia estas disposiciones no son plenamente conscientes para los agentes mismos, porque existen en un nivel preconceptual de interacción con el mundo y con las otras personas. Pero el punto que deseo sugerir es que la atribución de una creencia cualquiera implica la atribución de una forma de vida asumida como compartida. Podría objetarse que atribuimos creencias a bebes o niños muy pequeños incluso si no los consideramos todavía miembros de nuestra forma de vida. Sospecho que si pensamos que podemos atribuir creencias, así como el concepto de creencia, a un niño pequeño, deberemos también poder verlo como miembro de nuestra forma de vida, incluso si solo en un sentido muy rudimentario. Es muy probable, sin embargo, que la transición desde —de un lado— habilidades tácitas y preconceptuales hacia —del otro— creencias y prácticas sociales se realice de manera gradual y progresiva de forma que no haya una frontera entre estos dos extremos sino solo un continuo.

Según la condición (4), que no está explícita en todas las formulaciones del principio de caridad de Davidson, las creencias que la intérprete asume que comparte con el agente son creencias acerca de los objetos comunes, acerca de los significados de las palabras y acerca de los que constituyen propósitos razonables y deseables para sus acciones. Por ello, el principio de caridad incluye un supuesto fundamentalmente normativo. La intérprete no asumirá solamente que ellos comparten creencias acerca de cómo es el mundo sino acerca de cómo debería ser. La intérprete atribuirá al agente creencias descriptivas pero también valoraciones, con el objetivo de hacerlo inteligible.

La condición (4) incluye dos puntos. En primer lugar, la intérprete debe asumir que el agente actúa, en general y excepto casos de irracionalidad que discutiremos posteriormente, según su mejor juicio. En segundo lugar, ella debe asumir que ellos comparten, en líneas generales, propósitos a largo plazo. En consecuencia, ella deberá asumir que ellos comparten la creencia de que ciertas acciones son deseables porque son medios para alcanzar ciertos fines, los cuales son medios para alcanzar ciertos otros fines y así en adelante. Ella también deberá asumir que comparten creencias acerca de los fines a largo plazo. Naturalmente, la intérprete podrá estar en desacuerdo con algunos de estos medios y fines, pero deberá haber pasos en este encadenamiento de medios y fines —especialmente los más generales—, en que ella se encontrará de acuerdo con el agente. Este acuerdo, asumido como tal por la intérprete, le hará posible comprender la cadena de medios y fines de menor generalidad. De esta manera, el principio de caridad tiene un doble elemento normativo. Por una parte,

la intérprete no solo atribuirá creencias sino también valoraciones. Por otra parte, la intérprete estará compelida a atribuir todos estos elementos imbricados para poder comprender al agente.

Así, como vimos en el tercer capítulo, el holismo semántico contenido en el principio de caridad elimina las formas más radicales de relativismo e inconmensurabilidad en el nivel epistemológico, pero también lo hace en el nivel moral. Así como el discurso en general, también el discurso moral está gobernado por exigencias de consistencia y racionalidad. Pero tal vez la consecuencia más interesante sea una exigencia de solidaridad. Las condiciones de la interpretación nos exigen ver al otro como un semejante que es diferente. Por lo diferente que es, nos cuestiona y se convierte en un reto permanente para nuestras propias creencias y deseos, así como nos obliga moralmente a intentar entenderlo, porque nuestra autocomprensión está estrechamente ligada a la comprensión del otro. Del principio de caridad se deduce una obligación moral a tratar de entender al otro y a intentar incorporarlo a las posibilidades de nuestro propio discurso. Malpas lo dice de esta manera:

> Uno podría ponerlo en un lenguaje ligeramente diferente y decir que una de las implicaciones éticas del holismo es que el cuidado por nosotros mismos es inseparable del cuidado por los otros (1992, p. 187).

En efecto, el cuidado por uno mismo, tanto como la autocomprensión, pasa por el cuidado y comprensión del otro. La ética que se desprende del principio de caridad es una ética de la solidaridad y la caridad, esta vez no solo en un sentido epistémico sino también moral. Lejos de plantear el problema moral en términos de la pregunta por qué querría yo comprender al otro, para después intentar fundamentar una respuesta, la posición holista consideraría inevitable el desear comprender al otro y a uno mismo, mediante el reconocimiento de nuestro espacio compartido, como un solo proceso que es condición de posibilidad de toda acción en el mundo.

Sin embargo, aquí reaparece el fantasma que persigue al principio de caridad desde su nacimiento: la posibilidad de que sea una técnica filosófica de justificación del «imperialismo psíquico», como lo llama Nozick (1993 p. 153). Aparentemente esto se evidenciaría en la elección misma del término «caridad», de posibles connotaciones paternalistas. Ahora no desarrollaré en profundidad este punto, pues lo haré con más detalle en los próximos capítulos, pero sí me gustaría sugerir que, aunque no hay ningún vínculo explícito entre el principio de caridad y la virtud de la caridad, no puede ser tan arbitraria la elección del término. En efecto, la caridad es el cuidado desinteresado por el otro, pero es también el cuidado de uno mismo, y ambas formas de cuidado son inseparables. Se trata de una actitud de benevolencia en la interpretación y la interacción con el otro. Más que aludir a un tipo de paternalismo,

caridad implica *philía*. De hecho, el latín *caritas* es la traducción del griego *ágape,* que significa amor o cuidado, pero que además sugiere la idea de un espacio de enriquecimiento mutuo compartido, como lo fueron las comidas fraternales de los primeros cristianos.

Al hacer uso del principio de caridad, uno permite la comprensión del otro, pero también hace posible la comprensión de sí mismo, su autodescripción y, por tanto, también su transformación. Al tratar de comprender al otro, la intérprete comparará sus propias creencias y deseos con aquellos atribuidos al agente. Este proceso la obligará a poner en cuestión su propio sistema de estados mentales, lo que podrá conducir a una exigencia de tolerancia, adaptación y cambio, que permitirá la generación de un territorio común entre ambos interlocutores.

Comprender al otro no es, pues, reconstruir sus contenidos mentales en los nuestros, como sostendría una concepción internista de la mente, sino es más bien la actividad creativa de construir un territorio compartido: una comunidad de creencias, deseos, significados, valores, y objetos de la realidad. Este ámbito es construido pero también es hallado, porque pertenece al mundo que consideramos nos antecede en su objetividad. A su vez, es en este espacio donde justificamos nuestras más arraigadas convicciones, tanto en el terreno epistemológico como moral.

En el próximo capítulo continuaremos explorando aquellas habilidades contenidas en el fenómeno de la comprensión que nos permiten imaginar escenarios contrafácticos. Pero, sobre todo, me interesa aclarar la metáfora según la cual la comprensión involucra la ampliación de un territorio compartido.

Capítulo ocho
La comprensión como una actividad imaginativa

8.1. Comprensión y empatía

Escondidos bajo la palabra «comprensión» no hay uno sino muchos conceptos, así como hay muchas maneras de comprender algo y quizá sea una fútil tarea intentar buscar un rasgo común a todos ellos. Por eso mi interés principal es discutir un tipo particular de comprensión, el del comportamiento y los estados psicológicos de los agentes intencionales.

Como una cuestión metodológica cabría recordar que, cuando el objetivo es la comprensión, la interpretación es el instrumento, en tanto se trata de una actividad de atribución de estados mentales y acciones a un agente. La empatía es un elemento que debe estar presente en la interpretación para discriminar el tipo de estados mentales que serían atribuidos al agente. Sin embargo, si bien la empatía es condición necesaria para la comprensión, esta involucra elementos que no se reducen a la actividad empática, de manera que comprensión y empatía no son conceptos coextensivos. Por ejemplo, explicar el comportamiento de una persona en términos de regularidades puede ayudarnos a comprenderla mejor, lo que va más lejos de la pura empatía, aunque sin duda la empatía también está presente en la explicación del comportamiento en términos de regularidades nomológicas. Así, la comprensión intencional contiene, como un elemento necesario, aunque no suficiente, la capacidad empática[1].

En la primera parte de este capítulo me propongo analizar las versiones tradicionales de empatía en la comprensión del otro, con el objetivo de mostrar que suelen estar comprometidas con un modelo intencionalista, transposicional y monádico de lo mental. Posteriormente, intentaré reformular el concepto de empatía eliminando

[1] Para un interesante abordaje sobre las relaciones entre empatía y emociones véase Brunsteins, 2018. Para una revisión general de las conexiones entre el problema de la comprensión y las ciencias cognitivas véase Spaulding, 2018.

esos elementos que puedan sobrevivir en él. Deseo continuar desplegando la idea de que la comprensión del otro podría ser vista como la creación de un espacio compartido, con lo que intentaré explicitar y desarrollar algunas intuiciones que se encuentran en Wittgenstein y Gadamer e integrarlas con discusiones recientes en la filosofía de la mente y en la psicología experimental.

Una larga tradición filosófica sostiene que comprender al otro involucra la habilidad para identificarse con él imaginariamente. Hay varias versiones de esta tesis, pero un rasgo común es sostener que, para que la comprensión sea posible, la intérprete debe tener la capacidad de simular los estados mentales del agente bajo condiciones contrafácticas o debe tener la capacidad de simular que ella es él en circunstancias distintas. Estas habilidades han recibido distintos nombres en las diferentes escuelas filosóficas. Por ejemplo, *sympathy*, en Hume (1968, 2006) y Smith (1976); *Verstehen*, en Schleiermacher (1986, 1996) y Dilthey (1989); *simulation*, en Gordon y Goldman[2]; «empatía» en el uso psicológico habitual, *the replicating strategy*, en Heal (1986); y *re-enactment* en Collingwood (1946). Naturalmente estos términos no han sido empleados siempre de la misma manera y con frecuencia apuntan a objetivos diferentes, pero puede decirse que, en general, abordan un fenómeno semejante.

Hay, por otra parte, una importante diferencia entre empatía y simpatía, aunque son conceptos que con frecuencia se han confundido. Este es un tema sobre el que hay poco acuerdo. Algunos autores los usan intercambiablemente, mientras que otros distinguen diversos tipos de empatía. Tampoco hay acuerdo sobre si alguno de ellos es condición del otro. Emplearé una distinción cercana, aunque no idéntica, a la clásica distinción formulada por Wispé (1986, pp. 314-321), según la cual el objetivo de la empatía es el conocimiento del otro, mientras que el de la simpatía es el deseo de bienestar de otra persona. La palabra «empatía» alude a un fenómeno tanto cognitivo como emocional, mientras que «simpatía» a uno preferentemente emocional y evolutivamente más básico, al punto que se encuentra también en los bebes pequeños y en los animales superiores (De Waal, 1997). La empatía suele conllevar operaciones cognitivas complejas como la capacidad de representarse las representaciones ajenas y con frecuencia incluye un sentimiento de compasión. La simpatía no es necesariamente compleja desde un punto de vista cognitivo pero siempre implica elementos compasivos. Los bebes pequeños y los animales perciben —y con frecuencia comparten— los sentimientos ajenos, pero no necesariamente empatizan en el sentido de que se representen las representaciones ajenas. Esta idea se ve corroborada por los experimentos de la falsa creencia.

[2] Nozick, 1981, pp. 636-638, también cree que la comprensión del otro opera como un tipo de simulación bajo condiciones contrafácticas y que es una forma de inferencia por analogía.

La palabra «empatía» es de reciente uso. Titchener empleó *empathy* en su *Experimental Psychology of the Thought Processes* (1909), como una traducción al inglés del alemán *Einfühlung*, al discutir el uso que de este término hizo Lipps (1903). Titchener acuñó el término a partir del griego *empatheia*, que significa ser afectado por algo, y sugirió la siguiente distinción: empatía es la tendencia «*to feel oneself* into *a situation*», mientras que simpatía sería «*feeling together* with *another*» (Wispé, 1987; véase también Wispé, 1991; Scotto, 2004; Brunsteins, 2011).

«Empatía», *Verstehen* y *simulation* también han sido empleados de maneras diferentes por los autores. De hecho no estamos tratando de un solo concepto ni de conceptos intersecados, sino, a la manera wittgensteiniana, de conceptos superpuestos que tienen entre sí parecidos de familia. Sin embargo, por mor de claridad, es posible decir que un rasgo importante con frecuencia contenido en ellos es, en palabras de Blackburn, que se trata de una:

> […] proyección dramática que nos permite entrar en la posición de otra persona, viendo con nuestros ojos lo que desde los suyos debió haber sido que las cosas ocurrieran como ocurrieron —para que él dijera o escribiera o hiciera lo que hizo— (1995, p. 271).

Para Quine (1960, p. 219) también se trata de un acto «esencialmente dramático», en el que la intérprete se proyecta a sí misma al interior de la mente del agente. En efecto, para los griegos *drama* alude a un acontecimiento que merece especial atención y recuerdo por ser, como dice Aristóteles, «de carácter elevado y completo» (*Poética*, capítulo 6, 1449b). Se trata de un suceso que, siendo propio del agente, es universal porque pudo habernos ocurrido a nosotros y por eso nos identificamos con quien lo sufre, compadeciéndolo, admirándolo o compartiendo sus sentimientos.

Las versiones más antiguas de la noción de identificación, tanto simpática como empática, se pueden encontrar en dos lugares de Aristóteles. De un lado, en su concepción de la *katharsis* como el proceso por el que pasa el espectador de una tragedia al identificarse con los sufrimientos del personaje, para sentir con él sus desdichas y así expiar y eliminar sus propios padecimientos (*Poética*, capítulos 6, 13 y 14; *Política* VIII, 7). De otro lado, en su concepción de la compasión como la identificación con los pesares de otro, al asumir que podría haber sido uno mismo quien los padeciera (*Retórica* II, 8).

Sin embargo, los desarrollos más elaborados sobre estos temas recién los encontramos en Hume y Smith. Hume pensaba que la *sympathy* es un poderoso principio de la naturaleza humana que constituye el fundamento y el origen de la moral, ya que nos conduce a desplazarnos más allá de nuestra propia perspectiva y autointerés para poder capturar el beneficio de otra persona. Aunque Hume nunca definió

con precisión la simpatía, su idea era que se trata de una disposición para intuir los contenidos emocionales de los demás, independientemente de lo diferentes que ellos pudieran ser de nosotros mismos (1968, p. 316). Él sostenía que la simpatía está presente tanto en nuestra comprensión de otra persona como en nuestros juicios estéticos y morales. Hay aquí una conexión, que Hume no desarrolla, entre su concepción de la simpatía y su proyectivismo, es decir, su tesis de que tenemos una tendencia natural a proyectar nuestros sentimientos morales y estéticos sobre las personas y cosas. Por su lado, Smith es uno de los primeros en introducir en la discusión el concepto de imaginación, al afirmar que:

> Mediante la imaginación nos ubicamos en su situación, nos concebimos soportando los mismos tormentos. Es como si entráramos en su cuerpo y nos convirtiéramos en la misma persona que él, y, por tanto, nos formáramos cierta idea de sus sensaciones e incluso sintiéramos algo que, aunque más débil en grado, no es diferente de ellas (1976, pp. 47-48).

Los tempranos partidarios de la hermenéutica, especialmente Schleiermacher y Dilthey, sostenían que la comprensión (*Verstehen*) es un tipo de empatía, tal que comprender a una persona en una situación particular es poder revivir (*nacherleben*) o reconstruir en uno mismo las vivencias (*Erlebnis*) ajenas. Para Dilthey, la comprensión es el conocimiento psíquico subjetivo del otro a partir de signos exteriores. Esto implica, según él, revivir (*nacherleben*), reproducir (*nachbilden*) y transferir (*hineinversetzen*) (1949). Su idea es que lo que se comprende son vivencias significativas, en donde se entiende por significado «el modo especial de relación que dentro de la vida guardan las partes con el todo» (1949, p. 258). Aquí se está aludiendo al círculo hermenéutico que aparece con Schleiermacher y que contiene algunas de las intuiciones más importantes del holismo semántico y psicológico. Hay, además, algo de mucho valor: la intuición de que el significado es una relación y no una entidad mental, un objeto en sí mismo o una propiedad monádica.

Los enfoques de Hume, Smith, Schleiermacher y Dilthey pueden ser vistos como formas anticipadas de la teoría de la simulación, pero también son versiones del modelo transposicional según el cual la comprensión es posible cuando los estados mentales relevantes del agente son transpuestos a —y reproducidos en— la mente de la intérprete, lo cual presupone que los estados mentales tienen ya una determinación en la mente del agente antes del proceso de la interpretación y que de alguna misteriosa manera son transportados sin mayor alteración a la mente de la intérprete.

La formulación más precoz del modelo transposicional puede encontrarse en Locke (1982), quien consideraba que la función del lenguaje, que es compartido, es precisamente hacer públicos y trasladar de una mente a otra pensamientos que

son internos y privados. Locke piensa que la comunicación es exitosa si el hablante logra reproducir en la mente de la oyente pensamientos similares a los que él tuvo en el momento de la proferencia, es decir, si ambos atribuyen las mismas ideas a las mismas expresiones lingüísticas. Para este autor, los pensamientos son cadenas de ideas, en el mismo sentido que las oraciones son cadenas de palabras, y la comunicación es el proceso de transportar cadenas de ideas a través de cadenas de palabras, de la privacidad de una mente a la privacidad de otra.

El modelo transposicional estuvo asociado a la concepción de la empatía que era popular en los albores de la hermenéutica intencionalista, aquella según la cual comprender un texto es comprender las intenciones reales de su autor. El modelo transposicional de Dilthey está asociado a una concepción monádica y cartesiana de lo mental, en la que la mente es ese escenario interior y privado en que se constituyen nuestras ideas y representaciones, las cuales son objetos internos de los que nosotros somos espectadores privilegiados. En este modelo, los contenidos de los estados mentales se hallan determinados independientemente de la existencia de una intérprete, pues son entendidos como propiedades monádicas de una mente. Por ello esos contenidos son concebidos como aislados, en principio, de la existencia de una intérprete y de otras mentes. Así comprender al otro se torna un misterioso proceso de internamiento en la privacidad de la subjetividad ajena. Desde Peirce, Ryle, Heidegger, Wittgenstein, Gadamer y Davidson en adelante, una actividad filosófica favorita ha sido el mostrar las limitaciones del modelo cartesiano de lo mental y sus consecuencias nefastas para otros ámbitos de la filosofía. Queda la duda, sin embargo, de hasta qué punto hay dos Descartes. Uno es el Descartes histórico y el otro la influencia que él generó en la filosofía posterior y que resumimos con el nombre de «Descartes» para perfilar con más claridad nuestra concepción de la mente en oposición a esa tradición. Quizá, sin embargo, la expresión acuñada por Peirce (1988, pp. 88-122) «el espíritu del cartesianismo» sea más apropiada para el segundo.

Son muchas las objeciones que se han hecho a esta concepción de lo mental y no voy a pasarles revista ni me detendré a reconstruirlas. Solo diré que el modelo monádico y transposicional está comprometido con una concepción de lo mental centrada en la noción de privacidad de la que muchos autores recientes, y particularmente Wittgenstein (1988), se esmeraron en sacarnos.

8.2. Comprensión de conceptos y formas de vida

En su cuestionamiento a la concepción cartesiana de lo mental, Wittgenstein intuyó estas líneas de argumentación y prefirió sostener que la comprensión tiene lugar cuando se comparte una forma de vida. Para este autor la comprensión es la capacidad

de participar en prácticas, actividades o proyectos comunes, por eso el modelo de la comprensión es la participación en un juego regido por reglas implícitas que conocemos de manera tácita. Comprender no es solo un proceso mental sino sobre todo práctico, sostiene (1988, § 154). No requiere internarse en el mundo privado del otro sino ser capaz de compartir con él una forma de vida. Esta idea se encuentra, naturalmente, muy cerca de la metáfora de la fusión de horizontes de Gadamer, como veremos más adelante.

Las intuiciones más agudas de Wittgenstein acerca de la comprensión se producen cuando discute la naturaleza de la creencia religiosa. En efecto, un caso particularmente extremo de incomprensión se da cuando una persona no creyente trata de comprender al creyente, pues las oraciones del creyente tienen un significado diferente en tanto pertenecen a una forma de vida distinta. Por eso, cuando el creyente dice que habrá un juicio final y la no creyente lo niega, no necesariamente se están contradiciendo (1976). De igual manera, cuando la no creyente evalúa el significado y valor de verdad de la oración «Dios existe», suele asumir que es una oración sintética que pretende describir un hecho de la realidad que se puede dar o no, por lo que la oración puede ser verdadera o falsa. El creyente puede no asumir nada de eso. Para él, esta oración puede tener un significado básicamente moral y no metafísico ni fáctico. De hecho, es probable que haya sido en este sentido que Wittgenstein entendiera la religión.

Wittgenstein no explicitó los detalles de su intuición, de manera que voy a intentar hacerlo ahora sin pretender hacer exégesis. La no creyente no necesita creer lo que cree el creyente para comprenderlo, pero sí tiene que ser capaz de seguir el tipo de justificaciones y relaciones inferenciales que le permiten a él creer lo que cree. Supongamos que antes de interactuar con el creyente, la no creyente desconociera esas relaciones inferenciales y que después de interactuar con él las llegara a conocer. La no creyente habrá generado ciertas creencias que antes no tenía y que ahora tiene, y muchas de ellas serán compartidas con el creyente, por lo menos en lo correspondiente a esas relaciones inferenciales. A eso llamo generar un espacio compartido entre ambos.

Así pues para que la no creyente comprenda al creyente no es requerido ni posible que se interne en su privacidad, sino que entienda el significado de lo él que dice y hace. Pero el significado de cualquiera de sus creencias, por ejemplo «habrá un juicio final» o «Dios existe», solo podrá comprenderse si se conocen las otras creencias con las que estas están vinculadas. Pero como entender estas creencias requiere de compartir algunas de ellas, entender las creencias implica participar de una forma de vida. Es como entender un concepto. Los conceptos son resúmenes miniaturizados de complejos procesos sociales, de manera que comprender un concepto requiere especialmente de estar en condiciones de participar de esas prácticas sociales.

Así, comprender un concepto de otra cultura —por ejemplo, *gavagai* para la intérprete radical— es estar en condiciones de imaginar participar en las redes sociales que se encuentran encapsuladas en ese concepto. Esto es propiamente ampliar nuestra forma de vida para dar un espacio a otra, en la nuestra. Por otra parte, esas redes sociales articulan nuestra comprensión de la realidad y la realidad misma, en tanto esta también incorpora las estructuras sociales que nos hacen posible conocerla, de manera que comprender un concepto no es diferente de comprender la realidad expresada por este. Aquí se ve que los juegos de lenguaje, las formas de vida y la realidad misma son cognitivamente difíciles de separar. La comprensión requiere de cierta participación en todo ello.

Entonces, dado que los conceptos son resúmenes miniaturizados de procesos sociales complejos, comprender un concepto es ser capaz de imaginar participar en esos procesos. Eso es propiamente ampliar nuestra forma de vida —es decir los procesos sociales que habitamos— para incluir otros en los nuestros. Este es el elemento universal de la comprensión y su estructura, no el tema, materia u objeto de la comprensión. Aquí hay un punto de contacto con Gadamer, quien también defiende un elemento formal universal en la comprensión.

La participación en actividades comunes, como el juego y la fiesta, es también el modelo paradigmático de la comprensión para Gadamer (1977a). Él rechaza el concepto romántico de empatía y lo sustituye por la metáfora de la fusión de horizontes (1977b). Por eso, a diferencia de la hermenéutica intencionalista, Gadamer considera que el sentido no está contenido únicamente en las palabras del hablante o del texto, sino que debe ser «completado» por la situación histórica de la intérprete. Dice:

> El sentido de un texto supera a su autor no ocasionalmente sino siempre. Por eso la comprensión no es nunca un comportamiento solo reproductivo, sino que es a su vez siempre productivo. Bastaría decir que, cuando se comprende, se comprende de un modo diferente (1977a, p. 367).

Para la hermenéutica intencionalista, comprender adecuadamente es revivir y reproducir fielmente, representar correctamente los estados mentales del otro, sus deseos, intenciones o creencias. Pero el problema con esta concepción es que asume que el significado y los contenidos de los estados mentales están ya dados en el agente antes e independientemente de la interpretación. Por eso, en esa concepción, la intérprete es solo una receptora pasiva del sentido o de los estados mentales ajenos y no tiene ningún rol en su constitución. La hermenéutica gadameriana, así como la estética de la recepción (véanse, por ejemplo, Mayoral, 1987; Warnig, 1989; Jauss, 1992), objeta esos supuestos al considerar que la intérprete colabora en el proceso de la constitución de sentido. Gadamer no dice nada, sin embargo, acerca del proceso

de la constitución de los contenidos de los estados mentales, que en este libro he propuesto que emergen en tanto propiedades relacionales triádicas en una situación comunicativa. Para Gadamer un libro no es el papel y la tinta, sino el sentido que damos a sus palabras y el sentido no lo fija el autor sino es el producto de la tensión cooperativa entre autor e intérprete. Como cada lectura, cada interpretación hace del texto un producto nuevo: «El verdadero sentido contenido en un texto o en una obra de arte no se agota al llegar a un determinado punto final, sino que es un proceso infinito» (1977b, p. 368).

Dado que el sentido del texto no está contenido únicamente en lo que el texto tiene que ofrecer, sino que debe ser «completado» por la situación histórica de la intérprete, si el lector completa el texto entonces en la lectura hay más contenido que en el texto. Así, la contraposición entre la hermenéutica romántica intencionalista y la de Gadamer radica en que el hermeneuta romántico cree encontrar más en su lectura que en el texto, pero porque supone haber entendido el texto, en virtud de su ventaja histórica, mejor que quien lo escribió. Gadamer, por el contrario, cree encontrar más en su lectura que en el texto, pero porque al interpretarlo ha añadido a la lectura contenidos que no estaban en el texto. La distancia en el tiempo no es un obstáculo que deba ser salvado, como ingenuamente creían los románticos intencionalistas, sino un privilegio que debe ser aprovechado. Gadamer escribe como si la fusión de horizontes implicara algún tipo de acuerdo:

> Comprender lo que alguien dice es, como ya hemos visto, ponerse de acuerdo en la cosa, no ponerse en el lugar del otro y reproducir sus vivencias (1977b, p. 461).
>
> La conversación es un proceso por el que se busca llegar a un acuerdo (p. 463).
>
> Comprenderlo no quiere decir primariamente reconstruir una vida pasada, sino que significa participación actual en lo que se dice (p. 470).
>
> Es tarea de la hermenéutica explicar este milagro de la comprensión, que no es una comunión misteriosa de almas sino participación en un sentido comunitario (p. 361-362).

Estas citas convergen con las afirmaciones de Davidson (1984c, p. XVII) en que el objetivo de la interpretación no es el acuerdo (*agreement*), sino más bien hacer el desacuerdo inteligible, donde la inteligibilidad del desacuerdo implica un fondo de acuerdo tácito, pero que no es previo, sino que se genera en tanto la interacción comunicativa transcurre. En eso Gadamer y Davidson confluyen con la idea hermenéutica de una precomprensión como condición de posibilidad de la comprensión. Sin embargo, Davidson es más explícito en que se trata de una precomprensión dinámica que se construye en la interacción comunicativa. Para él esto es posible gracias al

fenómeno de la triangulación, en que hablante e intérprete se atribuyen mutuamente estados mentales en relación al mundo objetivo que comparten y asumen compartir.

Como hemos visto, Davidson piensa que la comprensión se da cuando elaboramos pequeñas teorías al paso sobre el fondo de nuestras teorías previas para intentar dar sentido al comportamiento del otro. Estas provisionales «teorías» son atribuciones de estados mentales, acciones y significados con las que entretejemos nuestra relación con el otro, o, expresado en lenguaje gadameriano son «anticipaciones de sentido».

Así pues, la posibilidad del acuerdo está dada por la generación de formas de vida que compartimos y así es como se construye la comprensión. No se trata de un horizonte que compartimos, ni solo de una fusión de ellos, sino de una multiplicidad de formas de vida superpuestas que son prácticas sociales que compartimos en tanto las constituimos.

Tengo la impresión, sin embargo, de que la noción de fusión de horizontes y la idea de compartir una forma de vida, centrales para el tema que nos ocupa, pueden ser planteadas con más claridad[3]. Además, no me parece que el concepto de empatía deba estar necesariamente comprometido con una hermenéutica intencionalista, con un modelo transposicional de la comunicación ni con una concepción monádica y cartesiana de la mente. Por ello, me propongo explorar las posibilidades de reformular el concepto clásico de empatía, mientras examino cómo puede verse la comprensión como la creación de un espacio compartido. Deseo sugerir que la comprensión no requiere revivir imaginariamente los estados mentales ajenos, sino que más bien tiene lugar cuando quien comprende expande su subjetividad constituida intersubjetivamente, creando un espacio para albergar al otro, con lo que también se transforma. En lo que sigue abordaré el debate acerca de la simulación para después desarrollar más ampliamente una de las propuestas centrales de este libro.

El modelo clásico de la simulación sostiene que la intérprete está en condiciones de comprender al agente, si ella es capaz de imaginarse en las circunstancias y los estados mentales de él. Esto requiere que ella pueda desplazarse creativamente desde su propia perspectiva a la del otro, para poder «ver» el mundo, así como para evaluar las situaciones y circunstancias, de una manera semejante a como ella cree que él lo hace. Esto le permitirá dar respuestas a las preguntas sobre por qué él hizo lo que hizo o por qué tiene los estados mentales que tiene.

Para Goldman (1992a, 1992b, 1993, 1995a, 1995b, 2006, 2013) la simulación ocurre cuando la intérprete formula preguntas de la siguiente forma: ¿Qué creería o sentiría yo, y qué hubiera hecho, si hubiera estado en su lugar? En estos casos uso mi propia personalidad con mis propios estados mentales, como un modelo

[3] Dos valiosos proyectos que podrían verse como intentos por hacerlo, aunque diferentes del mío, son Schatzki, 1966 y Brandom, 1994.

para simular los de él en sus propias circunstancias. Así, me pregunto qué haría yo —con mi propia biografía, experiencias, etcétera— y qué creería o sentiría si estuviera pasando por lo que creo que él está pasando. Luego, infiero por analogía que él haría o sentiría algo similar.

El modelo de Goldman presupone que primero realizo una introspección para conocer cuáles serían mis estados mentales en ciertas circunstancias imaginarias y luego infiero que el otro tendría estados mentales semejantes. Un problema con este modelo es que es difícil entender la idea de seguir siendo «yo», con mis propios estados mentales y mis experiencias, estando en una circunstancia diferente de las que normalmente constituyen lo que soy. El modelo de Goldman parece compartir el presupuesto cartesiano de que el yo tiene cierta consistencia previa e independiente de sus experiencias y circunstancias. Pero mientras más altere mis circunstancias reales, más problemático se vuelve entender qué soy «yo»[4]. Si imaginariamente cambio radicalmente mis circunstancias, se vuelve imprecisa la identidad de lo que puede seguir siendo llamado «yo». Por eso, en oposición al modelo de Goldman, Gordon (1986, 1992, 1995a, 1995b, 1995c, 1995d) ha sugerido que no es que yo simule ser yo mismo en las circunstancias del otro, sino más bien que simulo «ser» el otro. En este caso, simulo directamente ser él en sus propias circunstancias y con sus propios estados mentales, sin que haya una inferencia desde mis estados mentales hacia los suyos.

Veamos con más detalle las diferencias entre estas dos posiciones. Para Goldman la simulación requiere que la intérprete mantenga fijos sus estados mentales, mientras cambia las circunstancias externas. Para Gordon la simulación requiere que la intérprete cambie tanto sus estados mentales como sus circunstancias externas y, así, necesita de una transformación, más que de una transferencia o proyección. Según Gordon, el modelo de Goldman no me permitiría comprender al otro sino a mí mismo, si las circunstancias de mi vida hubieran sido diferentes. Este modelo sería fuertemente egocentrista, en tanto vería a los demás como variantes de uno mismo.

Pero no está claro que Gordon resuelva los problemas que el modelo de Goldman ha puesto al descubierto. Por una parte, la comprensión no requiere imaginar ser otro, sino en todo caso requiere imaginar ser otro manteniendo la propia perspectiva. Por otra parte, si ya es bastante difícil imaginar los estados mentales que yo tendría si estuviera en las circunstancias en que está otro, es decir, si ya es problemático mantener mi identidad personal en circunstancias contrafácticas, ¿cómo podría yo imaginar ser otro, manteniendo mi propia identidad? Gordon podría acentuar los problemas que Goldman no puede resolver.

[4] «Así, cuando fundimos nuestro ser real en roles irreales, generalmente no sabemos cuánta realidad hay que mantener constante» (Quine, 1960, p. 62).

La teoría clásica de la empatía, cuyos residuos se mantienen en la teoría de la simulación de Goldman y Gordon, está asociada a una concepción de lo mental que ve a los individuos monadológicamente, es decir, como sujetos independientes cuyos contenidos de estados mentales son propiedades monádicas que, en principio, no requieren de la existencia de una intérprete. Es esta concepción la que hay que superar y para ello es necesario ver a los contenidos de los estados mentales en la dimensión intencional de la interpretación como propiedades relacionales y no monádicas, es decir, no como propiedades privadas de los individuos sino como propiedades relacionales que emergen en situaciones interpretativas y que, por tanto, existen en los vínculos entre las personas.

Así, por ejemplo, desde una descripción física, cuando uno describe los procesos neuronales que acontecen en el cerebro de un individuo para explicar su comportamiento o su experiencia fenoménica, los eventos cerebrales son asumidos como propiedades monádicas de sistemas neurológicos que con frecuencia son causados por objetos externos, aunque estas descripciones físicas siguen estando subdeterminadas por la evidencia observacional disponible. Pero, al interior de una descripción intencional, como cuando hacemos una interpretación psicológica, los contenidos proposicionales de los estados mentales son propiedades relacionales que requieren de un agente a quien se adscribe el estado mental y de una intérprete que es quien lo adscribe desde sus propios criterios de interpretación. De esta manera, se puede entender que uno no requiera de la existencia de una intérprete para tener la experiencia fenoménica de un estado mental, pero sí es necesario ser una criatura interpretable para que se pueda determinar el contenido de sus estados mentales. Naturalmente, esto no significa que la intérprete tenga que estar ahí, sino que el agente debe ser potencialmente interpretable por una intérprete. Esta idea está asociada a la tesis wittgensteiniana según la cual solo podemos fijar los contenidos de nuestros estados mentales mediante reglas públicas de determinación[5].

Una manera apropiada de abordar nuestro problema es mostrar que no necesito internarme en la mente del otro, aunque sí debo imaginar tener sus estados mentales o ser él mismo. Pero esta actividad imaginativa es condición necesaria, aunque no suficiente. En condiciones ideales, la comprensión requiere que integre mis estados mentales con los suyos respecto a mundo que asumimos compartir, con lo que se crea un espacio compartido. Esto sigue siendo un acto de imaginación empática y también es un tipo de transformación, pero de una manera diferente. Al simular ser él, en muchos casos debo alejarme de mis creencias previas, con lo que extiendo mis estados mentales y creo un espacio que incorpora los del otro.

[5] Wittgenstein: «Un "proceso interno" necesita criterios externos» (1988, parágrafo 580).

En este proceso no pierdo mi perspectiva previa ni mis creencias acerca del otro, pero las transformo. Se trata, por tanto, de una transformación más radical de la que ve Gordon. Es este ámbito de transformación de mí mismo el espacio que he creado para dar un lugar al otro en su diferencia, lo que constituye propiamente la comprensión. Esta transformación es una extensión del yo, pues la idea no es perder la propia perspectiva sino ampliarla. La empatía solo es el punto de partida: al simular ser el otro, extiendo mi propio yo y creo un espacio compartido con él. Marcia Cavell está en una dirección semejante:

> La *empatía* no puede ser cuestión de que en alguna forma yo salga de mi propia mente y entre en la suya, sino que se basa más bien en descubrir y ampliar la base común que compartimos, usando mi imaginación en relación con la creencias y deseos que usted pueda tener con respecto a los cuales su comportamiento le parezca a usted más o menos razonable (2006, p. 66).

La interpretación se produce cuando la intérprete atribuye un sistema básicamente coherente de estados mentales al agente. Pero los contenidos de estos estados mentales atribuidos no pretenden ser aquellos que existen en el interior de la mente del agente con independencia de la intérprete. Por el contrario, siempre son relacionales, es decir, son los contenidos de los estados mentales del agente «para la intérprete». La coherencia que se busca no está en el agente en sí mismo sino en la relación entre agente e intérprete. Que los contenidos de los estados mentales sean relacionales triádicos significa que no son atributos monádicos de una mente, sino atributos que surgen y existen en la interacción entre individuos que se interpretan mutuamente en relación con el mundo que comparten. Pero insisto en que lo relacional es los contenidos proposicionales de los estados mentales, no su experiencia fenoménica, pues la experiencia subjetiva del deseo o de los afectos es claramente una propiedad monádica, subjetiva y privada. Por otra parte, en el ámbito de la descripción intencional los contenidos de los estados mentales son una relación y emergen cuando las personas interactúan interpretándose mutuamente. Adoptar esta tesis nos aleja definitivamente de entender la mente como una substancia cartesiana con propiedades monádicas. La tendencia natural que tenemos a sospechar de esta concepción relacional de la mente no es sino la evidencia de la poderosa presencia que aún tiene la concepción cartesiana en nosotros. Ramberg ha formulado esta idea de la siguiente manera:

> Las propiedades mentales, entonces, no son autónomas ni rasgos intrínsecos de alguna entidad; son relacionales. Ellas son individuadas y constituidas (en parte) por objetos que están más allá del sujeto. Las propiedades mentales de las personas son un sistema de relaciones entre la persona y su entorno. Que la mente involucre

> o esté constituida por lo que es más o menos distante de la persona es todavía intuitivamente extraño. La experiencia de esta extrañeza, sin embargo, procedente de nociones de contenido estrecho es un síntoma de la obstinación de una concepción de la mente como substancia (1997, p. 467).

Naturalmente, una persona aislada, pensemos en Robinson Crusoe antes de encontrar a Viernes, no solo tiene experiencias fenoménicas de estados mentales sino también estos tienen contenidos, lo que es posible porque él ya ha sido parte de una comunidad de personas que lo ha convertido en intérprete potencial, en potencialmente interpretable e incluso en autointérprete. Pero incluso ahora, cuando nosotros hablamos de él y consideramos su caso, estamos atribuyéndole contenidos de estados mentales que son relacionales en su naturaleza.

El caso de la autointerpretación es delicado. Hay un sentido en que uno no se interpreta a sí mismo, como sostuvieron Wittgenstein y Davidson, porque uno no hace oraciones-T para sí mismo dada la autoridad de la primera persona. Pero hay otro sentido más austero en que uno sí es un autointérprete, pues uno no solo experimenta estados mentales sino también los escruta y les atribuye significación. Uno tiene creencias, por ejemplo, acerca de sus propios estados mentales, su importancia y valor; acerca de su intensidad y de la manera en que se relacionan con otros estados mentales que uno también tiene. Con frecuencia esa valoración se constituye en la experiencia emocional, que es la textura misma de la subjetividad. Naturalmente, la interpretación que uno tiene de sus propios estados mentales ha sido constituida a partir de sus experiencias personales y a la forma de vida a la que pertenece. Con frecuencia uno se autointerpreta cuando piensa en un yo previo y trata de entender por qué hizo lo que hizo o tuvo los estados mentales que tuvo. También puede uno imaginar un posible yo futuro para preguntarse qué haría o qué sentiría en una circunstancia hipotética. Pero también podría preguntarse por qué tiene actualmente ciertos estados mentales o por qué no tiene los que piensa que debería tener. En todos esos casos, la autointerpretación no es muy diferente de la interpretación que hacemos de otro y puede presentar las mismas dificultades.

Al interior de la descripción intencional no hay hechos objetivos —*facts of the matter*— sobre los contenidos de lo que cree, desea, siente o hace alguien —lo que incluye a uno mismo— si no es en relación a una interpretación, lo que no quita realidad ni objetividad a los estados mentales o acciones, sino solo los contextualiza. Tomando el conocido ejemplo de Davidson, se puede decir que no hay un hecho objetivo de cuál es la temperatura de un objeto si no es en relación con una escala de medida. Una vez fijada la escala de medida podremos decir, por ejemplo, que es un hecho que cierto líquido se encuentra a 0° C y que también es un hecho que está a 32° F. Sería absurdo suponer que podría haber un hecho de cuál es la temperatura

del objeto independientemente de una escala de medida. De igual manera, puede ser un hecho que alguien crea, desee o haga algo en el contexto de cierta interpretación, sobre todo si esta es compartida por muchas intérpretes y agentes. Esto no da prioridad a la intérprete, sino a la relación entre intérprete y agente respecto al mundo que comparten. Así como diferentes escalas de temperatura pueden medir correctamente el calor que hace actualmente en Lima, aunque den cifras diferentes pero simultáneamente correctas —porque cada una de estas cifras debe entenderse en el contexto de un patrón de medida mayor—, de igual manera distintos manuales de interpretación o sistemas de creencias pueden ser simultáneamente correctos, siempre que apliquen los diferentes patrones de la manera adecuada.

Nada de eso significa que los estados mentales no tengan una realidad objetiva. La tienen, siempre que se entienda que esa realidad es determinada mediante una interpretación triangular que incluye al agente, a la intérprete y al mundo compartido. Pero otras interpretaciones podrían ser igualmente buenas. Así, bajo cierta interpretación puede ser un hecho que Fernando invierte en un viaje a Egipto porque ama la cultura antigua y cree que ese viaje le proporcionará experiencias que siempre deseó tener. Pero bajo otra interpretación también puede ser un hecho que Fernando ha gastado una cifra excesiva en un viaje que en este momento no es urgente, porque teme no tener la oportunidad de hacerlo en el futuro y cree que si no lo hace ahora lo lamentará posteriormente. Claramente ambas interpretaciones no son incompatibles, pero podría imaginarse escenarios en que lo fuesen. Supongamos que una intérprete piense que Fernando es un loco redomado que está malgastando los ahorros de su vida para ir a ver piedritas rodeadas de arena. En ese caso, habría que examinar las tres interpretaciones teniendo en cuenta criterios epistémicos para determinar cuál de ellas da cuenta de más información adicional, es más predictiva de nuevo comportamiento del agente, es más consistente internamente y con otras evidencias, cuál es más coherente con ciertos valores y presupuestos compartidos, como por ejemplo la importancia del conocimiento histórico y la transitoriedad de la vida, etcétera. Para determinar cuál interpretación es preferible, la opinión del propio agente es tan útil y tan importante como las interpretaciones de otras intérpretes igualmente informadas y agudas. Estas afirmaciones son una exploración del principio de indeterminación de la interpretación en relación al concepto de comprensión.

El punto es que la comprensión no requiere únicamente de imaginar lo que el otro está sintiendo sino también exige crear o enriquecer un espacio común, un vínculo o una conexión. La simulación es condición necesaria pero no suficiente. Toda comprensión del otro es una actividad creativa de enriquecimiento y autoconciencia de una relación, real o posible, cuya consecuencia última es una transformación personal. Pero esto no ocurre solamente porque uno se convierta imaginariamente

en el otro, sino porque en un importante sentido uno se transforma a sí mismo, lo que también transforma la relación.

La idea central en la tesis de que la comprensión es la creación de un espacio compartido es que al atribuirle al agente los estados mentales que nosotros creemos que tendríamos si fuéramos él o aquellos que creemos que él tiene dadas las circunstancias en que creemos que está, hacemos un esfuerzo por participar de su perspectiva y, así, compartimos algo de su espacio personal. Si él no sabe que lo estamos interpretando o incluso si se trata de un personaje de ficción, el espacio compartido es solo asumido. Sin embargo, si él se propone hacer lo mismo con nosotros quizá podamos crear el espacio compartido que originalmente asumimos tener y que será constituido en el proceso mismo de la interacción. Así pues, al interpretar al otro tejemos y articulamos relaciones entre nuestros estados mentales y los que le atribuimos—en relación con un mundo objetivo que compartimos y asumimos compartir— y que reinterpretamos, así como nos reinterpretamos a nosotros mismos y al otro en tanto la interacción avanza. Este proceso de reinterpretación del otro —que se produce cuando hacemos modificaciones en los estados mentales que le atribuimos—, de interpretación de nosotros mismos —que ocurre cuando modificamos las valoraciones de nuestros propios estados mentales gracias a la influencia del otro— y de reinterpretación de la realidad objetiva que compartimos —que acontece cuando uno o ambos modificamos nuestros estados mentales acerca de la realidad— tiene como consecuencia un retejido y una rearticulación de las relaciones triangulares entre intérprete, agente y mundo.

Al comprender al otro la intérprete no reconstruye los contenidos mentales del otro en su propia mente, como sostendría una concepción clásica de la empatía, reproduciendo en su escenario privado lo que ella cree que ocurre en el escenario interior ajeno. La comprensión es, más bien, la actividad creativa de construir un ámbito compartido: un territorio común de creencias, significados, deseos, valores y objetos de la realidad, donde las diferencias y discrepancias puedan hacerse perspicuas. Este terreno es elaborado pero también es hallado, y es en él donde confrontamos y justificamos nuestras diversas convicciones. Comprender a un agente es también ser capaz de contar una historia que exhiba las razones y motivos —creencias y deseos, conscientes o no— que él tuvo para actuar como lo hizo. Este tipo de narración puede o no estar interesada en elaborar generalizaciones como las que se realizan al explicar el comportamiento mediante regularidades de forma nomológica, en las que se subsumen relaciones causales entre estados mentales y acciones. De esta manera, una acción o estado mental es comprendido a la luz de otras acciones y estados mentales que nos permiten darle sentido.

8.3. Comprensión, teorías de la mente y simulación

La simulación es una prolongación del procedimiento usual que seguimos cuando lidiamos con nuestras propias circunstancias del mundo. Al bregar con las incertidumbres de mi propio futuro, al planificar mis acciones y tomar decisiones, debo ser capaz de imaginar posibles escenarios que me aguardan. Puedo preguntarme, por ejemplo, cómo debería reaccionar, qué debería creer o qué sentiría, si tal o cual situación me ocurriera. También puedo preguntarme cuál sería la acción razonable si lo que deseo es evitar ciertas consecuencias indeseables. Así, simulo ser yo mismo en diferentes circunstancias para poder planificar exitosamente mis acciones. No es que deje de ser lo que soy, más bien amplío mi ámbito de subjetividad, mis creencias, sentimientos y deseos, para incorporar aspectos que no son reales sino posibles. Poseer un simulador interno de este tipo tiene un alto valor de supervivencia, porque nos permite probar imaginariamente diferentes cursos de acción antes de que nos involucremos en ellos y corramos el riesgo del fracaso. Presumo que también empleamos este simulador para predecir el comportamiento de otras personas y elegir los mejores cursos de acción frente a ellas. Es posible que la comprensión involucre, en alguna medida, una extensión de estas capacidades, aunque no debe asumirse que el simulador interno sea anterior —filogenética u ontogenéticamente— al simulador que nos permite dar sentido al comportamiento ajeno. Es la misma facultad que se aplica de maneras diferentes, pero que fue seleccionada y evolucionó por las mismas razones.

No es probable, sin embargo, que la simulación sea la única estrategia que empleamos cuando comprendemos a las personas. También podemos hacerlas inteligibles al encontrar regularidades que gobiernan relaciones causales entre sus estados mentales y sus acciones, como sostiene la Teoría de la Teoría (T-T) (Stich & Nichols, 1995; Gopnik & Wellman, 1995; Fodor, 1995), pero pienso que la simulación también debe estar presente. Como vimos en el segundo capítulo, la T-T sostiene que la comprensión de otros requiere de una teoría de la mente. Tener una teoría de la mente es tener un conjunto de principios generalizadores, formulados en términos de regularidades nomológicas, que gobiernan las relaciones causales entre los estados mentales y las acciones de las personas.

La teoría de la simulación (T-S), por otra parte, normalmente sostiene que la posesión de habilidades simulativas es suficiente para la comprensión. Intentaré mostrar que, independientemente de si también está presente un conjunto de principios de generalizaciones nomológicas, la simulación es condición necesaria en la comprensión de las personas, aun si no lo es en su explicación, y esto porque uno podría explicar el comportamiento de una persona encontrando relaciones causales subordinadas a regularidades nomológicas, sin poder comprenderla. Se necesita la simulación

para poder aplicar una regularidad nomológica. Pero, además, la explicación del comportamiento solo requiere de un punto de vista externo, mientras que la comprensión exige un esfuerzo por ingresar al espacio del otro, a la fenomenalidad de sus estados mentales, así como a sus valoraciones y a la manera como él imprime significado a las cosas, y eso no sería posible si no se hubiera algún tipo de actividad de simulación.

Explicar el comportamiento de una persona incluye poder contestar a preguntas como ¿por qué realizó esas acciones? o ¿cuáles son sus creencias, sentimientos y deseos? Para contestar a estas preguntas podría bastar un sistema de regularidades nomológicas en las que se pueda subsumir conexiones causales entre estados mentales y acciones. Pero, aunque el conocimiento de estas regularidades podría permitir predecir el comportamiento de la persona, no bastaría para poder decir que lo hemos comprendido. La comprensión requiere, como condición necesaria, la posibilidad de redescribir el comportamiento y los estados mentales del otro en términos de los nuestros, es decir, poder integrarlos en un mismo espacio en crecimiento. Imaginar las razones que tuvo el otro para hacer lo que hizo nos obliga a alejarnos de nuestras propias creencias previas para poder ver las situaciones desde su punto de vista, lo que nos permitirá encontrarlo razonable desde nuestra perspectiva, es decir, nos permitirá reconocerlo como un agente racional para quien las acciones y situaciones resultan significativas en virtud de evaluaciones, propósitos y objetivos. Hacer esto es necesario para poder aplicar generalizaciones nomológicas en situaciones específicas, permitiendo excepciones e imaginando cursos de acción alternativos[6].

Más aún, comprender al otro supone apreciar la manera como él se percibe como individuo y esto incluye saber por qué cree él que hizo lo que hizo, independientemente de si coincidimos con su evaluación. Blackburn lo caracteriza de esta manera:

> Estudiar los rasgos mentales de las personas es estudiar su autocomprensión, es decir los conceptos bajo los cuales se incluyen a sí mismos y a sus acciones: los conceptos que determinan sus planes, actividades y pensamientos […]. Cuando llego a comprender por qué tú actuaste como lo hiciste no me interesa ubicarte en alguna red causal nomológica, sino ver el objetivo de tus acciones. Comprenderte es una actividad distinta, no reducible a ver tu comportamiento solo como parte de lo que generalmente ocurre, parte de un patrón científicamente repetible. Comprenderte es como deliberar qué hacer en situaciones posibles, lo que hago al «recentrarme», imaginando la situación y respondiendo desde dentro (1995, p. 276).

[6] «Parece ser un hecho, congruente con la teoría de la simulación, que cuando la experiencia de vida de un agente es muy diferente a la nuestra, nos resulta más difícil predecir o explicar su comportamiento» (Goldman, 1992b, p. 190).

Ahora bien, podría objetarse que la simulación involucra algún grado de teorización, algún tipo de teoría tácita acerca de la gente y sus reacciones (Dennett, 1998b [1987]). Es posible. Incluso podría ocurrir que esas teorías tácitas pertenezcan a una estructura previa de habilidades no conscientes —lo que Searle llama un *background of intentionality* y que no está demasiado lejos de la noción heideggeriana de la precomprensión—, que desarrollamos en parte como consecuencia de nuestro desarrollo biológico y en parte al pertenecer a una cultura en particular. Es posible que las habilidades simulativas comiencen como una preestructura de habilidades no conscientes; no me parece objetable llamar a eso una teoría tácita, aunque es un uso bastante austero de teoría. Sin embargo, lo que distingue a la simulación de cualquier teoría empírica acerca de la naturaleza es que la segunda está fundamentalmente concernida con la predicción, mientras la primera está comprometida con capturar la perspectiva y significación que el otro da a las cosas, así como la manera en que él se ve a sí mismo. Por ello, la explicación en términos de regularidades nomológicas no es ni necesaria ni suficiente para la comprensión, aunque podría ser parte de ella. De otro lado, la simulación es necesaria para la comprensión aunque no es suficiente para la explicación.

La simulación presupone la capacidad de metarrepresentación. Esto es, la intérprete no solo debe representarse el mundo sino también debe poder representarse las representaciones que los demás tienen del mundo, sin perder por ello su propia perspectiva, mediante la formación de creencias de segundo o más grados. Esta habilidad ha sido extensamente estudiada en niños por Joseph Perner (1991) y otros autores. Según Perner, el niño pasa de ser un representacionalista —una habilidad desarrollada alrededor de los dos años— a ser un metarrepresentacionalista rodeando los cuatro años. Este cambio se puede constatar empíricamente con los experimentos denominados de la «falsa creencia» a los que ya nos referimos.

Como hemos visto en el segundo capítulo, los experimentos sugieren que hay procesos neurológicos que se activan antes de los tres o cinco años y que permiten la simulación. Esto se vería reforzado por la evidencia de que los niños con autismo entre los seis y los dieciséis años suelen desaprobar la prueba de la falsa creencia, incluso si tienen un coeficiente intelectual promedio. Los niños autistas carecen de vínculos afectivos con las personas y normalmente no distinguen entre objetos y sujetos, habilidades que parecen requerir de algún grado de empatía. De otro lado, los niños con síndrome de Down, con mucho menor coeficiente intelectual, no suelen tener problemas con la prueba de la falsa creencia (Baron-Cohen, Leslie & Frith, 1985, pp. 37-46; Dunn, 1991). Los niños autistas también tienen problemas para comprender oraciones contrafácticas (Harris, 1995), para comprender metáforas

(Happé, 1995, pp. 275-295; MacKay & Shaw, 2004, pp. 13-32), así como para desarrollar la capacidad de actuar, es decir, de concebirse a sí mismos en diferentes situaciones posibles (Leslie, 1988). Se sabe que estos niños tienen poca capacidad para imaginar situaciones de ficción, es decir, situaciones que involucran cursos de acción alternativos y realidades ilusorias (Harris, 1995; Currie, 1005). También parecen tratar a las personas y a las cosas por igual.

Algunos autores (Fonagy, Gegerly, Jurist & Target, 2002) prefieren usar «mentalización» para la capacidad de dar sentido al comportamiento propio y ajeno vía la atribución de estados mentales. También se llama a esto la facultad de comprender o dar sentido a la incomprensión —*understanding misunderstanding*—. Esta actividad, que puede ser consciente o inconsciente, se desarrolla espontáneamente en el infante; sin embargo, su desarrollo puede verse afectado por el tipo de apego que tenga con sus cuidadores tempranos.

Fonagy y otros (2002, p. 64) sugieren que un apego temprano inseguro probablemente originará una mentalización deficiente. En casos extremos, puede producir comportamientos antisociales al forzar al individuo a ver a los demás no como agentes intencionales sino en términos no humanos, como miembros de un cuerpo social, un grupo o una posición social. La mentalización surge en el infante a partir de la habilidad de su cuidadora de sintonizar con él, leyendo y modulando sus estados mentales. A su vez, la capacidad del infante de reconocer los estados mentales de los demás y los de él mismo está estrechamente ligada a la función reflectiva paterna. Más aún, una adecuada función reflectiva paterna permitirá que el niño sepa modular sus estados mentales apropiadamente en circunstancias inesperadas. La «función reflectiva» es el referente operacional de la capacidad mentalizadora, lo que puede ser medido en los niños. Una escala de la función reflectiva permite medir la habilidad que tenga el niño en reconocer sus propios estados mentales y los ajenos, en descripciones narrativas de situaciones interpersonales (Slade, 2005).

Ciertamente, hay un sentido en que los niños normales menores a los cuatro años se vinculan con las personas aún sin requerir de la habilidad desarrollada de la metarrepresentación cognitiva. Los niños pequeños están en condiciones de triangular con la madre y con algún objeto común. Así, por ejemplo, si el bebe ve que la madre mira hacia algún objeto distante con interés, el niño sigue la mirada de la madre y reacciona ante las reacciones de ella. De esta manera, el niño comparte el punto de vista ajeno sin perder el propio. Todo esto sugiere que la empatía se desarrolla a partir de habilidades preconceptuales y no conscientes de las que el niño se vuelve progresivamente más consciente, aunque probablemente haya un fondo en estas habilidades que siempre forma parte del proceso de la comprensión. Como acabamos de ver, es probable que estas habilidades progresen mediante la combinación de cierto tipo

de educación que sea sensible a los sentimientos de las demás personas, combinado con un desarrollo neurológico cuyas raíces están en mecanismos adaptativos de origen evolutivo. Estas raíces pueden ser encontradas en lo que suele llamarse *motor mimicri*: los mecanismos subliminales que hacen que un individuo imite el comportamiento físico de otros individuos en ciertas circunstancias, especialmente las expresiones faciales y las posturas corporales. Considérese, por ejemplo, el mecanismo que conduce a un grupo de animales a dejar lo que está haciendo para mirar todos en una misma dirección, en caso que uno de ellos así lo haya hecho. Estos comportamientos tienen su base en relaciones de apoyo mutuo entre animales de la misma especie, lo que presta un servicio a la supervivencia. Cuando estos comportamientos de apoyo no se realizan de padres a hijos, reciben el nombre de «conductas auxiliares». Darwin estudió este tipo de comportamiento en *The Descent of Man* (1994 [1871]) y señaló que se trata de un tipo de simpatía, lo que probablemente permitiría, en especies con otro tipo de evolución, el desarrollo de una conciencia o sentido de lo moral. Esto, por otra parte, podría brindar fundamento empírico a las tesis de Hume de que el origen y el fundamento del juicio moral se encuentran en el sentimiento moral de la simpatía (Sherman, 1998).

Hay una venerable lista de filósofos que han postulado la existencia de estructuras preconceptuales, prelógicas y preconscientes que son condición de posibilidad de la comprensión. La lista incluye a Ryle (*knowing how*), Polanyi (*tacit knowledge*), Heidegger (la precomprensión)[7], Wittgenstein[8], Dreyfus, Haugueland y Searle (*the background of intentionality*)[9], entre otros. Polanyi describe estas habilidades como aquellas en que:

> [E]l sujeto llega a conocer una operación práctica, pero no sabe cómo lo logró. Esta suerte de percepción subliminal tiene la estructura de una habilidad, pues una habilidad combina acciones musculares elementales que no son identificables según relaciones que no podemos definir (1966, p. 8).
>
> [Al comprender a las personas] usamos los estados anímicos del rostro sin poder decir, excepto muy vagamente, mediante qué signos lo hemos logrado conocer (p. 5).

[7] Antes de Heidegger, la idea de un fondo no proposicional y preintencional fue discutida por Husserl, 1973a y 1973b, mediante la noción de «mundo de la vida» (*Lebenswelt*).

[8] Wittgenstein solía aludir a esta noción con una serie de metáforas: el lecho del río, la estructura animal, los cimientos de la casa, las prácticas sociales compartidas, etcétera (1988).

[9] Searle distingue entre lo que él llama *the network* y *the background*. Lo primero es el sistema de estados mentales al que un estado mental en particular pertenece. Lo segundo es el sistema de habilidades y prácticas no proposicionales y preintencionales que hacen posible la existencia de la *network* (1983, pp. 141-159).

Estas son las habilidades que Ryle tenía en mente cuando propuso la célebre distinción entre *knowing how* y *knowing that*:

> Si la comprensión no consiste en inferir ni adivinar, los supuestos precursores interiores de acciones manifiestas, ¿qué es? Si no requiere del dominio de una teoría psicológica con la habilidad de aplicarla, ¿qué conocimiento requiere? […]. La comprensión es un tipo de *knowing how.* El conocimiento requerido para comprender realizaciones inteligentes de algún tipo es cierto grado de competencia en realizaciones de este tipo (1949, p. 54).

Citaré a Dreyfus para resumir la tesis de la existencia de un *background of intentionality* que hace posible la comprensión:

> Y así como podemos aprender a nadar sin adquirir una teoría, ya sea consciente o inconsciente, de la natación, adquirimos un fondo de prácticas sociales [*social background practices*] al desarrollarnos en ellas, no al formar creencias ni aprender reglas (1980, p. 7).
>
> Lo que hace a este fondo [*background*] no es creencias, ya sea explícitas o implícitas, sino hábitos y costumbres, incorporadas en la suerte de habilidades sutiles que exhibimos en nuestra interacción diaria con las cosas y las personas (p. 7).

Dreyfus no considera estas prácticas como creencias porque posiblemente tenga en mente una concepción representacionalista de las creencias, pero si las entendemos como disposiciones para actuar, no habría razón para no entender el *background* como un sistema de creencias no conscientes.

Es, pues, posible que nuestras habilidades metarrepresentacionales comiencen como habilidades no conscientes. También es posible, como estos autores sostienen, que estas habilidades permanezcan como un fondo previo a lo largo de nuestras vidas, siendo el contexto necesario para toda comprensión consciente y conceptual. Pero lo relevante ahora es que todo esto sugiere que las habilidades de simulación, mentalización y metarrepresentacionales se adquieren y cultivan progresivamente cuando el niño se acostumbra a ponerse en el lugar del otro para imaginar una vida que no es la suya pero que podría haberlo sido, y eso ocurre cuando el niño se enfrenta a situaciones nuevas, imprevistas y desconcertantes. De esta manera, el niño llega a aprender que hay diferentes aunque compatibles interpretaciones de la misma situación.

Así pues, la compleja actividad de la comprensión involucra al viejo fenómeno de la empatía, ahora reformulado como simulación, que se puede seguir entendiendo como la capacidad para vivir imaginariamente la vida de otro. Pero esta metáfora no debería verse como la habilidad para reconstruir en nuestro escenario interior lo que creemos que está ocurriendo en el suyo. La comprensión es más bien la actividad

de ampliar nuestro horizonte o forma de vida compartida, sintiéndonos así afectados por la diferencia, la diversidad o el sufrimiento ajeno, pues el otro es quien nos confronta radicalmente impulsándonos al autocuestionamiento. Esta confrontación nos transforma interiormente y nos mueve a ampliar esa subjetividad que se ha conformado en nuestras relaciones con los demás para así crearle al otro un lugar en nosotros mismos.

En este punto surge una pregunta para la cual aún no tenemos una respuesta definitiva, aunque sí varios atisbos: ¿qué surgió primero en el ámbito filogenético, la simulación o el lenguaje? Dado que no existen fósiles lingüísticos ni psicológicos, no nos queda más que hacer suposiciones sobre la base de correlaciones con la ontogénesis o con otras evidencias secundarias. Hasta donde sabemos, el cerebro inició su explosivo crecimiento hace tres millones de años, pasando de 450 cc a 1450 cc, en la actualidad. De acuerdo con Dunbar (2003, 2009), el crecimiento del cerebro fue causado por la necesidad de adaptarse a sociedades complejas, lo que aumentó los grados de intencionalidad de nuestra atribución psicológica. El origen de esta carrera habría sido la necesidad de cooperar y competir, lo que exigía mecanismos de cognición social suficientemente complejos. Así, entonces, las habilidades de simulación y metarrepresentación debieron haber sido anteriores a la evolución de un lenguaje plenamente formado, con semántica, sintaxis y habilidades pragmáticas, lo que debió haber ocurrido entre 200 000 y 100 000 años antes del presente. Pero es de suponer que antes de que apareciera un lenguaje plenamente formado existió una o más formas de protolenguaje, así como formas de comunicación animal no lingüísticas. Parece razonable suponer, por tanto, que hacia el comienzo del crecimiento exponencial del cerebro (hace unos tres millones de años) hubo formas de comunicación animal no lingüísticas y también mecanismos básicos de atribución psicológica, lo que seguramente requería de formas rudimentarias de simulación y metarrepresentación, como las que podemos encontrar en bonobos y chimpancés. Casi con certeza, estas formas de comunicación y de atribución psicológica se potenciaron mutuamente, lo que permitió un crecimiento aún más rápido del cerebro y estableció las bases para la aparición del lenguaje. Para cuando ya hubo un lenguaje relativamente maduro (entre 200 000 y 100 000 años antes del presente), la complejidad de atribución psicológica (probablemente hasta en tres grados de intencionalidad) inició otra carrera coevolutiva, en este caso entre lenguaje y simulación, también potenciándose mutuamente. No creo que se pueda decir, por tanto, que la simulación fue anterior ni posterior a los mecanismos de comunicación no lingüísticos, aunque sí sería razonable que fuera anterior al lenguaje ya formado. Sin embargo, cuando surgió el lenguaje, con todas las funciones que actualmente

tiene, generó modificaciones cerebrales y cognitivas importantes. Entre otras cosas, el lenguaje permitió:

- Verbalizar anticipadamente los escenarios posibles y las opciones disponibles, lo que permitió entenderlos mejor y operar con ellos de manera imaginaria, como una suerte de prueba piloto sin los riesgos que implica la acción. Esto hizo posible una simulación más precisa y exitosa.
- Facilitar la comunicación y, por tanto, la transmisión a los otros de las prácticas sociales beneficiosas. Por tanto, posibilitó también la transmisión cultural y la propagación de los memes exitosos para la supervivencia.
- Categorizar el mundo compartido (externo) y la vida subjetiva (interna) con el fin de construir conceptos públicos para hablar tanto de lo externo como de lo subjetivo. Esto afinó nuestra percepción del mundo y de nuestros estados mentales subjetivos.
- Verbalizar los estados mentales propios y ajenos, lo que permitió simularlos con mayor detalle y experimentarlos subjetivamente de manera más nítida. Esto también hizo posible la comunicación de los estados mentales propios de manera más precisa y detallada.
- Expresar nuestros estados mentales, así como los ajenos, lo que permitió construir y verbalizar estados mentales en varios niveles de intencionalidad, probablemente más de cuatro niveles. Por ejemplo: «Me duele la cabeza», «Me preocupa que me duela la cabeza», «Me entristece que me preocupe que me duela la cabeza». Pero también: «Jorge le dijo a María que a él le preocupa que ella piense que él cree que le duele la cabeza».
- Emitir nuestros estados mentales en varios niveles de intencionalidad, lo que potenció la metacognición y esta, a su vez, la conciencia autobiográfica o extendida, también llamada autoconciencia, subjetividad o yo.
- Verbalizar nuestros estados mentales en varios niveles de intencionalidad, lo que facilitó la aleoatribución y autoatribución psicológica, y potenció, a su vez, la aleocomprensión y autocomprensión.
- Explicitar las relaciones e inferencias lógicas. El lenguaje permitió la construcción de teorías explicativas sobre el mundo, conformadas por conceptos.

De esta manera, la cognición, en sus diversas formas, el lenguaje y, posteriormente, la aparición de la cultura fueron coevolucionando y potenciándose mutuamente.

A lo largo de todos estos capítulos hemos hablado de los estados mentales. Ahora es el momento de preguntarnos directamente qué es la mente, si la entendemos como una red estructurada de estados mentales. Cuando comprendemos a una persona no solo comprendemos su mente sino también la comprendemos con nuestra mente y, evidentemente, el fenómeno del autoconocimiento es el proceso mediante el cual una mente se conoce a sí misma. Por eso debemos preguntarnos qué es, pues, lo mental.

Capítulo nueve
¿De qué hablamos cuando hablamos de la mente?

9.1. El problema de las relaciones entre la mente y el cuerpo

Interpretamos el comportamiento de agentes intencionales, es decir, de criaturas cuyas acciones han sido causadas por sus estados mentales. A su vez, los estados mentales estructurados constituyen la mente y la interpretación es un proceso realizado por una mente. ¿Pero qué es la mente? ¿Cuál es su estatuto ontológico? ¿De qué manera existe? ¿Cómo se relaciona con el cerebro y con el resto del cuerpo? Responder a estas preguntas significa comenzar a contestar la interrogante sobre qué es la vida mental, aquel rasgo del ser humano que creemos que nos diferencia del resto de la naturaleza, aunque bien podría ser que encontremos mamíferos no humanos de los que también pueda decirse que tienen una mente rudimentaria, aunque esta tenga características diferentes de la nuestra. Lo que es menos probable, aunque no imposible, es que encontremos algún grado de subjetividad en especies no humanas, es decir, criaturas dotadas de mentes con la capacidad de reflexionar sobre ellas mismas, que se tienen a sí mismas como objeto de su actividad cognitiva y emocional. Esto requeriría de varios grados de intencionalidad y, por tanto, de la capacidad de tener estados mentales —tanto cognitivos como afectivos— sobre otros estados mentales igualmente cognitivos y afectivos[1].

Los problemas clásicos de la naturaleza de lo mental y de las relaciones entre la mente y el cuerpo han sido abordados desde la antigüedad clásica. En la filosofía griega las dos posiciones más importantes al respecto son la de Platón, desarrollada

[1] Los candidatos naturales serían los chimpancés, bonobos —chimpancés pigmeos—, babuinos y otros primates no humanos (Cheney & Seyfarth, 2007; De Waal, 2013). Los pulpos también parecen tener una conciencia compleja que procede de que sus neuronas están distribuidas por todo su cuerpo, especialmente por sus tentáculos (Godfrey-Smith, 2017).

en el diálogo *Fedón* (2010), quien sostiene un dualismo de substancias y objeta expresamente una forma de emergentismo; y la de Aristóteles en el tratado *De Ánima* (1983), quien defiende una forma de naturalismo no reductivista, cercano a lo que hoy llamaríamos un monismo ontológico de aspecto dual. Sobre esa posición volveremos al final de este capítulo.

Desde los comienzos de la Edad Media —hacia el siglo V d.C.— los filósofos solían identificar «mente», «alma» y «espíritu» uniendo dos grupos de propiedades. Por una parte se atribuía al alma las características que los griegos solían adscribir a lo que ellos llamaban *psyché*, como voluntad, conciencia y autoconciencia, la facultad de razonar y el lugar de las emociones. En efecto, Aristóteles definía la *psyché* —que fue traducida como *anima* por los latinos— como *arché tes kinéseos kai aisthesis*, lo que en castellano sería principio o fundamento del movimiento y de la experiencia sensorial (Quintanilla, 1990).

De otro lado, los filósofos medievales también atribuían al alma el rasgo de la inmortalidad —de origen judeocristiano—, pues, aunque muchos filósofos griegos creían que el alma es inmortal, y Platón es el representante paradigmático, no todos ellos lo creían. Es recién hacia el siglo XIX que los filósofos vuelven a distinguir ambos grupos de propiedades —las de origen griego y las de origen judeocristiano—, y reservan las primeras para lo que hoy llamamos «mente» o «aparato psíquico» y las segundas para el concepto religioso de «alma». Como es claro, en este libro estamos hablando de las primeras y no de las segundas.

A lo largo de la modernidad, desde aproximadamente el siglo XVI hasta el XIX inclusive, un tema central de reflexión filosófica fue lo que hoy llamaríamos filosofía de la mente: el problema de las relaciones entre mente y cuerpo; la naturaleza de la sensibilidad, la conciencia y la autoconciencia; las características del entendimiento y la razón, las emociones y pasiones. La mayor parte de filósofos de ese período escribieron sobre estos temas, como los títulos de sus obras lo evidencian: Descartes (1936, 1965, 1968), Locke (1982), Berkeley (1974), Hume (2002), Leibniz (1992), Kant (2009), Hegel (1966). Sin embargo, es recién a mediados del siglo XX que la filosofía de la mente surge como una disciplina filosófica por derecho propio y *El concepto de lo mental*, de Ryle (1949), fue uno de los hitos en su nacimiento.

A lo largo de los siglos XX y XXI, la filosofía de la mente aborda los viejos problemas filosóficos asociados con la naturaleza de la mente, pero con el añadido de que lo hace en permanente intercambio con disciplinas empíricas como las ciencias cognitivas, las ciencias de la evolución, la lingüística, las neurociencias, el psicoanálisis y la psicología experimental, entre otras. Al día de hoy, pretender hacer filosofía de la mente sin tener en cuenta esos desarrollos sería una tarea ilusa.

En líneas generales, entenderemos por mente los procesos psicológicos que tienen por lo menos una de las dos siguientes características:

(1) Involucran experiencias fenoménicas, es decir, son procesos conscientes o que pueden en principio llegar a ser conscientes. Para expresar esta característica, que implica la experiencia de sentir algo, a veces se usa en castellano las expresiones «darse cuenta» o «percatación», como traducciones de las palabras inglesas *sentience* o *awareness*.

(2) Están dotados de intencionalidad o capacidad representacional, es decir, están dirigidos a algo diferente del estado mental mismo. En inglés se suele usar el neologismo *aboutness* para designar esta propiedad.

Hay estados mentales que poseen conciencia e intencionalidad; otros, solo conciencia o solo intencionalidad, pero no los hay que carezcan de ambos atributos. Ejemplos de estados mentales conscientes e intencionales son los sentimientos y las emociones; de estados mentales conscientes pero no intencionales son los dolores; de estados mentales no conscientes pero intencionales pueden ser las creencias inconscientes. Mientras los dolores son experiencias fenoménicas aunque carecen de intencionalidad, las creencias tienen un contenido proposicional que va más allá de la creencia misma, pero no necesariamente implican una experiencia fenoménica.

Así, entonces, para que un estado físico pueda también ser descrito como un «estado mental» es necesario que tenga por lo menos una de estas dos propiedades: experiencia fenoménica o intencionalidad. En consecuencia, podremos decir que una criatura posee una mente si, además de tener un soporte físico como un cerebro o algo semejante, posee estados fenoménicos o intencionales. Es materia de discusión si, además de nuestra especie, hay primates no humanos dotados de mente y a partir de qué edad se puede decir que un infante humano tiene una mente. Con frecuencia estos debates se entrampan en decisiones estipulativas que suelen despertar discusiones acaloradas, pero poco fecundas, pues en gran medida depende de la amplitud de los criterios que uno desee usar.

Ahora bien, el problema de las relaciones entre la mente y el cuerpo involucra varias dimensiones, entre ellas las más importantes son la ontológica y la causal. La primera tiene que ver con la existencia o no de dos substancias diferentes, y la segunda con la manera en que mente y cuerpo interactúan causalmente entre sí. Abordaremos ambas dimensiones simultáneamente analizando las diversas posiciones que se han ofrecido en la filosofía de la mente reciente.

9.1.1. Dualismo de substancias

Aunque hay varias formas de dualismo, lo central del dualismo de substancias es sostener que el ser humano es un compuesto de dos substancias o entidades ontológicamente diferentes, cuerpo y mente, siendo el cuerpo material pero la mente no. Un primer problema con esta posición es que si el cuerpo es material —y por tanto espacio-temporal— pero la mente no lo es, no queda claro cómo fluye la mente en el tiempo y, sobre todo, cómo puede estar conectada —y de qué manera— con algo que sí es espacio-temporal como el cuerpo. Se podría sostener que la mente es solo temporal y no espacial, como pensaban Agustín y Kant —es decir que se distiende en el tiempo, pero no se extiende en el espacio—, pero eso contradiría todo lo que se cree en la física contemporánea sobre la imposibilidad de separar espacio y tiempo. Por otra parte, también surge la pregunta de cómo podría conocerse la existencia y características de una substancia que, por definición, está más allá de todos nuestros instrumentos naturales y artificiales de observación. Adicionalmente, si la mente no es física, ¿de qué manera está asociada a un cuerpo que sí lo es? Existen dos formas clásicas de dualismo ontológico de substancias.

9.1.1.1. Dualismo de substancias interaccionista

Esta forma de dualismo sostiene que ambas entidades —mente y cuerpo—interactúan causalmente entre sí. Afirma la existencia de la mente como la única forma de explicar características observables del cuerpo que aparentemente no podrían ser explicadas de ninguna otra manera, como por ejemplo la voluntad, la conciencia, el libre albedrío o la racionalidad. El caso paradigmático de esta posición es Descartes, quien era tanto un dualista como un mecanicista. Es decir, él consideraba posible explicar mecánicamente la mayor parte de fenómenos físicos y psicológicos, incluyendo por ejemplo las emociones (1965), pero creía que los demás fenómenos mencionados no podían ser explicados mecánicamente y que su explicación requería de postular la existencia de una mente no mecánica y no física a la manera de una entidad inferida. Como una objeción a esta posición puede decirse que quizá en el pasado haya sido necesario postular la existencia de entidades no físicas para explicar fenómenos físicos, pero difícilmente se diría hoy que necesitamos postular la existencia de una substancia no material para explicar el funcionamiento del cuerpo (Quintanilla, 2013).

Otro problema con esta forma de dualismo es que no puede explicar cómo dos substancias ontológicamente diferentes pueden interactuar causalmente entre sí, dado que aparentemente solo pueden interactuar causalmente objetos ontológicamente semejantes. Pero podría oponerse a esa objeción que presupone un concepto

demasiado clásico de causalidad. ¿Qué razón hay para que una entidad no física no pueda ser causa de una física o viceversa? En verdad ninguna, excepto nuestra intuición de extrañeza. Pero una objeción a la contra objeción es que la causalidad es un concepto físico y, por tanto, no podría aplicarse a algo no físico. Por otra parte, si asumimos que esta interacción puede explicarse, habría que explicar cómo se pueden individualizar mentes que no son espacio-temporales. Es decir, ¿en qué sentido mi mente está vinculada a mi cuerpo y es diferente de otra mente, si no es un objeto espacio-temporal?

9.1.1.2. Dualismo de substancias paralelista

Como consecuencia de los problemas anteriormente señalados, este tipo de dualismo rechaza la posibilidad de interacción entre cuerpo y mente. Dos defensores clásicos de esta posición son Leibniz (1972) y Malebranche (1972-1986). Según el primero, los eventos físicos y los mentales transcurren en líneas paralelas, pero sincronizadas gracias a una armonía preestablecida por Dios, la cual da la impresión de que interactúan, pero en realidad eso no ocurre. De acuerdo con el segundo, quien denominó a su posición «ocasionalismo», Dios interviene para posibilitar la interacción entre cuerpo y alma, cuando esta es necesaria. La tesis general es que Dios es el único agente causal y que los seres humanos son solo ocasiones para que Él participe en la creación.

9.1.2. Fisicalismo

El fisicalismo es una forma de monismo que considera que la realidad está toda ella constituida únicamente por entidades físicas. Esta posición no tiene que negar la existencia de la mente, simplemente niega que la mente sea un objeto no físico. Así, habitualmente el fisicalismo sostiene que «mente» es el nombre que le ponemos al cerebro —o al sistema nervioso central— cuando desconocemos su funcionamiento, mientras que cuando podemos explicarlo en términos físicos simplemente lo llamamos cerebro. Algunos pioneros del fisicalismo son David Armstrong (1966) y Keith Campbell (1987), quienes llamaron a su posición *central state materialism,* pues no distinguen entre materialismo y fisicalismo, distinción que otros autores hacen para diferenciar entre, de un lado, la posición que sostiene que todo lo que existe es material y, del otro, la que afirma que todo lo que existe es aquellas entidades con las que estarían comprometidas nuestras mejores teorías físicas en condiciones ideales. Para algunos autores el concepto de fisicalismo es más amplio que el de materialismo, pues dirían que existen —o pueden existir— entidades físicas que no son materiales. En todo caso, el fisicalismo es una familia de concepciones que da lugar a varias versiones.

9.1.2.1. Fisicalismo reductivista

Esta posición no solo sostiene que lo que llamamos mente no es otra cosa que el cerebro sino también afirma que toda explicación psicológica puede en principio ser reducida a una explicación física. Afirma, sin embargo, que la explicación psicológica podría perder significado relevante al ser reducida a una explicación física. Es decir, la mente se podría reducir al cerebro desde un punto de vista ontológico y epistémico, pero no semántico. Si el fisicalismo no es solo epistémicamente reductivista sino también semánticamente reductivista suele recibir el nombre de eliminativismo.

9.1.2.2. Fisicalismo eliminativista

Esta es la versión más radical del fisicalismo, en tanto sostiene que el discurso y la explicación psicológica podrían, en principio, ser eliminados totalmente para ser substituidos por el discurso y la explicación física, sin pérdida de capacidad explicativa ni de significado. Se trata, por tanto, de un reductivismo ontológico, epistémico y semántico. Los mayores representantes del eliminativismo son Paul Churchland (1979a, 1979b) y Patricia Churchland (1986, 2002).

Hay, sin embargo, algunos problemas generales con las diversas formas de fisicalismo reductivista. Por una parte, todos los proyectos que intentaron reducir los fenómenos intencionales a fenómenos físicos —o las explicaciones psicológicas a explicaciones puramente físicas— han fracasado, lo que ha conducido a que haya cierto acuerdo en que los métodos explicativos de las disciplinas físicas y los de las disciplinas intencionales son diferentes y no pueden ser reducidos entre sí. Los fenómenos humanos que más se resisten a ser reducidos a algo físico serían la normatividad y la intencionalidad. Por otra parte, el fisicalismo reductivista presupone el representacionalismo, es decir, la tesis de que hay en principio una descripción definitiva y correcta de la realidad que sería la descripción física. Esta posición es objetable porque convierte una descripción de la realidad en la descripción privilegiada. Además, presupone que la descripción física es, en algún sentido, más real o correcta que la descripción intencional, lo que es ya una petición de principio.

9.1.3. Funcionalismo

Esta es una versión de fisicalismo no reductivista. Originalmente propuesto por Hilary Putnam y posteriormente abandonado por él mismo, sostiene que cuando usamos el vocabulario psicológico describimos fenómenos físicos en virtud a la función que realizan y no en virtud a lo que son. A su vez, estas funciones se describen desde el punto de vista de los objetivos de la acción humana en general. La idea es que es posible explicar y reducir funcionamientos más complejos en términos del funcionamiento

de procesos más específicos. En otras palabras, los nombres de estados mentales no describen entidades físicas ni mentales sino funciones de entidades físicas. El punto es que no interesa tanto la materia de la que esté hecho el cerebro sino la función que desempeñan sus partes y su estructura. Así, en principio, una computadora organizada con suficiente complejidad podría desarrollar conciencia y mente si tuviera procesos específicos que funcionaran de la manera requerida. Como se imaginará, este es el fundamento teórico de la inteligencia artificial: los proyectos para crear procesos cognitivos complejos en soportes computacionales no humanos.

Algunas formas de funcionalismo han postulado una teoría de la identidad tipo-tipo (*type-type*) entre eventos físicos y eventos mentales, con lo cual cada tipo de evento mental —un determinado tipo de recuerdo, un determinado tipo de emoción, un determinado tipo de creencia, etcétera—, sería idéntico a un determinado tipo de evento físico. Sin embargo, en general esa posición ha sido abandonada en favor de una teoría de la identidad caso-caso (*token-token*), la cual propone que cada caso o instancia de evento mental —por ejemplo, una instancia de recuerdo o una instancia de creencia— es idéntico a una instancia de evento físico, aunque no hay una identidad entre un tipo de evento físico y un tipo de evento mental[2]. La idea en la teoría de la identidad caso-caso es que cualquier estado mental es idéntico a un determinado estado del cerebro, pero no necesariamente distintos casos de estado mental son idénticos al mismo tipo de estado cerebral.

En ambos casos, la mente sobreviene al cuerpo, en el sentido de que los términos que describen estados mentales no tienen el mismo significado que los términos que describen estados físicos, pero los estados mentales sobrevienen de los estados físicos. La idea en la «sobreveniencia» o superveniencia —*supervenience*— es que dos cuerpos que tienen exactamente la misma organización física en todas sus instancias deben generar los mismos estados mentales, con lo cual esos estados mentales sobrevendrían de los estados físicos. Así, dos eventos físicamente idénticos deben ser mentalmente idénticos también, pero no viceversa, y cualquier cambio físico acarreará un cambio psicológico, pero no viceversa. Por tanto, dos eventos que tengan las mismas propiedades físicas deben tener las mismas propiedades mentales, pero no viceversa. La idea es que lo mental depende de lo físico, pues sin propiedades físicas no podría haber propiedades mentales, pero lo mental no es reducible a lo físico, pues las propiedades mentales no pueden ser reducidas a propiedades físicas.

[2] La distinción entre tipo y caso —*type* y *token*— es de gran utilidad en filosofía. Un tipo es una clase de objetos que comparten un rasgo común. Un caso o instancia es cada uno de esos objetos. Sea, por ejemplo, el conjunto A = {5, 5, 5}, ¿cuántos números 5 hay en él? En el conjunto A hay un tipo de 5, pero tres casos o instancias de 5.

Un ejemplo de esto es la teletransportación. Si se destruyera un cuerpo humano y se creara otro idéntico en todos los detalles celulares y atómicos, en principio y si el fisicalismo es correcto, se debería producir la misma experiencia subjetiva (Parfait, 1984). Esto es así en el mismo sentido en que si una pintura es hermosa, otra pintura que tuviera exactamente la misma disposición de colores y trazos también tendría que serlo, incluso si físicamente no es la misma. De esta manera, en el mismo sentido en que la belleza sobreviene de la disposición de los colores y trazos, la mente sobreviene de los estados físicos del cuerpo. Esta posición parece una alternativa viable de fisicalismo no reductivista, en tanto cree que los estados mentales son idénticos a los físicos, pero las explicaciones psicológicas no se reducen a explicaciones físicas.

9.1.3. El emergentismo

El emergentismo, en su versión débil, es una posición que fue desarrollada en el siglo XIX con el nombre de epifenomenalismo. Sostiene que la fenomenalidad de lo mental emerge o procede de estados físicos y que los estados mentales son físicos, aunque de una mayor complejidad que no puede ser reducible a lo físico.

El emergentismo fuerte, por otra parte, sostiene que un sistema complejo de propiedades físicas puede permitir la emergencia de propiedades no físicas. En ese sentido, el emergentismo fuerte puede identificarse con una forma de dualismo de propiedades —como veremos pronto— o incluso con un dualismo de substancias, si sostiene que una substancia no física puede llegar a emerger de una substancia física. Esta posición sostendría que a partir de objetos físicos de cierto tipo emergen objetos de un tipo diferente, como una suerte de salto cualitativo en que una forma de substancia surge a partir de otra como producto de una particular combinación de esta última. Los emergentistas débiles dirían que así como la vida emerge de la materia inorgánica, en la medida en que la vida es algo cualitativamente diferente de la materia inorgánica pero igualmente física, la mente emerge del cuerpo, en la medida en que la mente es una entidad física. En cambio, los emergentistas fuertes sostendrían que lo que emerge es ontológicamente diferente de lo físico o que, en todo caso, tiene propiedades que no son físicas. Uno de los autores que ha discutido con más detalle las diversas formas de emergentismo es Searle (1983, 1992, 1995).

9.1.4. El monismo de aspecto dual

Esta posición, también llamada «teoría de los dos puntos de vista» o «teoría de los dos aspectos» se remonta hasta el monismo de Baruch Spinoza (2002), quien pensaba que mente y cuerpo son la misma realidad bajo denominaciones diferentes, de la misma manera que Dios y la naturaleza lo son. Es posible, sin embargo, que una versión incluso más antigua se pueda encontrar en Aristóteles, quien pensaba que

la substancia es indivisible en tanto compuesto *hylemórfico* de materia y forma. Bajo cierta interpretación de Aristóteles, se podría decir que forma y materia son dos maneras diferentes de describir la substancia y no dos substancias diferentes.

Versiones más elaboradas sostienen que, en última instancia, «mente» y «cuerpo» son términos que pertenecen a dos descripciones o perspectivas diferentes de la realidad, la cual no es ni física ni mental, pues puede ser descrita de una u otra manera, o de ambas. Así lo mental y lo físico no serían substancias ontológicamente diferentes sino descripciones o vocabularios diferentes con los que podemos explicar distintos aspectos de la realidad, la cual, según el tipo de monismo que se asuma, es ontológicamente neutra —como sostiene el monismo neutral— o es por lo menos, en algunos lugares, como el cerebro humano, al mismo tiempo física y mental, pues tiene ambos tipos de propiedades.

Una forma de monismo de aspecto dual sería la posición del segundo Wittgenstein (1988), quien pensaba que el lenguaje psicológico es un juego de lenguaje diferente al físico, sin que eso signifique que ambos juegos de lenguaje refieran a objetos ontológicamente diferentes. Richard Rorty, en su célebre *La filosofía y el espejo de la naturaleza* (1989) también pensaba que la mente es solo un modo de hablar o un juego de lenguaje, pero no defendía una ontología monista, simplemente no defendía ninguna ontología. Por esa razón, su posición fue considerada en cierto momento eliminativista, aunque después él afirmó que se trataba más bien de un fisicalismo no reductivista.

En la filosofía del siglo XX, el monismo de aspecto dual sostuvo al inicio que los estados mentales tienen un aspecto físico externo, observable públicamente y objeto de experiencia científica, y también un aspecto interno conocido por introspección que sería propiamente la experiencia fenoménica. Esta posición se encuentra, por ejemplo, en Thomas Nagel (1996, 1998), aunque se puede encontrar una versión algo más antigua, pero quizá más clara, en Peter Strawson (1959, 1985, 1995). Este autor sostiene que la persona tiene dos aspectos o propiedades, una física y la otra mental, ambas irreducibles en la otra. De esta manera, la mente no sería un epifenómeno o propiedad emergente del cerebro sino un aspecto de la persona en su totalidad, y las descripciones física y psicológica serían irreducibles una respecto de la otra[3].

9.1.5. El dualismo de propiedades

Esta posición es muy cercana al monismo de aspecto dual aunque con sutiles diferencias. Ambas concepciones son monistas y no reduccionistas, pero mientras la primera sostiene que la realidad es al mismo tiempo física y mental —por lo menos en determinados lugares de ella, como los cerebros— o es ontológicamente neutral,

[3] Otras versiones de la teoría de aspecto dual se pueden encontrar en Hampshire, 1971 y en O'Shaughnessy, 1980.

el dualismo de propiedades sostiene que la realidad es básicamente física, pero algunos sistemas físicos complejos también tienen propiedades mentales que no son reducibles a las físicas. De esta manera, el dualismo de propiedades es compatible con el emergentismo fuerte. David Chalmers (1996, 2010, 2012) ha presentado una reciente y poderosa defensa de esta posición al sostener que la conciencia está conformada por propiedades no físicas.

9.1.6. Las concepciones de la mente encarnada —*embodied mind*— y de la mente extendida —*extended mind*—

David Chalmers y Andy Clark (1999, 2008, 2016) han discutido la posibilidad de que la mente no se limite al cerebro. Si uno entiende la mente como todo aquello que permite la cognición y la vida afectiva, no habría razón para ubicar la mente en el cerebro y no en todo el cuerpo[4].

Por otra parte, en tanto uno puede pensar o sentir a través o gracias a artefactos que no son parte de su cuerpo —por ejemplo una computadora donde uno almacena gran parte de su memoria— no habría razón para sostener que los límites de lo mental son los límites de lo cerebral. Esto conduciría a que se difuminen las fronteras de lo que llamamos la mente. Aunque esta posición es actualmente defendida sobre todo por Chalmers y Clark, uno puede encontrar sus raíces en William James (1983), quien en 1890 propuso que el yo no se reduce a la suma de estados mentales, sino que incluye los vínculos afectivos e incluso los objetos físicos con los cuales uno está de diversas formas comprometido.

9.1.7. Pansiquismo

Esta posición sostiene que «los componentes básicos de la materia tienen propiedades mentales» (Nagel, 1996, p. 181), de manera que la naturaleza en su totalidad podría tener algún grado de conciencia. Por extraña que parezca, esta posición ha sido defendida en varias ocasiones a lo largo de la filosofía y recientemente ha generado nuevo debate gracias a la obra de filósofos de la mente como Chalmers (2015).

[4] Como sostiene Godfrey-Smith (2017), el pulpo sería el candidato más claro para evidenciar una mente encarnada, porque tiene neuronas prácticamente en todo su cuerpo, especialmente en los tentáculos, lo que le permite tener tacto, olfato y gusto a través de ellos. Más aún, los pulpos tienen un número de neuronas inusualmente alto para los moluscos, que llega aproximadamente a 500 000 —el mismo número que tienen los perros y el doble de los gatos— lo que los convierte en una de las especies con más alto número de neuronas en proporción al tamaño de su cuerpo. Aunque el cerebro humano controla todo el cuerpo, está ubicado solo en la cabeza y tiene 86 000 0000 neuronas aproximadamente. Cada neurona tiene un promedio de 7000 conexiones sinápticas. A lo largo del desarrollo del niño se reduce significativamente el número de conexiones en lo que se llama «poda cerebral», pues ciertas conexiones se fortalecen y otras se debilitan hasta desaparecer, según la actividad que tengan.

9.2. Algunas cuestiones ontológicas

Como resultará patente, con frecuencia las distintas posiciones en torno del problema mente/cuerpo se superponen entre sí y generan relaciones de compatibilidad e incompatibilidad confusas y complejas. Adicionalmente, con frecuencia no es claro lo que en estas discusiones se entiende por «existir». Es evidente que el monista fisicalista piensa que el único sentido de existencia es el de las entidades que o pueden ser descritas por las ciencias físicas o son ontológicamente reducibles a entidades que pueden ser descritas por las ciencias físicas, y que el dualista ontológico cree que hay por lo menos otro sentido posible. Pero no queda claro qué entiende por «existir» el dualista de propiedades, y lo mismo podría decirse del monista de aspecto dual, del emergentista y de algunas modalidades de funcionalista.

Algunas versiones del monismo de aspecto dual sostienen que, en última instancia, «mente» y «cuerpo» son términos que pertenecen a dos descripciones diferentes de la realidad, la cual no es ni física ni mental y puede ser descrita de una u otra manera. Se puede entender esta afirmación de por lo menos tres maneras diferentes. En una primera lectura, el monismo de aspecto dual se acercaría al llamado «monismo neutral» de James (1996 [1912]) y Russell (1914a, 1914b). Según esta posición, lo mental y lo físico no serían substancias ontológicamente diferentes sino descripciones o vocabularios diferentes con los que podemos explicar diferentes aspectos de la realidad, la cual sería en sí misma ontológicamente neutra. En esa misma línea, de inspiración kantiana, se encuentra Thomas Nagel (1996, 1998). El problema con esta posición es que, por una parte, es una forma de noumenismo que sostiene que es imposible conocer la realidad en sí misma, pero, de otro lado, afirma que de alguna misteriosa manera sí podemos saber que la realidad es una y ontológicamente neutra, es decir, presupone un acceso privilegiado a la realidad desde el cual se afirma que no se puede saber nada de ella excepto que es una y no doble, lo que es contradictorio[5].

Una segunda lectura del monismo de aspecto dual sostendría, siguiendo la crítica davidsoniana al llamado tercer dogma del empirismo que, en tanto no es posible separar la descripción de lo descrito, la realidad descrita por los vocabularios físico y mental es simultáneamente física y mental, pues tiene ambos aspectos al mismo tiempo. La imagen wittgensteiniana del pato/conejo puede ilustrar esta posición, ya que el mismo dibujo es un pato y es un conejo, aunque también puede

[5] Solms y Turnbull, 2002, han adoptado esta interpretación noumenista del monismo de aspecto dual en su proyecto de construir una ciencia neuropsiconalítica, es decir, de integrar los desarrollos de las neurociencias con algunas tesis centrales del psicoanálisis para darles un fundamento empírico. El proyecto mismo es interesante y valioso, pero los autores no abordan los problemas filosóficos asociados a una lectura noumenista del monismo de aspecto dual.

ser una línea negra sobre un fondo blanco. El objeto es todas esas cosas al mismo tiempo en diferentes sentidos. Así algunos aspectos de la realidad —básicamente los cerebros humanos— serían al mismo tiempo físicos y mentales, porque tendrían ambos aspectos.

Según una tercera lectura del monismo de aspecto dual, de corte instrumentalista, podría sostenerse que no hay un concepto de existencia, sino varios, pues existen y son parte de la realidad todos los objetos con los que nuestros sistemas de creencias más explicativos están ontológicamente comprometidos. Así es como podemos afirmar que existen volcanes, neutrinos, adverbios, clases sociales, creencias y dolores de cabeza. Seguramente también existen los agujeros negros y quizá las supercuerdas. Probablemente también existe el inconsciente dinámico freudiano, si es necesario para explicar formas de comportamiento humano de otra manera inexplicable. Es verdad que estos objetos solo podrían existir asociados a objetos físicos, por ejemplo, los adverbios a hablantes, las clases sociales a grupos de individuos, y el inconsciente a cerebros, con lo cual el concepto físico de existencia tiene cierta prioridad respecto de los otros. Pero lo importante es que habría que admitir que la realidad se compone de entidades o propiedades que en sí mismas no son físicas, aunque estén asociadas a objetos físicos, como los adverbios y las clases sociales, en tanto sobrevienen a los objetos físicos. Si aceptamos todo lo anterior, tendremos que admitir que también existen las mentes, en caso que tengamos que postularlas para explicar objetos de otra manera inexplicables, como las experiencias fenoménicas o los *qualia*.

En esta misma línea de reflexión, también sería útil volver a discutir la posible distinción entre materialismo y fisicalismo. Si bien muchos autores los usan intercambiablemente, probablemente «fisicalismo» sea un término preferible a «materialismo» porque incluye como parte de la realidad física a objetos que podría sostenerse no son materiales. Es decir, podría defenderse que cuando hablamos de materia nos referimos a estructuras complejas y no a sus componentes mínimos ni, eventualmente, a las propiedades emergentes que de ella surgen. En cambio, cuando hablamos de «realidad física» nos referimos a todo lo que hipotéticamente podría ser descrito o postulado por nuestras teorías físicas más sofisticadas en condiciones ideales. Así, el concepto de realidad física parecería más amplio que el de materia y lo contendría. Pero si seguimos esa línea de argumentación, dado que los conceptos de adverbio o clase social no pertenecen a ninguna teoría física, pero sí estamos dispuestos a sostener que existen adverbios y clases sociales porque proceden de teorías explicativas no físicas que están ontológicamente comprometidas con esos objetos, no habría razón para rechazar la existencia de mentes en pie de igualdad con la existencia de cuerpos físicos. Esto parecería conducir a un pluralismo ontológico en general y, en particular, a un dualismo ontológico en lo relativo al problema mente-cuerpo.

Así, esta discusión parecería sugerir que el monismo de aspecto dual tiene dos caminos: o colapsa en alguna forma de monismo emergentista o deja de ser monismo para ser un pluralismo ontológico.

Sin embargo, el monista de aspecto dual todavía puede argumentar en su favor. En primer lugar, podría decir que lo que existe es todas aquellas entidades con las que están hipotéticamente comprometidas las ciencias físicas en condiciones ideales y las propiedades emergentes de esas entidades. De esa manera, no solo existen electrones y árboles, sino también mentes, clases sociales y adverbios, todos ellos con los mismos derechos ontológicos, aunque sobrevinientes de objetos naturales. Es decir, existen las mentes y los adverbios, pero porque existen los cerebros y las comunidades de hablantes.

Otro camino que podría tomar el monista de aspecto dual es sostener que si bien todos los objetos de conocimiento podrían ser descritos en términos de relaciones causales físicas, una explicación que use nociones intencionales como «creencia» y «deseo» podría ser más potente en la explicación. Con esto ocurriría algo semejante a cuando confrontamos una teoría geocentrista y una heliocentrista del sistema solar. Ambas explicaciones son igualmente válidas, en tanto ambas pueden dar cuenta del sistema solar en términos causales y nomológicos, pero la teoría heliocentrista tiene más virtudes epistémicas que su rival, sin que eso signifique que sea la única explicación válida del sistema solar. De igual manera, podría decirse que el vocabulario intencional es más explicativo del comportamiento humano que un vocabulario puramente físico, sin que se asuma que uno de ellos es la única explicación válida. Si admitimos todo eso, podríamos aceptar la necesidad de usar conceptos psicológicos como «creencia» y «deseo» sin tener que aceptar que tienen una existencia diferente a los procesos neuronales de los que son descripciones diferentes pero ontológicamente idénticos. Un problema con este último argumento, sin embargo, es que el geocentrismo es reducible al heliocentrismo, pero el monista de aspecto dual desea que la explicación intencional no lo sea a una explicación física.

Un problema no aclarado por las versiones tempranas del monismo de aspecto dual es el de las relaciones causales entre ambos aspectos. Como vimos en el primer capítulo, para abordarlo Davidson (1980) desarrolló el monismo anómalo, el cual sostiene que mente y cuerpo son conceptos que pertenecen a dos descripciones diferentes, la psicológica y la física, respectivamente. Ambas descripciones lo son de la persona, que es tanto física como mental, es decir, que puede ser descrita como física pero también como mental, sin que ninguna de esas descripciones sea reducible a la otra. Sin embargo, para esta posición, las relaciones causales son independientes de estas descripciones. En otras palabras, no hay interacción entre la descripción física y la mental, de la misma manera que no hay interacción entre un mapa geográfico

y uno político de un país, aunque ambos mapas lo son del mismo país y las relaciones causales del mismo objeto, así como las regularidades de esas relaciones son independientes de ambas descripciones. Ocurre, sin embargo, que cuando describimos esas relaciones causales, normalmente lo hacemos bajo la descripción física y así podemos hablar de leyes físicas. Pero las leyes físicas no son sino descripciones lingüísticas de ciertas regularidades entre relaciones causales. En otras palabras, solo hay una realidad y esta puede ser descrita mediante el vocabulario físico o el psicológico, los cuales tienen criterios y propósitos explicativos diferentes. Dado que se trata de dos vocabularios y no de dos realidades, no interactúan causalmente, es decir, no hay interacción causal entre la mente y el cuerpo. Las relaciones causales se dan en la realidad, independientemente de si se usa la descripción física o la mental. El caso es que esas relaciones causales son descritas posteriormente por el vocabulario psicológico, con lo que tenemos regularidades psicológicas, o por el vocabulario físico, con lo cual tenemos regularidades físicas, que pueden dar lugar a lo que solemos llamar «leyes físicas».

La tesis de Davidson se llama monismo anómalo porque no es una forma de dualismo ni tampoco de fisicalismo reductivista, y sostiene que solo hay leyes bajo la descripción física, de manera que no hay leyes psicológicas ni tampoco psicofísicas. Estas ideas se encuentran en *Essays on Actions and Events* (1980), especialmente en los artículos «Actions, Reasons and Causes» y «Mental Events».

Sin embargo, como vimos en el primer capítulo, esta posición ha sido objetada por varios autores, entre otros Honderlich (1982), Sosa (1984), Kim (1984, 1989) y Moya (1990), entre otros. Ellos sostienen que esta posición colapsa en una forma de epifenomenalismo en la que la eficiencia causal radica en las propiedades físicas del evento y no en las propiedades mentales, de manera que estas últimas resultan inútiles y redundantes.

Una posible defensa de Davidson sería que no hay prioridad causal de lo físico frente a lo psicológico, pues lo que llamamos «regularidades causales físicas» son simplemente una descripción diferente de lo que llamamos «regularidades causales psicológicas».

Algunos también han sospechado que en esta posición se esconde cierto noumenismo kantiano. Una salida a ese problema puede encontrarse en un giro pragmatista en el que ambos vocabularios son descripciones de la realidad, la cual sería al mismo tiempo física y mental. Este sería un giro pragmatista porque está en la línea del pluralismo de William James. Si tomamos esa dirección, habría que decir que hay relaciones causales físicas —las descritas en ese vocabulario—, relaciones causales psicológicas —las descritas en ese vocabulario— y relaciones causales psicofísicas —las descritas en ambos vocabularios—. Lo que no habría es leyes psicofísicas ni leyes psicológicas, es decir, regularidades con alto grado de determinación al interior de esos vocabularios.

Una objeción de principio contra el monismo de aspecto dual es que parece presuponer que conoce lo que realmente existe, al afirmar que esta realidad no es ni física ni mental; ya sea porque sabe lo que es y sabe que no es ni física ni mental, o porque no sabe lo que es, aunque paradójicamente sí sabe que no es ni física ni mental. Lo segundo parece una forma de noumenismo escéptico kantiano; lo primero es simplemente una forma de dogmatismo. Sin embargo, el monismo de aspecto dual no tiene necesariamente que colapsar en una de esas posiciones. Esta concepción puede ser asociada a una forma de pluralismo epistemológico en la que se sostiene que la realidad puede ser descrita de múltiples formas, cada una con objetivos y criterios de validez diferentes. Esta forma de pragmatismo implica que no hay una descripción que determine la naturaleza última de la realidad, pero varias descripciones pueden ser simultáneamente válidas —o incluso verdaderas— en tanto explican aspectos de la realidad diferentes entre sí.

El monismo anómalo no es incompatible con el funcionalismo, en tanto ambas posiciones son formas de fisicalismo no reductivista. Lo central en el funcionalismo sería asumir que los estados mentales —y, por tanto, lo que llamamos mente— son funciones del cerebro, no el cerebro mismo, pero tampoco, en sentido estricto, son epifenómenos del cerebro. El funcionalismo es no reductivista porque considera que las explicaciones psicológicas no son reducibles a explicaciones físicas. Tampoco parece que el monismo de aspecto dual sea incompatible con una forma débil de emergentismo, siempre que uno tenga claro que la emergencia debe verse en el plano temporal. Es decir uno podría considerar que la materia inerte carecía en cierto momento, digamos antes de la aparición de la vida, de propiedades biológicas. En cierto momento de complejidad de la materia emergieron las propiedades biológicas como nuevas propiedades emergentes de la materia. Análogamente en cierto momento la materia con propiedades biológicas, es decir los organismos vivos, produjeron, como rasgos emergentes, propiedades sociales, grupales, políticas y, en algún momento, incluso cognitivas y conscientes. Así, aquellos rasgos que consideramos característicos de lo mental, como la conciencia o la intencionalidad, habrían sido propiedades emergentes a partir de sistemas complejos no mentales como, por ejemplo, los cerebros. La complejidad cerebral habría dado lugar a nuevos sistemas complejos dotados de propiedades nuevas, en este caso psicológicas. La emergencia de propiedades mentales o psicológicas se habría producido tanto en la dimensión filogenética —es decir en la evolución de la especie— como en la dimensión ontogenética —el desarrollo del niño—. Pero nada de esto parece, en principio, incompatible con la tesis de la superveniencia y del monismo de aspecto dual. No es impensable que algunos sectores de la realidad tengan al mismo tiempo propiedades físicas y psicológicas, en los que estas no están en pleno nivel de igualdad porque

las propiedades psicológicas emergieron de sistemas complejos que ya tenían propiedades físicas. Análogamente, tampoco es absurdo afirmar que algunos sectores de la realidad tengan al mismo tiempo propiedades físicas y sociales, o físicas, sociales y psicológicas.

La idea, entonces, es que lo que existe es eventos estructurados causalmente entre sí, de manera previa e independiente de las múltiples descripciones que podamos hacer de ellas. Podemos describir esos eventos de muchas maneras, según los objetivos explicativos que tengamos. Podemos describirlos, por ejemplo, usando el lenguaje físico, con lo que tendremos una realidad física conformada por partículas subatómicas, átomos, moléculas, objetos espaciotemporales macroscópicos, etcétera.

También podemos describir esos mismos eventos con un lenguaje intencional, ya sea individual o colectivo. Según el primero diremos, por ejemplo, que Boris cree o desea algo. Así tendremos acciones y estados mentales. Si usamos un lenguaje intencional colectivo tendremos comportamiento social basado en relaciones de poder y autoridad, pero también podremos describir procesos sociales, históricos, lenguas, etcétera. Tanto los electrones, los árboles, las clases sociales como los adverbios y la caída de Constantinopla en el año 1453 son eventos que son parte de la realidad, aunque descritos de maneras diferentes para explicar o lograr finalidades distintas.

La mayor parte de eventos del universo solo tienen propiedades físicas y, por tanto, solo pueden ser descritos mediante un lenguaje físico. Pero algunos pocos eventos en el universo, además de tener propiedades físicas tienen propiedades intencionales —es decir, tienen la capacidad de representar algo diferente de ellos mismos y hasta quizá algún grado de agencia—, lo que permite que se les pueda describir también mediante un lenguaje psicológico, sociológico, histórico, etcétera. Estos eventos son tanto físicos como mentales, pues tienen propiedades físicas y mentales. Sus propiedades físicas son semejantes a las de cualquier otro evento del universo y están gobernadas por las mismas leyes naturales. En cambio, sus propiedades mentales han emergido a partir de sistemas físicos complejos de los cuales supervienen. Esto significa que sin aquellas propiedades físicas no habrían podido emerger las propiedades mentales y que todo cambio físico producirá un cambio mental, pero no al revés.

El punto, entonces, es que los eventos tienen el estatuto ontológico de sus propiedades y que puede haber eventos que sean, al mismo tiempo, físicos y mentales, lo que da lugar a ambos tipos de descripciones, de la misma manera como el conocido dibujo de Wittgenstein del pato-conejo es tanto un pato como un conejo, según cómo queramos o podamos verlo. Análogamente, estos eventos serán descritos de una y otra forma según los objetivos que uno tenga, ya sea, por ejemplo, curar una gastritis o hacer psicoterapia. A esa ontología podemos llamar «dualismo de propiedades»,

pero claramente no es un dualismo de sustancias. Podríamos también denominarla «monismo de sustancias» en tanto hablamos de un solo tipo de eventos con dos tipos diferentes de propiedades.

Este es un tema complejo que ha dado lugar a mucha discusión y este no es el lugar para revisarla, pues el objetivo prioritario de este libro es aclarar la naturaleza de la comprensión y no la ontología de lo mental. Sin embargo, sí me parece importante decir que es el evento como tal, con sus propiedades físicas y mentales, el que tiene rol causal para generar otros eventos, independientemente de las maneras como lo describamos. No obstante, hay que admitir que Davidson es poco claro en este punto. Por momentos se expresa como si la descripción intencional no añadiera nada al evento físico (1993, p. 12), lo que permite la objeción de que la descripción intencional es causalmente ineficiente. Por eso creo que es necesario incorporar al dualismo de propiedades la idea de que las propiedades mentales emergen en un sentido débil a partir de las propiedades físicas. Estos eventos, que incorporan tanto propiedades físicas como las propiedades mentales emergentes, son los que causan otros eventos y tendrían un rol causal diferente si no incorporaran esas propiedades mentales. Una descripción puede o no tomar en consideración las propiedades mentales y, por tanto, puede o no describir al evento como solo físico o como también intencional, pero el evento incorporará ciertas propiedades y no otras, independientemente de la descripción. Las propiedades mentales son, por tanto, reales, tanto como las físicas, aunque dependan y sean supervinientes de estas últimas. Sin embargo, como ya he señalado en varias ocasiones, estos eventos, con sus propiedades, pueden ser descritos de múltiples formas y dan lugar a cierto margen de indeterminación de la interpretación. Y es en esas situaciones interpretativas que se constituye el contenido del estado mental, el cual puede variar en las distintas interpretaciones que se generen. Así, por ejemplo, tres intérpretes diferentes podrían asignar al mismo agente tres sistemas de interpretación diferentes y cada uno de estos sistemas atribuirá distintos estados mentales con diferentes contenidos proposicionales. Esto conducirá a que para la intérprete 1 el agente tenga las creencias p, q, r; mientras que, para la intérprete 2, él tenga las creencias s, t, u; y para la intérprete 3 él tenga las creencias v, w, x. Los contenidos que cada una de ellas atribuya al agente serán diferentes y por eso habrá un grado de indeterminación de la interpretación. Sin embargo, los estados mentales serán los mismos solo que descritos de diferente manera, como —para usar el célebre ejemplo davidsoniano— si midiéramos la temperatura de una habitación usando grados Celsius o grados Fahrenheit. La temperatura es la misma aunque la medida sea diferente. Análogamente, los eventos físicos y mentales serán los mismos, aunque sean descritos, medidos e interpretados de diferente manera.

Ahora bien, no sería preciso decir que las propiedades mentales emergieron de las propiedades físicas, sino que las propiedades mentales emergieron de sistemas complejos que tenían ya propiedades físicas. Pero todo esto ocurre en la dimensión temporal, ya sea filogenética u ontogenética. En la dimensión conceptual de las relaciones entre mente y cuerpo, el enfoque debe ser otro y eso es lo que ocurre con los fenómenos psicosomáticos.

9.3. ¿Cómo son posibles los fenómenos psicosomáticos?

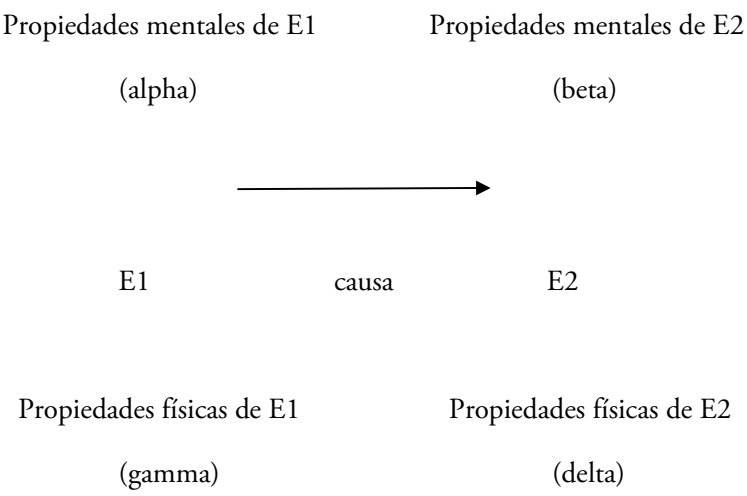
Intuitivamente sabemos que hay relaciones causales entre el cuerpo y la mente. Esto es, fenómenos psicológicos generan efectos corporales —una depresión puede causar una gastritis, por ejemplo—,y fenómenos físicos generan efectos psicológicos —por ejemplo, un desbalance químico puede generar una depresión—. Ahora bien, solo puede haber interacción causal entre objetos que son diferentes entre sí, porque un objeto no interactúa causalmente consigo mismo. Se sigue que si aceptamos la interacción causal entre la mente y el cuerpo, tendremos que aceptar alguna forma de dualismo de sustancias. Pero si mente y cuerpo son ontológicamente diferentes, tampoco pueden interactuar causalmente. Esto nos conduce a un entrampamiento porque, como hemos visto, el dualismo ontológico tiene demasiados problemas conceptuales sin resolver, aunque la existencia de fenómenos psicosomáticos nos parece intuitivamente cierta. Así, si abandonamos el dualismo de sustancias y postulamos alguna forma de monismo, parecería que tendríamos que abandonar también la posibilidad de relaciones causales entre lo mental y lo corporal, lo que parece intuitivamente inaceptable. Pero si aceptamos el dualismo de substancias tampoco podemos explicar que haya interacción causal entre entidades ontológicamente diferentes. ¿Cómo podemos resolver esta aporía? Me propongo, en lo que sigue, sugerir una manera.

Si aceptamos una forma de monismo de aspecto dual o un dualismo de propiedades, tendremos que decir que los eventos mentales son idénticos a los eventos físicos. Esto es, un estado o evento mental cualquiera —por ejemplo, una creencia, una emoción o un dolor— es idéntico a una determinada configuración neuronal del cerebro o, en general, a un determinado estado del cuerpo. No es que el estado mental sea causado por la configuración neuronal ni que la configuración neuronal sea efecto de un estado mental, es más bien que el estado mental es idéntico al estado físico. Ahora bien, hay relaciones causales entre los eventos independientemente de cómo sean estos descritos. Ciertos eventos causan otros eventos, que pueden ser solo físicos o también mentales. Pero si son mentales deben ser también físicos. Dicho de otra manera, si un evento es solo físico y no mental —el colesterol alto,

por ejemplo— puede causar solo eventos físicos, pero también podría causar eventos mentales. Ahora bien, si causa eventos mentales también causa eventos físicos, porque los eventos mentales son idénticos a los eventos físicos dado que no hay eventos mentales que no sean físicos, aunque ciertamente hay eventos físicos que no son mentales. De igual manera, si un evento es mental —y por tanto también físico, por ejemplo, la depresión— puede causar solo eventos físicos —el estado corporal, tanto estomacal como neural, propio de una persona con gastritis— o eventos físicos que también son mentales —el estado cerebral propio de una depresión, por ejemplo, una reducción en los niveles de serotonina—.

Vistas así las cosas, en sentido estricto no hay interacciones causales entre mente y cuerpo; lo que hay es interacciones causales entre eventos físicos que solo pueden ser descritos como físicos y eventos físicos que también pueden ser descritos como mentales. La relación causal se da entre los eventos, independientemente de cómo sean estos descritos, pero es posible formular la relación causal como una relación entre la mente y el cuerpo, siempre que uno describa este evento no solo como físico sino también como mental.

El siguiente esquema podría graficar la situación. Imaginemos dos eventos, E1 y E2, y supongamos que E1 causa E2. En la práctica, toda relación causal es multicausal, es decir, no es un evento aislado el que causa a otro, sino un conjunto de eventos el que causa por lo menos otro evento, pero por claridad hablaremos de dos eventos. Cada uno de ellos puede ser descrito ya sea solo como físico —por ejemplo, la circulación de la sangre— o como físico y mental —por ejemplo, un dolor de muelas—. Así también, cada una de estas descripciones lo es de un aspecto del evento, sea el aspecto físico o el mental. Tenemos, entonces, el siguiente esquema:



Llamemos «alpha» a las propiedades mentales de E1, «beta» a las propiedades mentales de E2, «gamma» a las propiedades físicas de E1 y «delta» a las propiedades físicas de E2. Donde:

- Alpha: la experiencia fenoménica de la depresión, es decir, la sensación de tristeza.
- Beta: la experiencia fenoménica de un intenso dolor de estómago.
- Gamma: el estado corporal, especialmente cerebral, propio de una persona deprimida, por ejemplo, una reducción de los niveles de serotonina en el cerebro.
- Delta: el estado corporal, tanto estomacal como cerebral, propio de una persona que padece de gastritis.

Es claro que podemos describir las relaciones causales de E1 y E2 de múltiples maneras. Podemos decir que alpha —la experiencia fenoménica psíquica de la tristeza— causó beta —el dolor de estómago—. O que alpha causó delta. También podemos decir que gamma causó a beta o que gamma causó a delta. Ciertamente podemos decir, lo que sería más correcto, que alpha y gamma causaron a beta y delta, es decir, que la depresión —que es idéntica a un tipo de configuración cerebral, por ejemplo, de disminución de serotonina— causó el dolor de estómago y la gastritis. Coloquialmente diríamos que la gastritis causó el dolor de estómago, pero sería más apropiado decir que el dolor de estómago es una manifestación o una expresión de la gastritis, más que su efecto.

En cualquier caso se verá que, en sentido estricto, no hay interacción causal entre la mente y el cuerpo, solo hay interacción causal entre eventos que pueden ser descritos como físicos y eventos que también pueden ser descritos como mentales. En sentido estricto, la causalidad es una relación entre eventos independientemente de cómo se los describa. Pero la explicación física y la interpretación intencional son formas de describir los eventos para satisfacer objetivos y finalidades concretas, de manera que podemos elegir cierta descripción de un evento para afirmar que causó otro evento al cual nos referimos bajo cierta descripción de él. Por ello sería correcto decir que hay fenómenos psicosomáticos, que una depresión puede producir dolor de estómago o gastritis, y que la disminución de los niveles de serotonina en el cerebro puede causar dolor de estómago o gastritis.

Por otra parte, resultará claro que uno podría intervenir en alpha —por ejemplo, con psicoterapia— o en gamma —con fármacos que disminuyan la recaptación de la serotonina— y generar efectos en beta —el dolor— o en delta—aliviando la gastritis—. Sin duda, una doble intervención —en alpha y gamma— resultaría más eficiente para generar resultados en beta y delta.

En el ejemplo anterior he analizado una relación causal entre lo que llamamos psíquico —la experiencia fenoménica de la tristeza—, pero que también es un fenómeno físico —la configuración del cerebro de una persona deprimida—, y un efecto que clásicamente sería físico —una gastritis—, pero que en este caso también tiene una dimensión psíquica —la experiencia fenoménica del dolor—. Aunque también sería posible imaginar relaciones causales entre eventos únicamente físicos —un infarto cerebral— y efectos que son tanto físicos como psíquicos —la perdida de la memoria, que es también un daño en ciertos tejidos del cerebro—. Igualmente, uno podría imaginar que la depresión cause angustia, con lo que estamos describiendo la relación causal entre el aspecto psíquico de un evento E1 y el aspecto psíquico de otro evento E3. El punto, sin embargo, es que independientemente de la manera en que describamos al evento —si psíquico o físico—, toda descripción psíquica de un evento tiene que estar asociada a una descripción física de ese evento, mas no al revés. Hay fenómenos físicos que no tienen aspecto psíquico —por ejemplo, la circulación de la sangre—, pero no hay eventos psíquicos que carezcan de un aspecto físico.

Es posible decir que el cerebro es básicamente un conjunto de conjuntos de redes neuronales que son, asimismo, procesos físicos gobernados por relaciones causales que instancian regularidades. Estas redes tienen propiedades físicas, predecibles y explicables nomológicamente. Sin embargo, ellas generan propiedades emergentes que sobrevienen a las propiedades físicas, pero que no se reducen a ellas. Tales propiedades emergentes tampoco pueden ser predecibles ni explicables nomológicamente. Los diversos procesos psíquicos constituyen estas propiedades emergentes como, por ejemplo, la conciencia o la intencionalidad. Así, entender la mente como un sistema complejo de propiedades emergentes es compatible con una forma de dualismo de propiedades, aunque no con un fisicalismo reductivista. Adicionalmente, la mente en tanto sistema complejo de propiedades emergentes se constituye en la interacción y el intercambio con otras mentes en el contexto de sistemas aún más complejos, como son los sistemas sociales. Es precisamente en esas interacciones intersubjetivas que se constituyen y estructuran los estados mentales y las mentes en tanto sistemas de estados mentales. Pero ahora deberemos abordar una de las maneras en que los estados mentales están estructurados, que es la racionalidad.

Cuarte parte
La racionalidad y sus límites

Capítulo diez
¿Qué es la racionalidad?

10.1. Interpretación y racionalidad

Comprender a alguien supone encontrar la racionalidad de sus estados mentales y su comportamiento. Encontrar a alguien sistemáticamente irracional sería lo mismo que no poder entenderlo, es decir, no hallar sino ininteligibilidad. Con frecuencia atribuimos irracionalidad a los agentes, pero esto debe hacerse sobre un fondo mayor de racionalidad para que la irracionalidad sea inteligible, es decir, razonable. Es más, cuando atribuimos irracionalidad a alguien lo hacemos precisamente para encontrarlo *grosso modo* racional, es decir inteligible. Pero los conceptos de racionalidad e irracionalidad tienen una larga y compleja carga filosófica que es necesario desenredar. En esta parte del libro trataré de redondear mis tesis acerca de la comprensión, atando los cabos que hayan quedado sueltos en los capítulos previos.

La palabra castellana «razón» procede de latín *ratus* que, al ser el participio pasado de *reor,* significa ordenar, articular o relacionar, fundamentalmente pensamientos e ideas. Los latinos utilizaron estas palabras y sus derivadas para traducir los términos griegos logos y *diánoia* que tienen un significado semejante. Como se sabe, el verbo en infinitivo *legein* significa recoger, ordenar, articular y, por extensión, narrar o dar cuenta y justificación de algo. Es este último sentido el que terminó primando en el uso filosófico, con lo que la razón termina siendo no solo la justificación de algo sino también la facultad de aquella persona que está en condiciones de justificar algo. Este uso lingüístico está vinculado a dos presupuestos que me parecen objetables. En primer lugar, el supuesto de que la razón de algo es solo un tipo de justificación lógica, fundamentalmente de creencias o procesos cognitivos. En segundo lugar, el supuesto de que la razón es una facultad o propiedad monádica que tienen algunas criaturas. Este segundo presupuesto condujo, en la tradición filosófica, a interrogar por la naturaleza de esa facultad que, de ser universal, sería la característica distintiva

de los seres humanos y, de ser particular, sería el rasgo específico de algunos tipos de cultura. Así, más adelante surgiría una sospechosa pregunta que en algún momento se convirtió en tema clásico de discusión: ¿Hay una racionalidad universal que compartamos todos los seres humanos o existen diversas racionalidades, así como existen distintas culturas?

En este capítulo deseo discutir ambos supuestos. Sostendré, en primer lugar, que la expresión «razón» no debe aludir únicamente a un tipo de justificación lógica de contenidos cognitivos sino, más comprehensivamente, a la articulación entre creencias, deseos y acciones, con lo cual esta debe entenderse fundamentalmente en su dimensión práctica, pues el objetivo último de las creencias y los deseos es hacer posible la acción. Mi interés es sugerir una manera de superar la dicotomía moderna entre la teoría y la práctica, y mostrar el carácter esencialmente práctico de la creencia y de la teoría misma. En segundo lugar, afirmaré que la racionalidad no es una facultad ni una propiedad monádica de individuos, sino una propiedad relacional entre intérpretes y el mundo que comparten, que emerge en situaciones comunicativas específicas. A partir de aquí se verá que no tiene sentido preguntar si hay una racionalidad o varias, dado que no se trata de una facultad que pueda o no ser compartida. Finalmente, sostendré que la concepción de racionalidad que estoy defendiendo —que es parte de una teoría de la interpretación— tiene consecuencias en torno al relativismo conceptual. Con el desarrollo de temas ya abordados en el tercer capítulo, intentaré mostrar que una adecuada comprensión de la comunicación humana hace insostenibles las formas clásicas del relativismo conceptual, al mismo tiempo que permite una forma mitigada pero omnipresente de indeterminación de la interpretación.

Como he sostenido a lo largo del libro, el argumento principal se inspira en un desarrollo del principio de caridad realizado por Davidson, pero pretende ir más lejos que él. La idea es que toda comprensión del otro requiere, como condición de posibilidad, verlo como básicamente racional, donde la racionalidad es la articulación consistente entre las creencias, los deseos y las acciones que la intérprete le adscribe en una situación comunicativa dada. La comprensión también requiere asumir que las creencias del otro son en su mayor parte verdaderas y que él tiende a actuar en conformidad con el bien, todo según los propios criterios de la intérprete, que es quien está atribuyendo esos estados mentales al agente. Dado que los estados mentales que la intérprete atribuirá al agente serán aquellos que la intérprete misma cree que tendría si fuera el agente en las circunstancias en que ella cree que él se encuentra, sería imposible para ella interpretar a un agente a quien ella considere básicamente irracional. Asimismo, sería imposible para ella atribuirle a él creencias mayoritariamente falsas o adscribirle propósitos sistemáticamente incompatibles con

sus propias creencias y deseos —los de ella—. A partir de aquí, se puede mostrar cómo la idea misma del relativismo conceptual no se sostiene, si el relativismo conceptual está asociado a alguna doctrina acerca de la inconmensurabilidad entre sistemas de creencias o a la tesis de que no existen criterios objetivos que nos permitan determinar el valor de verdad de nuestras creencias o el valor moral de nuestras acciones. La razón de ello sería que en las interacciones comunicativas se constituyen no solo los significados y los estados mentales, sino también los criterios objetivos mediante los cuales determinamos los valores de verdad de las creencias y los valores morales de las acciones. Pero llegaremos a esa conclusión recién al final de este capítulo.

Como hemos visto, una intérprete podrá comprender a un agente solo si ella está en condiciones de atribuirle un sistema interconectado de estados mentales, significados y acciones que se interdefinen mutuamente, de manera que el contenido de cada uno de estos se determina en su relación con los otros. Por ejemplo, el contenido de una creencia se determina con relación con las otras creencias, los deseos, lo significados y las acciones atribuidos. Así, la intérprete tendrá que atribuir sistemas básicamente consistentes de estados mentales y acciones, porque en el momento que comience a descubrir inconsistencias entre ellos los contenidos mismos de estos elementos comenzarán a desvanecerse. Por ejemplo, si yo entiendo lo que alguien cree porque sé lo que quiere y lo que hace, en el momento en que vea una inconsistencia entre lo que quiere y lo que hace, tendré que revisar mi concepción de lo que cree o desea. La intérprete determinará la inconsistencia entre estos elementos si ella reconoce que estos son incompatibles para la realización de un posible curso de acción y que, por tanto, no podrían culminar en un tipo de comportamiento reconocible por ella como intencional.

Es factible decir, por tanto, que la atribución de racionalidad emerge en una situación comunicativa como una exigencia de consistencia entre los estados mentales y las acciones que la intérprete se ve obligada a adscribir al interpretado para poder hacer inteligible su comportamiento. Pero no es que la racionalidad sea simplemente la coherencia de estos elementos en el agente, pues eso implicaría que los contenidos de los estados mentales y las acciones existen con independencia de una situación interpretativa, y eso es precisamente lo que estoy intentando negar. Los estados mentales surgen en la interacción comunicativa, como una obra creada solidariamente por agente e intérprete. Como he señalado en varios momentos, sostengo que los estados mentales no son propiedades monádicas de individuos, como sostendría una concepción internista de la mente, sino propiedades relacionales triádicas entre individuos y el mundo que comparten. Por ello, no debería uno preguntarse si los estados mentales son construidos o descubiertos, pues son ambas cosas al mismo tiempo. Decir que los estados mentales son atribuidos no debería conducirnos a suponer que la intérprete

los inventa o los impone en el agente. La intérprete atribuye estados mentales en el sentido en que los adscribe al agente, pero estos no proceden de la subjetividad de la intérprete sino del espacio compartido generado en la comunicación intersubjetiva. De igual manera, sería apropiado decir que la racionalidad es construida pero también descubierta en la interacción comunicativa entre intérprete y agente.

Esto es análogo a la constitución de los significados. El significado de una expresión no es una propiedad que esta expresión posea por sí misma, independientemente de una situación interpretativa, sino una atribución que surge en la interacción comunicativa entre hablante e intérprete. Es pues, en sentido estricto, una obra compartida. Si estas situaciones comunicativas se hacen regulares en una comunidad de hablantes, es decir, si se generan prácticas sociales compartidas en las que el significado emerge regularmente entre hablantes e intérpretes, entonces diremos que el significado ha sido constituido en aquella comunidad de hablantes y los diccionarios comenzarán a registrarlo. Análogamente, la racionalidad tampoco es una propiedad del agente como tal, sino una propiedad que emerge en la interacción entre intérprete y agente y que se aplica, en sentido estricto, a cada uno de los elementos de un sistema con relación al todo. Es decir, una creencia, un deseo o una acción pueden ser racionales o irracionales según sean consistentes o no con la red de estados mentales previamente atribuidos al agente. Un aspecto del comportamiento del agente puede ser considerado racional o no con relación a la totalidad de estados mentales y acciones atribuida, pero no tendría sentido decir que una persona es totalmente irracional, porque eso significaría que no estamos en condiciones de atribuirle un sistema articulado de estados mentales, con lo cual no estaríamos reconociendo siquiera conducta intencional. Por otra parte, afirmar de una persona que es plenamente racional es solo una confusa manera de decir que hemos dado inteligibilidad a su comportamiento en función de nuestras propias creencias y deseos, y de aquellos que creemos que tendríamos si fuéramos él.

Lo que puede ser racional o irracional no es un sistema de creencias en su totalidad ni tampoco una persona en sí misma, sino sus estados mentales y acciones, los cuales son atribuidos por una intérprete. De aquí se desprenden dos cosas. Por una parte, la racionalidad no es una propiedad monádica del comportamiento de un individuo ni de su visión del mundo, esto es, no se puede decir de un individuo o de una cosmovisión que sean racionales o no en sí mismos. La racionalidad es, más bien, una propiedad relacional que surge en la interacción comunicativa entre un agente y una intérprete, en relación con el mundo que comparten. En estas circunstancias, la intérprete considerará que alguna acción, creencia o deseo atribuido por ella al agente es racional o no por comparación con el resto de estados mentales y acciones que ella le debe atribuir para hacerlo inteligible—que deben ser básicamente

consistentes entre sí para que tenga sentido decir que ella lo está interpretando— y frente al contexto mayor de la realidad objetiva compartida. Así entonces, la racionalidad es también un atributo que la intérprete adscribe holísticamente a los estados mentales y acciones del agente cuando lo encuentra a él inteligible. Pero, naturalmente, mientras las acciones de un agente pueden ser asumidas como racionales por una determinada intérprete —si se consideran los estados mentales que ella le atribuye—para otra intérprete tales acciones podrán ser irracionales si ella le ha atribuido estados mentales diferentes que no son vistos por ella como consistentes entre sí. Resultará claro, entonces, que la racionalidad tampoco es una propiedad de las acciones o estados mentales del agente en sí mismos ni de la arbitraria interpretación que haga la intérprete, sino de la interacción entre ambas cosas y en confrontación con el mundo; es, en este sentido, que es una propiedad relacional triádica.

Una criatura racional es aquella a la que se le puede atribuir razones como causas y motivaciones de su comportamiento, es decir, una que «tiene la capacidad de reconocer, evaluar y ser movida por razones y, entonces, tener actitudes sensibles a los juicios» (Scanlon, 1998, p. 23). ¿Pero qué es una razón? Una actitud cognitiva que brinda una justificación cuya fuerza muestra que una posición es preferible a otra, no solo para uno sino para cualquiera que tenga la disposición para evaluar las posiciones y razones que están en juego. Cuando uno justifica racionalmente algo, asume que está normativamente impelido a preferir una opción frente a otra, y que cualquiera que tenga la posibilidad de hacer esta evaluación deberá también sentirse impelido a optar por esa posición. En ese sentido, la fuerza de las razones es vinculante. Desde un punto de vista naturalista, la evolución de nuestros cerebros en complejos contextos sociales ha hecho que seamos criaturas que nos sintamos obligadas a dar y recibir razones, y a comportarnos según esas razones. Eso ocurre incluso cuando actuamos irracionalmente, como veremos más adelante. Es decir, incluso cuando tenemos estados mentales irracionales o actuamos irracionalmente, hay causas y razones que explican y dan sentido a esa irracionalidad.

Decir, pues, en sentido coloquial y genérico, que una criatura es racional es igual a decir que es interpretable en términos de estados mentales y acciones, esto es, que es interpretable intencionalmente. Decir de una criatura que es no racional significaría simplemente que no es interpretable intencionalmente. Por otra parte, decir de alguien que es irracional equivaldría a decir que, si bien es interpretable intencionalmente, hemos detectado un estado mental o un tipo de comportamiento que es incompatible con el fondo mayor de consistencia que creemos haber encontrado en él. Esto mostraría que solo alguien que es generalmente racional puede ser irracional. Ahora bien, el sentido en que habitualmente se dice genéricamente de los seres humanos que son racionales —y no, por ejemplo, de los animales— es que podemos

encontrar en los primeros, pero no en los segundos, conducta intencional interpretable en términos de razones, creencias y deseos. La tradición solía entender esto como explicable en virtud de una facultad solo presente en los humanos. Mi propuesta es que la racionalidad no implica necesariamente una facultad de ese tipo, sino la posibilidad de crear un espacio de interpretación intencional común, lo que dará lugar a una forma de vida compartida o, para ponerlo en lenguaje de hermenéutico, a una fusión de horizontes. En condiciones normales, este espacio compartido solo se crea con otros humanos. Sin embargo, hay un sentido en que podemos atribuir deseos y creencias a algunos animales, y generar también espacios compartidos con ellos. Pero no es que la atribución de estados mentales tenga como consecuencia la generación de espacios compartidos, se trata de un mismo e inacabable proceso. En principio, no hay ninguna razón por la que no debamos atribuir estados mentales a algunos animales, pero sería imposible atribuir a un animal la riqueza de estados mentales que podríamos atribuir a un humano, fundamentalmente porque es relativamente reducida la forma de vida común que podemos crear y compartir con el animal, mientras que con el humano podemos crear espacios comunes cada vez más ricos y complejos, que no solo transforman nuestra comprensión de él sino también de nosotros mismos y de ambos como comunidad.

Es condición de posibilidad para la interpretación de una criatura, un texto o una forma de conducta, que le atribuyamos un sistema de estados mentales y acciones a partir de los nuestros propios y sobre la base de nuestras creencias acerca de los estados mentales que creemos que tendríamos si estuviéramos en el lugar del otro. Se sigue de aquí, evidentemente, que sería imposible interpretar a una criatura cuyo comportamiento no fuese reconocido como intencional por nosotros o cuyos deseos y creencias fuesen inconmensurables con los nuestros. Sería también imposible afirmar que podríamos reconocer en alguien comportamiento intencional aunque carezcamos de todo criterio para determinar sus creencias, deseos y acciones. Tampoco sería posible reconocer los estados mentales de alguien sin que tengamos criterios compartidos —y, por tanto, objetivos, porque no dependen de nuestro gusto o capricho— para determinar los valores de verdad de sus creencias y los valores morales de sus acciones, pues tanto los criterios para determinar sus estados mentales como los criterios para determinar estos valores de verdad y valores morales han sido constituidos en una situación comunicativa. Después de todo, los criterios que empleamos en estos casos también están determinados por las creencias —nuestras y del agente— que emergen de esa situación comunicativa.

Pero como hemos visto, hay un grado de indeterminación de la interpretación que recorre tanto los niveles semánticos como psicológicos de manera que, en principio, diferentes intérpretes podrán atribuir distintos significados, estados mentales

y acciones al mismo agente, e incluso una intérprete podrá encontrar racionalidad en un tipo de comportamiento en el que otra ve irracionalidad. Pero esto se deberá a que las intérpretes habrán atribuido al agente distintos sistemas de estados mentales y acciones, con lo que el contenido mismo de cada estado mental y de cada acción será diferente, pues los contenidos de estos elementos están determinados por las relaciones que tienen con los otros elementos en una red holista interconectada. En principio, ambas interpretaciones podrían ser correctas aunque, por supuesto, una podría ser mejor que la otra si es que se muestra como más explicativa en función a ciertos criterios de explicación que puedan llegar a ser compartidos o que sean asumidos por un tercer punto de vista, que será el que sostenga la preferencia de una de estas interpretaciones sobre la otra.

Hasta aquí es importante subrayar que lo que puede o no ser racional es un estado mental o una acción frente al contexto mayor del sistema de estados mentales y acciones atribuido a un agente y, de manera derivada, a un individuo en una situación comunicativa[1]. Cuando interpretamos a un agente, reconocemos su comportamiento como un conjunto de eventos físicos que también pueden ser descritos como acciones intencionales. Si ya hemos reconocido intencionalidad, seremos capaces de atribuirle creencias y deseos porque, en condiciones normales, asumimos que son estos los que causan las acciones del individuo. En otras palabras, sería imposible reconocer comportamiento intencional sin poder atribuir al mismo tiempo estados mentales que lo hagan inteligible: ¿cómo sabríamos que cierto comportamiento físico es intencional si no es porque hemos reconocido ciertos propósitos y creencias que, a nuestro juicio, lo han causado?

Asimismo, sería imposible atribuir a alguien un estado mental inconsistente con el fondo de estados mentales ya atribuidos, pues si no fuésemos capaces de encontrar una estructura consistente en el comportamiento del agente no tendríamos ninguna razón para suponer que su comportamiento es intencional y movido por propósitos. Interpretar al otro como fundamentalmente irracional, como un creyente de falsedades o como alguien cuyos propósitos son sistemáticamente absurdos simplemente no deja nada que pueda ser interpretado, pues desvanece el contexto mismo que hace posible la interpretación. De ser así, se esfumaría la inteligibilidad misma de sus acciones, con lo cual su comportamiento dejaría de ser interpretable como intencional para ser explicable como un conjunto de meros eventos físicos.

[1] En tanto las emociones incorporan creencias en su propia constitución, también pueden ser racionales o irracionales. Por ejemplo, sería irracional que alguien mantuviese emociones negativas frente a otra persona después de enterarse —esto es, después de formarse la creencia de que— de que no fue ella quien le hizo cierto perjuicio.

Es condición de posibilidad para que un tipo de comportamiento sea inteligible que sea *grosso modo* racional ante nuestros ojos[2].

Ahora bien, como también hemos visto, un estado mental o una acción no pueden existir aislados. Por ello, para que se atribuya a un agente un estado mental o una acción irracional, estos deben ser parte de un subsistema de creencias que es a su vez inconsistente con un sistema de creencias mayor. Pero aquí parece surgir una paradoja. Si el principio de caridad sostiene que es condición de posibilidad de la interpretación asumir la racionalidad del otro y si, además, la identidad de un estado mental depende de sus interconexiones y por tanto de su consistencia con otros estados mentales, ¿cómo podría alguien atribuir irracionalidad a un agente? La respuesta es que incluso cuando la intérprete atribuye una dosis de irracionalidad a alguien lo hace con el fin de encontrarlo fundamentalmente racional. Es decir, si la única posibilidad para considerar que el sistema atribuido al agente es básicamente consistente es atribuyéndole un caso de irracionalidad, la intérprete atribuirá esta irracionalidad como una manera de seguir encontrándolo básicamente racional. Pero, para que esto sea posible, será necesario que el estado mental o la acción irracional no sea una atribución aislada, porque se desvanecería del sistema, sino parte de un subsistema internamente consistente pero inconsistente respecto del sistema mayor. Esto sugerirá que lo que la intérprete está atribuyendo es un caso de división de la mente, en el que algunos subsistemas de creencias son internamente consistentes pero inconsistentes entre sí, lo que causa en el individuo acciones o estados mentales que no los justifican. Pero la naturaleza de la irracionalidad es un tema en sí mismo sobre el que volveremos en el último capítulo, de manera que retornemos ahora al problema que nos concierne.

Yendo más lejos de lo que postula el principio de caridad, es posible decir que interpretar al otro es ser capaz de reconocernos en su conducta, esto es, imaginarnos ser él en condiciones contrafácticas, atribuyéndole las creencias que nosotros suponemos que tendríamos si estuviéramos en sus circunstancias. Esto implica verlo como un miembro de nuestra forma de vida o de una forma de vida afín. Malpas (1992) va en la misma dirección. Él ha asociado la noción de sistema de creencias con el concepto husserliano de «horizonte» y el heideggeriano de «proyecto». Sugiere entender un sistema de creencias como un grupo de proyectos de vida asociados entre sí. Esto es plenamente consistente con mi sugerencia de entender un sistema de creencias como un sistema de disposiciones para actuar integrados, es decir,

[2] Este elemento constitutivo del principio de caridad puede encontrarse ya en la concepción kantiana y hegeliana de la historia, donde la idea es que asumir la historia como teleológica y racional es condición de posibilidad de encontrarla inteligible. Véase Kant, 1987 y Hegel, 1974.

como un sistema de prácticas sociales o como una forma de vida. Estos proyectos pueden contener relaciones de inclusión e incompatibilidad. En ocasiones, algunos de ellos ocultan e inhiben a otros, lo que permite echar mano del concepto heideggeriano de *alétheia* y además relacionarlo con la noción freudiana de represión. Dice Malpas:

> Es como si lo psicológico se fragmentara en una miríada de proyectos a veces contradictorios, a veces flexiblemente asociados. Alguna racionalidad es preservada, y esto es lo que permite lograr un grado de comprensión sobre lo que está ocurriendo (1992, p. 79).

Esta idea se puede desarrollar mucho más. Como hemos visto, hay diversas formas de describir una creencia. Esta puede ser descrita como una configuración neuronal, como un estado mental representacional o como un tipo de actitud proposicional. Pero una descripción particularmente útil para la teoría de la interpretación es aquella que la ve como una disposición para actuar y para aceptar otras disposiciones para actuar. Esta es una extensión de la célebre definición de creencia que Charles Sanders Peirce desarrolló y tomó de Alexander Bain. Ahora bien, si aceptamos eso, podemos entender un sistema de creencias como un sistema de disposiciones para comportarse, es decir, como una forma de vida. Pero es importante reconocer que esta concepción de la creencia está tan involucrada con la acción que nos permite superar la vieja dicotomía mentalista moderna que ubica a las creencias y las teorías como, en principio, desvinculadas o desligadas de la praxis y la acción. Si una creencia es una disposición para actuar —y para aceptar otras disposiciones para actuar—, entonces las teorías, en tanto sistemas de creencias o visiones del mundo, son también sistemas de disposiciones para comportarse. Las teorías son formas particulares de comportamiento y toda forma de comportamiento intencional expresa teorías implícitas, incluso si estas no son plenamente conscientes para el agente en cuestión. En el modelo de interpretación que estoy defendiendo, los conceptos mismos de teoría y de racionalidad están anclados en la praxis y ella los define. En este punto, es parte de mi intención reformular conceptos filosóficos tradicionales —como teoría, razón o creencia— de tal forma que nos permitan superar dicotomías propias del paradigma moderno de la subjetividad en las que el sujeto, como muestran tanto Heidegger como Taylor, se encuentra desvinculado y no comprometido con su entorno real. Esta es también una manera de superar la concepción representacionalista del conocimiento propia de la modernidad, en la que este último es concebido como una forma de representación mental antes que como un tipo de participación constitutiva y transformadora de la realidad. En última instancia, de lo que se trata es de superar la dicotomía entre razón teórica y razón práctica. La razón no puede ser teórica

ni práctica, porque la teoría es práctica en su esencia. De hecho, me parece que la mejor respuesta a la pregunta sobre por qué debiéramos ser racionales es porque queremos gobernar nuestras vidas y no tener vidas acráticas.

Mi impresión es que la concepción tradicional de la racionalidad, como desgajada de la praxis, ha tenido como consecuencia que se identifique erróneamente racionalidad con visión del mundo, y entonces algunos se sienten llamados a preguntar si hay una sola racionalidad o varias. Algunos sostienen, por ejemplo, que hay una racionalidad occidental diferente a una andina o una oriental. Me parece que esto refleja una grave confusión respecto del concepto de racionalidad. Identificar racionalidad con cosmovisión conduce a una completa trivialización del concepto mismo de racionalidad, porque, ya que en principio hay un número indeterminado de cosmovisiones, también lo habría de racionalidades. Por contraposición a ello, sugiero una concepción formal y no material de racionalidad, en la que lo que puede o no ser racional no es el sistema de creencias en su totalidad, sino un elemento de este en relación con el todo.

Mi intuición, entonces, es que el concepto de racionalidad alude a la articulación entre las diversas dimensiones o ámbitos de la vida humana, los que se encuentran integrados en el comportamiento del individuo y en su interacción con los otros individuos. Es de esta manera que las diversas facetas de la vida de una persona confluyen y se integran. Los conflictos entre nuestros diversos proyectos y las inevitables incompatibilidades entre los subsistemas de creencias que habitan en nosotros deben resolverse en nuestra acción, es allí donde en última instancia se integra la identidad personal. Si, por el contrario, el sujeto no puede resolver sus contradicciones en un tipo de comportamiento suficientemente articulado, se convierte en un sujeto acrático. En una situación extrema de desarticulación, el individuo dejaría de ser interpretable como agente de acciones intencionales, es decir, sería imposible reconocer una red integrada de estados mentales que haya causado sus acciones, con lo cual no podríamos atribuirle siquiera irracionalidad. En una situación como la descrita, nos veríamos obligados a abandonar el lenguaje intencional para tener que describir a la criatura —ya no sujeto— como un conjunto de eventos físicos gobernados por las leyes de la naturaleza. Pero la intencionalidad tampoco es una propiedad que los eventos tengan en sí mismos, es más bien una propiedad de algunos eventos cuando son descritos mediante un lenguaje intencional. Desde una descripción intencional, mostramos que las acciones son eventos físicos realizados por los agentes en conformidad con sus razones y otras motivaciones. De esta manera, producimos una interpretación del comportamiento del agente que será también una explicación causal de sus acciones sobre la base de los estados mentales que le hemos atribuido.

10.2. Eliminando el relativismo

Volvamos nuestra mirada nuevamente al tema del relativismo, porque podría objetarse que la noción de racionalidad que estoy sugiriendo conduce a aceptar que pudieran existir sistemas de creencias, o formas de vida, internamente consistentes pero inconmensurables entre sí, lo que conduciría a una forma de relativismo conceptual. En el tercer capítulo sugerí algunos argumentos destinados a mostrar por qué el relativismo no se sostiene. Ahora sostendré que la concepción de racionalidad que defiendo muestra la imposibilidad del relativismo.

La primera pregunta que habría que formular es qué es el relativismo conceptual. Este podría definirse como la tesis según la cual hay en principio una diversidad de esquemas conceptuales con los que describimos y categorizamos la realidad, sin que existan criterios objetivos para establecer cuál de ellos es el correcto o si alguno lo es. Una manera alternativa de formular el relativismo es diciendo que no hay criterios objetivos para determinar el valor de verdad de nuestras creencias o el valor moral de nuestras acciones.

Comenzaré concentrándome en la primera formulación. Como vimos, esta ha sido objetada de manera clásica por Davidson (1984d [1974]). Su estrategia es mostrar que este tipo de relativismo presupone una innecesaria distinción ontológica entre esquema conceptual y contenido, que sería la versión actual de la distinción moderna entre sujeto y objeto, teoría y realidad o la distinción kantiana entre fenómeno y cosa en sí. Según esta distinción, nuestros esquemas conceptuales categorizan un contenido previo no interpretado. La idea de Davidson es que la distinción es objetable en tanto presupone la inteligibilidad del concepto mismo de realidad no interpretada, pero no nos detendremos en este tema. Para los fines que nos interesan, lo importante es que la distinción haría posible la doctrina según la cual los distintos esquemas conceptuales guardarían relaciones de inconmensurabilidad entre sí. Davidson objeta simultáneamente la distinción ontológica y el concepto de inconmensurabilidad, y entiende por este último la intraducibilidad entre distintos esquemas conceptuales. Su argumentación es fina y demoledora, aunque define inconmensurabilidad como intraducibilidad, lo que, como vimos en el capítulo tres, es controversial (véase Bernstein, 1983). Por eso, contra esta primera formulación del relativismo tomaré otra línea argumentativa.

El problema con la primera formulación radica en considerar los distintos esquemas conceptuales como, en principio, aislados entre sí, de suerte que cada uno de ellos tendría sus propios criterios de verificación, con lo que se verificarían las mismas creencias asignando distintos valores de verdad. De esta manera, una proposición considerada verdadera en un esquema conceptual podría ser falsa en otro,

sin que existan criterios objetivos para determinar el valor de verdad de la proposición ni la corrección de los esquemas. El inconveniente de esta posición es que describe las características de los esquemas conceptuales desde el exterior de ellos, es decir, desde otro esquema conceptual que pretende no estar comprometido con los esquemas analizados, de manera que nosotros podríamos entender ambos esquemas estando incapacitados para saber cuál es el correcto. Esto presupone que podríamos tener una actitud aséptica y no comprometida con los esquemas conceptuales, de forma que pudiésemos entenderlos sin atribuirles las creencias que nosotros consideramos verdaderas. También presupone que podríamos entender un esquema conceptual sin conocer los criterios por los cuales en ese esquema una proposición es considerada verdadera o falsa. Pero eso es un error. Se sigue de mi argumentación en este libro, que toda interpretación se hace desde un esquema conceptual, o una forma de vida, ya cargado con una serie de presupuestos y no desde un punto de vista *sub especie aeternitatis* desde el cual comprendemos los esquemas en disputa y juzgamos que son inconmensurables entre sí. Este es un interesante punto de contacto con la hermenéutica de Gadamer. Toda interpretación se hace desde un conjunto de anticipaciones de sentido y emerge a partir de la fusión entre el horizonte de la intérprete y el del interpretado. Toda interpretación es, pues, inevitablemente comprometida.

El relativista pretende hablar de lo que ocurre en el escenario como un espectador que se encuentra fuera de este. Yo insistiría en que toda interpretación se hace desde un punto de vista involucrado con los personajes mismos. Por otra parte, como hemos visto, para entender un esquema conceptual es necesario considerarlo básicamente racional y atribuirle muchas de nuestras propias creencias, no solo acerca del mundo sino también acerca del significado de las palabras. Por eso, sería imposible que yo entendiese dos esquemas conceptuales distintos —habiéndolos hecho inteligibles precisamente al haberles atribuido muchas de mis propias creencias— sin que considere que ellos tienen suficientes creencias en común como para que se pueda establecer criterios de verificación que también son compartidos. Estos criterios de verificación permitirían, en principio, establecer los valores de verdad de las creencias y los valores morales de las acciones de manera objetiva, en los que «objetivo» significa no arbitrario ni subjetivo, sino más bien válido para todos o muchos de los puntos de vista involucrados en la interpretación. En otras palabras, el hecho de que yo reconozca los comportamientos ajenos como conducta intencional y como reflejo de formas de vida implica que he determinado creencias, significados, deseos y acciones que en alguna medida también son compartidos por mí. Si he podido hacer eso, entonces hay creencias, significados, deseos y acciones atribuidas a ambos esquemas que también son compartidos por ellos —desde mi punto de vista—

lo que hace posible la existencia de criterios de verificación comunes. Por supuesto, podría ocurrir que yo reconozca elementos compartidos en ambos esquemas, pero que ellos no se reconozcan mutuamente como compartiéndolos. Eso es concebible, aunque solo podría ocurrir si ellos no reconocen que el otro es un esquema conceptual, es decir, que el otro esquema expresa conducta intencional. Pero esto solo mostraría las limitaciones interpretativas de esos esquemas, no probaría ningún tipo de relativismo. Además, si para interpretar al otro —sea este un individuo o una cultura diferente de la nuestra— es necesario asumirlo básicamente racional y atribuirle un sistema integrado de creencias, significados, deseos y acciones, que son los que nosotros creemos que tendríamos si fuéramos él, entonces se sigue, inevitablemente, que no podemos interpretar su sistema de creencias como inconmensurable con el nuestro. El solo hecho de considerar un conjunto de eventos físicos como acciones intencionales implica atribuir a su agente estados mentales y, en consecuencia, posibilitar la interpretación. La intuición central de Davidson, entonces, es que el concepto mismo de inconmensurabilidad presupone un punto de vista privilegiado desde el cual confirmamos la disyunción entre los sistemas en conflicto, o entre uno de ellos y el nuestro. Pero, como ya se ha visto, ambas posibilidades serían insostenibles.

Lo que resulta inaceptable es el concepto de inconmensurabilidad total, no el de inconmensurabilidad parcial. De hecho, todos los sistemas de creencias pueden ser considerados como parcialmente inconmensurables, no solo unos respecto de otros sino también respecto de nosotros e incluso respecto de ellos mismo en un momento posterior en el tiempo. Sin embargo, como muestra Davidson, si decimos que hay inconmensurabilidad parcial estamos implícitamente admitiendo que hay conmensurabilidad parcial, y esto es simplemente conmensurabilidad, con lo cual sería mejor utilizar simplemente el concepto de indeterminación de la interpretación, que para todos los efectos significa lo mismo que inconmensurabilidad parcial y no nos conduce a estos malentendidos. Lo que sostengo, por tanto, es que el fenómeno de la indeterminación de la interpretación está presente en cualquier proceso comunicativo, aunque esto no conduce a ningún sentido de inconmensurabilidad ni tampoco a la forma de relativismo conceptual que estamos analizando.

Ahora pasaremos a examinar la segunda formulación del relativismo. Esta sostiene que no hay criterios objetivos para determinar los valores de verdad de una creencia o el valor moral de una acción. Así, estos valores se determinarían mediante los criterios existentes al interior de un sistema de creencias, pero, al no haber criterios para determinar la corrección o no de un sistema de creencias como tal —porque esto solo podría fijarse desde un sistema mayor que incluya el sistema de creencias en cuestión—, no existirían los criterios objetivos que estamos buscando.

Lo primero que señalaré es que este tipo de relativista es, en realidad, un fundacionalista. Él piensa que para que haya conocimiento objetivo es necesario que se compruebe, más allá de toda duda concebible, la verdad de una creencia. Además, presupone que esta demostración debe estar fundada sobre una base indubitable, universal, inconcusa y definitiva. La tradición ha considerado distintos candidatos que podrían cumplir esta función: la experiencia sensible, las ideas innatas, el *cogito*, las categorías del entendimiento, etcétera, todos ellos con dudoso éxito. Al descubrir el relativista que esta base no existe, concluye que el conocimiento objetivo es imposible y que da igual tener una creencia u otra, porque no hay cómo darles una justificación última. La primera objeción que hay que hacer a este relativista es que cualquier tesis que él afirme resultaría siendo tan válida como su opuesta, con lo cual la tesis relativista misma se autodestruye. Pero no utilizaré ahora esta argumentación porque hay versiones más sofisticadas de relativismo que podrían haberla superado con éxito (Olivé, 1988b). Más bien sugeriré que el problema con esta forma de relativismo es que tiene sus pretensiones epistemológicas demasiado altas, tan altas como el fundacionalista. Si la justificación de una creencia no alcanza sus exigencias, es rápidamente descartada. Aquí parecen estar en juego los conceptos mismos de conocimiento y objetividad. Sospecho que es posible entender estos conceptos de una manera más humilde, no fundacionalista, que permita aceptar el conocimiento objetivo como una auténtica posibilidad humana. Pero no me detendré en el análisis de estos conceptos pues nos llevaría demasiado lejos[3], ahora volveré a la formulación que nos interesa.

El relativista sostiene que no hay criterios objetivos para determinar el valor de verdad de una creencia y esto, entre otras cosas, porque la misma proposición puede ser probada como verdadera en un sistema y como falsa en otro, con lo cual toda justificación es interna a un sistema, y cuando hay conflictos entre sistemas no es posible determinar cuál es el correcto. Una objeción a esta afirmación es que si una oración es probada como verdadera en un sistema y como falsa en otro, estamos hablando de una misma oración gramatical pero de dos proposiciones diferentes en tanto el significado de una oración, es decir la proposición contenida en ella, está determinado por sus relaciones con las otras proposiciones del sistema. Al decir esto de ninguna manera me estoy comprometiendo con la existencia de proposiciones en el sentido convenientemente objetado por Quine. Estoy distinguiendo simplemente entre la cadena gramatical a la que solemos llamar oración y el significado que podría atribuirle una intérprete bajo las condiciones apropiadas.

[3] Para una discusión más detallada del concepto de justificación, véase Quintanilla, 2001b.

Entonces, bajo una concepción holista del significado como la que yo asumiría aquí, es correcto sostener que tanto el significado como el valor de verdad de una creencia se fijan solo al interior de un sistema de creencias. Yo no puedo saber cuál es el valor de verdad de una creencia si previamente no sé cuál es su significado; ya que su significado está parcialmente determinado por el resto de creencias del sistema, su valor de verdad también lo estará. Ahora bien, esto no impide que haya criterios de verificación objetivos, por el contrario, lo hace posible. Esta es una posición cercana a la de Dewey, quien solía decir que ciertas acciones están objetivamente bien en ciertas circunstancias pero objetivamente mal en otras. Una manera análoga de poner esto sería diciendo que una oración puede ser objetivamente verdadera en un sistema de creencias y objetivamente falsa en otro, siempre que quede claro que esto ocurre porque lo que es verdadero o falso es la proposición que está siendo expresada por la oración, con lo cual se trata de dos creencias distintas expresadas por la misma oración.

Se podría poner las cosas de esta manera: el valor de verdad de una oración es relativo a un lenguaje, teoría o sistema de creencias, aunque esto puede ser objetivo. El valor de verdad de una oración depende de dos cosas: de su significado y de la manera como es el mundo. Pero el significado de una oración solo puede ser determinado en un sistema de creencias, luego su valor de verdad solo puede ser determinado en un sistema tal. Es, en ese sentido, que el valor de verdad de una oración es relativo a un sistema de creencias. Pero, además, el sistema de creencias en cuestión puede tener criterios objetivos, internos a ese sistema de creencias, para determinar el valor de verdad de una oración que tiene significado en ese sistema. Entonces, aunque el valor de verdad de una oración sea relativo a un sistema de creencias, puede ser objetivo.

Emplearé el mismo ejemplo discutido en el capítulo tres. Supongamos que me preguntara cuál es el valor de verdad de la oración «La masa es invariable». Esa oración, en sí misma, carece de valor de verdad, porque no se ha estipulado cuál es su significado. Para empezar el significado de la oración «La masa es invariable» depende del significado de la palabra «masa», y este, a su vez, dependerá de si tomamos como marco de referencia la mecánica de Newton o la teoría de la relatividad de Einstein. Para Newton la masa es una constante que se define como la cantidad de materia producto de la densidad por el volumen. Para Einstein, por el contrario, la masa es una variable que está en relación con la velocidad, en la medida en que su mínimo valor es el mismo de la masa newtoniana y aumenta su máximo valor infinitamente mientras el móvil se acerca a la velocidad de la luz. Entonces, la oración «La masa es invariable» tiene significados distintos si se la interpreta en un contexto newtoniano o einsteiniano. Como el valor de verdad de una oración depende de su significado y de la manera como es el mundo, entonces solo se puede asignar un valor de verdad

a la oración «La masa es invariable» en relación con —es decir, relativo a— un sistema de creencias, en este caso una teoría científica. Sin embargo cada una de estas teorías, la de Newton y la de Einstein, tiene criterios objetivos para determinar el valor de verdad de una oración que tiene significado en esa teoría. Así, entonces, es objetivamente verdadero que la masa es invariable en relación con la mecánica de Newton y es objetivamente falso que la masa sea invariable en relación con la teoría de la relatividad de Einstein. Quedará claro, naturalmente, que la oración «La masa es invariable» tiene significados diferentes, es decir da lugar a creencias distintas, en la mecánica de Newton y en la teoría de la relatividad de Einstein. El punto es que ni el valor de verdad de una creencia ni el valor moral de una acción son arbitrarios, subjetivos o dependen del gusto de cada quien. Pueden ser objetivos, aunque en relación con un contexto dado[4].

La idea, hasta aquí, es que puede haber criterios compartidos para determinar los valores de verdad de una creencia y los valores morales de una acción, de tal suerte que estos criterios constriñen a todo agente racional a aceptar ciertas conclusiones y no otras, siempre que estos agentes acepten el sistema de creencias —o el esquema conceptual— en relación con el cual estos criterios son válidos. En este punto alguien podría sostener que la noción de objetividad que estoy empleando es insuficiente y que a esa persona no le interesa saber si una oración es objetivamente verdadera en un sistema de creencias dado, sino si es objetivamente verdadera sin más, es decir, objetivamente verdadera para todos los sistemas de creencias. Mi respuesta a esa posible objeción es que eso sería como preguntarse cuál es el significado de una oración en sí misma, independientemente de un lenguaje o en cualquier lenguaje dado. Es obvio que solo se puede fijar el significado de una oración en un determinado lenguaje; como el valor de verdad de una oración depende de su significado, será igualmente obvio que el valor de verdad de una oración solo se puede fijar en un lenguaje o sistema de creencias. Debo insistir, sin embargo, en que los valores de verdad de las creencias de un sistema son siempre atribuidos por el sistema de la intérprete y generados en el espacio común producido por la situación comunicativa, en relación con el mundo compartido, con lo cual no se trata de sistemas aislados como compartimientos estancos.

Ahora bien, quizá en este punto se quiera saber qué ocurre cuando tenemos dos sistemas de creencias diferentes y necesitamos saber cuál es el verdadero. Lo primero que diría es que se puede atribuir valor de verdad a una creencia solo en relación con

[4] También podría añadirse, aunque esto se aleja de nuestros intereses presentes, que puede ser un hecho moral que cierta acción esté objetivamente bien en ciertas circunstancias, así como también puede ser un hecho que p, si a la oración de que «p» le asignamos un valor de verdad objetivo en cierto sistema de creencias. Véase Quintanilla, 2009c.

un sistema de creencias al cual pertenece, o de un sistema de creencias en relación con otro mayor al cual pertenece, lo que muestra que «verdadero» también es un predicado relacional y no monádico.

Pero no tendría mucho sentido preguntar si toda una visión del mundo es verdadera o falsa si no tenemos un sistema mayor que fije los criterios de evidencia o verificación. En la práctica, podemos fijar el valor de verdad de una creencia al interior de un sistema y luego preguntarnos si este sistema es preferible o no a otros alternativos, apelando a virtudes epistémicas como fecundidad, predictibilidad, consistencia, etcétera. Pero este es un tema sobre el que no nos detendremos ahora.

Ahora bien, alguien podría decir que esto puede ser aceptable en contextos donde los sistemas de creencias tienen criterios muy precisos de asignación de valores de verdad, como es el caso de las ciencias naturales, pero no en áreas donde los criterios son menos precisos, como por ejemplo la ética o las ciencias sociales. Yo diría que, en cualquier caso, el valor de verdad de una oración es relativo a un sistema de creencias, pero también diría que esto «puede» ser objetivo —no siempre lo es— y que «puede» ser más o menos objetivo según si los criterios de asignación de valores de verdad son más o menos precisos en el sistema de creencias en cuestión.

10.3. Formas de vida y relativismo

Todo esto nos conduce a otro punto por tratar. Como ya hemos visto, el sistema en el que se fijan los valores de verdad de las creencias es, en realidad, un sistema de prácticas sociales compartidas o una forma de vida. Esta forma de vida se constituye intersubjetivamente en las interacciones comunicativas de los agentes intencionales. Se trata del espacio común en el que confluyen los estados mentales, significados y acciones de los agentes que interactúan, produciendo así una fusión de horizontes. A partir de estos espacios compartidos se constituyen los criterios objetivos que permitirán determinar los valores de verdad de las creencias y los valores morales de las acciones. Se trata de criterios objetivos porque, al ser constituidos intersubjetivamente, constriñen a los agentes que interactúan según los criterios que ellos mismos han colaborado en constituir. El punto es que en las interacciones comunicativas no solo se fijan los significados y los estados mentales de los agentes, sino también los criterios con los cuales se determinan estos significados y estos estados mentales, así como los valores de verdad de las creencias y los valores morales de las acciones. Pero, ciertamente, no se trata de afirmar que las proposiciones posean en sí mismas valores de verdad objetivos, la idea es más bien que los valores de verdad de las creencias pueden construirse, así como también descubrirse, intersubjetivamente con un grado de validez objetiva que constriña a todos los involucrados en la constitución de las creencias en cuestión.

Como ya se habrá sospechado, las tesis que estoy desarrollando aquí están muy cerca de algunas intuiciones wittgensteinianas. En relación con la objetividad de la verdad, Wittgenstein rechaza una posición correspondentista, así como un convencionalismo ingenuo. Ambas posiciones comparten el error de imaginar el lenguaje como desconectado de la realidad y de las prácticas sociales humanas. En *Investigaciones filosóficas* pregunta: «¿[d]ices, pues, que la concordancia de los hombres decide lo que es verdadero y lo que es falso?». Su respuesta es: «Verdadero es lo que los hombres *dicen;* y los hombres concuerdan en el lenguaje. Esta no es una concordancia de opiniones sino de formas de vida» (1988, § 241). En efecto, las proferencias verbales son acciones intencionales al interior de juegos de lenguaje, los cuales son sistemas de prácticas sociales inseparables de formas de vida. Los acuerdos acerca de los significados de las oraciones y sus valores de verdad no son independientes ni están desconectados de estas formas de vida. Por eso no puede decirse que sean acuerdos arbitrarios. Tampoco puede decirse que las formas de vida sean arbitrarias. Las formas de vida son el fondo o el marco de referencia frente al cual las oraciones son verdaderas o falsas, son el contexto último. No tiene sentido decir que sean arbitrarias porque no hay un contexto ulterior frente al cual puedan o no serlo. Un sistema de creencias o una creencia son fundados o arbitrarios frente a un fondo o marco de referencia anterior. Si no hay tal marco de referencia para una forma de vida, no tiene sentido preguntar si esa forma de vida es arbitraria o está fundada; una forma de vida está más allá de la fundamentación o la arbitrariedad. Dice Wittgenstein que una forma de vida es «algo que está más allá del ser justificado o no justificado; como si existiera, como algo animal» (1972, aforismo 359). Así es como las formas de vida se crean en las relaciones intersubjetivas entre las personas, de manera que se trata siempre de un fondo en transformación y cambio.

Para concluir este capítulo, diré que la concepción de la interpretación que he defendido tiene como propósito colaborar en la superación de ciertos presupuestos propios de la modernidad:

(1) Una concepción internista del sujeto que debe ser reemplazada por una visión intersubjetivista.

(2) Una concepción intelectualista de la racionalidad y la comunicación que debe ser reemplazada por una concepción práctica que involucre al individuo en su totalidad.

(3) Una idea referencialista y representacionalista del lenguaje que debe ser reemplazada por una concepción, más acorde con el giro lingüístico y la hermenéutica, en la que el lenguaje participa en la constitución de lo que conocemos y no es un instrumento meramente designativo sino un fenómeno multidimensional.

Considero que la versión davidsoniana del principio de caridad es imperfecta y debe ser complementada. Lo que intento añadir al principio de caridad es:

(1) Una teoría de la simulación.

(2) Una teoría de la cooperación interpretativa, en la que tanto el significado como la constitución misma de los estados mentales, en el terreno epistemológico y ontológico, son una creación compartida por intérprete y agente.

(3) Una concepción de la comunicación como la constitución de un espacio compartido en el ámbito de prácticas sociales o formas de vida, antes que como la mera reproducción de los significados y estados mentales del otro.

Pero hay un punto que debe ser aclarado. Yo no diría —y sospecho que Davidson tampoco— que el principio de caridad agota el concepto mismo de comunicación. El principio de caridad no es una teoría resumida de la comunicación ni pretende explicar qué es lo que ocurre cuando dos personas se comunican. El principio de caridad es una explicitación de algo que es condición necesaria —aunque no suficiente— para que la comunicación sea posible. En otras palabras, el principio de caridad establece, a la manera de un argumento trascendental, lo que la intérprete debe presuponer para que pueda interpretar al agente. El punto es que estos presupuestos tienen suficiente riqueza como para extraer a partir de allí consecuencias en torno a la comunicación, la racionalidad y el relativismo. En este orden de ideas, hay algunos otros puntos de detalle que me gustaría comentar.

El principio de caridad ha sido acusado de ser intelectualista y de estar comprometido con una forma de imperialismo psíquico. Abordaré esos argumentos en el siguiente capítulo. Ahora solo recordaré que no es que la intérprete imponga sus propios estados mentales en el otro, sino más bien que ella parte de imaginar ser el otro en condiciones contrafácticas para así poderle atribuir los estados mentales que ella cree que tendría si fuera el agente. Este es el elemento proveniente de la teoría de la simulación que deseo integrar con el principio de caridad. Con frecuencia la intérprete hará atribuciones de las que no es plenamente consciente, guiada por elementos afectivos y emocionales. Pero las atribuciones no tienen por qué tener un cariz intelectualista desgajado de lo emocional. La persona que se compadece de otra —es decir, que padece con la otra— lo hace porque puede imaginarse por lo que está pasando la otra persona y entonces le atribuye los estados mentales que cree que tendría si fuera él. Es la capacidad de simular ser el otro en condiciones contrafácticas lo que permite a la intérprete atribuir ciertos estados mentales y no otros. Cuando uno empatiza con otra persona, incluso en un nivel afectivo muy básico, hay una gama de atribuciones que le está haciendo, muchas de las cuales no han sido elaboradas plenamente

incluso por uno mismo. Hasta ahí nos movemos en la perspectiva de la intérprete. Sin embargo, tan pronto la interacción comienza a discurrir, la intérprete tendrá que modificar sus hipótesis interpretativas para poder crear un espacio común con el agente. Pero la expresión «hipótesis» no debería tener una carga intelectualista. Estas, como las anticipaciones de sentido de Gadamer, no son solo cognitivas sino también afectivas. Así, la comunicación se vuelve una obra compartida que no solo permite que ambos interlocutores se comuniquen, pues también los transforma. Me parece que en esto no hay nada que pueda ser acusado de imperialismo psíquico. La clave es ver las teorías como sistemas de creencias y las creencias como disposiciones para actuar. De esta manera, las teorías serían, en el fondo, sistemas de prácticas sociales compartidas, explícitas o tácitas, conscientes o no, que inevitablemente involucran elementos emocionales y afectivos. Las diversas atribuciones de estados mentales y significados se dan holística y simultáneamente. Pero lo importante es que se dan en tanto intérprete y agente interactúan; es en ese sentido que constituyen una obra compartida y no una imposición de la intérprete hacia el agente. No solo proyectamos en el otro nuestras creencias y deseos sino también nuestros sentimientos y emociones al imaginar lo que sentiríamos si fuésemos el otro.

Finalmente, me parece que si se considera que el principio de caridad peca de etnocéntrico, es necesario distinguir —siguiendo a Rorty (1991b, p. 30, n. 13)[5]— entre dos tipos posibles de etnocentrismo. Hay uno inevitable y deseable, y otro indeseable e imperialista. El sentido indeseable es el que lleva a una sociedad a sentirse epistemológica o axiológicamente superior a las otras. De ese sentido debemos alejarnos. Pero, en otro sentido, toda sociedad es etnocéntrica porque no puede evitar ver a las otras si no es desde su propia perspectiva, y no puede evitar creer que sus creencias son verdaderas y sus valores son correctos. Este sentido no solo es inevitable sino también deseable, porque le permite ver a los otros como semejantes con los que tiene lazos de solidaridad y responsabilidades morales. Sería muy difícil sentir responsabilidad moral por alguien tan diferente de nosotros que no pudiéramos siquiera imaginar su sufrimiento o su desdicha, es decir, por alguien a quien no pudiéramos atribuir estados mentales. Pienso, entonces, que si bien el principio de caridad no tiene consecuencias imperialistas, sí puede tener ciertas mitigadas consecuencias etnocéntricas, las que podrían ser inevitables e incluso moralmente deseables. Esto es lo que abordaremos en el siguiente capítulo, lo que nos permitirá volver sobre la pregunta que atraviesa este libro.

[5] Para una discusión en mayor profundidad sobre este punto véase Quintanilla, 2015.

Capítulo once
Racionalidad y etnocentrismo

11.1. Atribuir irracionalidad para comprender

En este capítulo analizaré la tesis según la cual comprender al otro —sea un individuo o una colectividad— tiene como condición necesaria, aunque no suficiente, encontrar la racionalidad que gobierna su comportamiento. Pero, sobre todo, discutiré si esta tesis tiene consecuencias etnocéntricas.

Por racionalidad —como hemos visto en el capítulo anterior— no debería entenderse los procesos de razonamiento característicos que subyacen a una visión del mundo o a un sistema de creencias, como cuando se dice, por ejemplo, que hay una racionalidad occidental diferente a una racionalidad andina u oriental. Por racionalidad se entenderá una exigencia formal de consistencia entre las creencias, deseos, afectos y acciones de un agente, desde el punto de vista de una intérprete. Así la concepción de racionalidad que asumo es relacional y no monádica, porque emerge necesariamente en una situación comunicativa dada. Tampoco es una concepción teórica de la racionalidad sino práctica, porque no es solo la coherencia entre creencias lo que es relevante, sino la consistencia entre estados mentales y acciones.

Con frecuencia nuestro comportamiento es irracional. Como veremos en el próximo capítulo, un estado mental o una acción es irracional si es inconsistente con nuestro sistema de estados mentales consciente. Eso ocurre, por ejemplo, si hacemos lo que sabemos que no es nuestra mejor opción o lo que no deseamos hacer, si creemos que lo que sabemos es falso, si valoramos algo que sabemos que no tiene valor para nosotros o si tenemos una emoción que creemos que no estamos justificados en tener.

Pero, aun cuando atribuimos a alguien una creencia o acción irracional —en un aspecto puntual de su comportamiento o de su vida—, lo hacemos para comprenderlo, es decir, para capturar algo de su subjetividad y hacerlo inteligible ante nuestros ojos, así como para encontrar la racionalidad que subyace a su comportamiento.

Esa es la paradoja de la irracionalidad: solo un ser racional puede ser irracional. Solo el ser humano tiene el privilegio de poder ser absurdo, dice Hobbes, y Davidson añade: «la irracionalidad es una fisura al interior de la casa de la razón» (2004c [1982], p. 169). Por eso, solo podemos atribuir irracionalidad a alguien si le hemos atribuido un marco mayor de racionalidad que subyace a su comportamiento y sus estados mentales, con lo cual la atribución de irracionalidad tendrá que ser siempre excepcional. Suponer que la totalidad del comportamiento de un individuo es irracional equivale a no reconocer comportamiento en lo más mínimo, es no haber comenzado siquiera a interpretarlo.

Esta discusión es particularmente importante porque nos plantea la necesidad de superar una dicotomía presente en la filosofía contemporánea entre, de un lado, la tesis de que hay una racionalidad supra histórica que convenientemente resulta siendo la occidental —con las nefastas consecuencias morales que tiene esta forma de etnocentrismo— y, de otro lado, la tesis de que hay tantas racionalidades cuantas visiones del mundo, lo que, como hemos visto, hace que se trivialice el concepto mismo de racionalidad.

Pero, como el asunto es más complicado, tenemos que internarnos en sus detalles. Plantearé nuevamente algunos aspectos del principio de caridad. Como he sostenido en los capítulos anteriores, no defiendo la versión clásica de este principio ofrecida por Davidson, sino propongo una formulación que incluye como un pilar fundamental la noción de simulación, esto es, la capacidad de imaginar ser el otro bajo condiciones contrafácticas. Pero lo que discutiré en este capítulo es si este modelo puede conducir a alguna forma de etnocentrismo, como creen sus críticos o si, por el contrario, es una buena manera de superar la dicotomía entre un etnocentrismo indeseable y el relativismo cultural.

Como hemos visto, la versión davidsoniana clásica del principio de caridad sostiene que es condición de posibilidad para que la intérprete pueda comenzar a interpretar al agente que ella asuma que él es básicamente racional, que la mayor parte de sus creencias son verdaderas y que él tiende a actuar en concordancia con el bien, todo según los criterios de la intérprete. En otras palabras, ella deberá asumir que hay consistencia básica entre los estados mentales y las acciones del agente, y que ambos comparten la mayor parte de estados mentales. Al hacer esto, ella estará proyectando una parte de su propio sistema de estados mentales al agente. Después de comenzar a interactuar, sin embargo, lo más probable es que ella tenga que modificar algunas de sus atribuciones para poder hacer inteligible el comportamiento inesperado del agente. La intérprete necesitará, por tanto, un criterio para determinar qué estados mentales debe atribuirle y en qué circunstancias. Davidson no ofrece este criterio, por lo que sugiero que uno adecuado es la noción de simulación. Utilizar la noción

de simulación como criterio para la atribución permitirá mostrar que la intérprete proyectará en él solo aquellos estados mentales que ella cree que tendría si fuese el agente, en las circunstancias en que ella cree que él se encuentra. Para ello tendrá que imaginarse ser él en las circunstancias en que ella cree que él se encuentra. Una de las objeciones centrales contra el principio de caridad es que está comprometido con una forma de etnocentrismo. En lo que sigue discutiré si este es un riesgo real. Al hacerlo, explicaré con más detalle las características del principio de caridad en la versión que me interesa defender.

Podría objetarse que al asumir el principio de caridad no se está dejando lugar para aceptar la diferencia radical ya que, según algunos, este principio no admite la posibilidad de entender al otro como totalmente diferente de uno. Al proyectar nuestros estados mentales en el otro —podría decirse— lo estamos convirtiendo en un *alter ego* de nosotros mismos y, entonces, no estamos capturando lo que él realmente es sino solo una fabricación de nuestra propia imaginación. La intérprete se estaría engañando a sí misma al creer que está comprendiendo al agente, cuando en realidad solo estaría imponiendo en él sus propios estados mentales. Mientras que en la interpretación individual esto es una forma de egocentrismo, en la interpretación intercultural es una forma de etnocentrismo; en ambos casos no es comprensión sino imposición. Si estos cargos estuvieran justificados, ciertamente constituirían argumentos devastadores contra una teoría de la interpretación basada en el principio de caridad.

Jonathan Lear (1991) piensa que la obsesión de este principio por encontrar o imponer consistencia en el agente podría convertir a la intérprete en incapaz de comprender la complejidad real de la mente del otro. La objeción de Lear está dirigida principalmente contra la versión de Quine de este principio y, en alguna medida, contra la de Davidson. Ya que mi propia versión del principio tiene más amplitud que las anteriores, pienso que puede afrontar exitosamente las objeciones de Lear, como intentaré sostener. Con la finalidad de probar su caso, Lear retorna a las fuentes y cita a Quine en relación al célebre experimento mental de la interpretación radical:

> Simplificando, sin duda, supongamos que se sostiene que algunos nativos aceptan como verdadera una cierta oración que tiene la forma «p y no p». O, para no simplificar demasiado, que aceptan como verdadera una oración de la forma «q ka bu q», cuya traducción inglesa tiene la forma «p y no p». Pero ahora, ¿qué tan buena traducción es esta, y cuál puede haber sido el método de los lexicógrafos? Si alguna evidencia puede contar contra la adopción del lexicógrafo de «y» y «no» como traducciones de «ka» y «bu», ciertamente la aceptación de los nativos de «q ka bu q» como verdadero cuenta abrumadoramente. Se nos deja con la ininteligibilidad de la doctrina de que hay personas prelógicas. La prelogicidad es un rasgo inyectado por los malos traductores (1991, p. 191).

Esto es lo que Lear comenta sobre el texto citado de Quine:

> La prelogicidad no es necesariamente un rasgo inyectado por los malos traductores: podría ser un rasgo reconocido por antropólogos sensibles a las irrupciones de lo arcaico en la vida diaria. Decir a priori que los nativos de ninguna manera podrían aceptar oraciones que expresen contradicciones es, en efecto, rehusarse a reconocer lo que la vida mental arcaica es: mente en actividad (p. 191).

Creo que Lear está equivocado en este punto. El principio de caridad no tiene que sostener que la intérprete no puede interpretar las proferencias o acciones del agente como inconsistentes. De hecho, en muchos casos será necesario atribuir a las personas inconsistencias e irracionalidad localizadas para poder darles sentido, es decir, para que sean básicamente inteligibles. En un sentido lógico, dos creencias que expresan proposiciones son inconsistentes si violan el principio de no contradicción. Pero, en un sentido más amplio, el comportamiento de un agente es inconsistente si, a los ojos de la intérprete, muestra incompatibilidades entre sus creencias, deseos y acciones. Más aún, estados mentales y acciones serán considerados incompatibles si parecen excluirse mutuamente en la interpretación que la intérprete hace del agente. En el texto citado, Lear parece estar interesado solo en la consistencia lógica, aunque me parece que tiene en mente —o por lo menos debería— el sentido más amplio; en mi argumentación incluiré ambos sentidos.

Lo que el principio de caridad sostiene es que una intérprete puede reconocer una inconsistencia en el comportamiento del agente solo contra el marco de lo que ella considera el comportamiento general del agente. Ya que las creencias y otros estados mentales expresan disposiciones para comportarse, la intérprete no podría ver las acciones del agente como básicamente inconsistentes o como dirigiéndose hacia objetivos incompatibles, en la mayor parte de casos, porque eso distorsionaría su creencia de estar expuesta a acciones y objetivos. Para poder atribuir una acción al agente la intérprete debe ser capaz de reconocer un patrón de comportamiento intencional y, para que eso sea posible, ella debe ser capaz de encontrarlo básicamente consistente. Si la intérprete no es capaz de encontrar un marco de consistencia en el comportamiento del agente es porque ella no está en condiciones de encontrar suficiente comportamiento intencional que pueda ser relacionado con sus propios estados mentales. Esto podría deberse a su incapacidad para interpretarlo apropiadamente o a que a sus ojos no hay nada que interpretar. Naturalmente, una intérprete más creativa podría encontrar comportamiento intencional allí donde una menos creativa no es capaz de hallarlo.

Algunas formas de comportamiento inconsistente podrían ser causadas por enfermedades mentales severas. En estos casos todavía podría verse al agente realizando

acciones básicas, pero la atribución de acciones intencionales más complejas probablemente se distorsionará y volverá problemática. Así, la intérprete tendrá que desplazarse desde intentar comprender al agente mediante la atribución de estados mentales conscientes, a intentar explicar algunos aspectos de su comportamiento como causados por estados mentales inconscientes. En casos más extremos, la intérprete deberá dejar de comprender intencionalmente su comportamiento para explicarlo solo físicamente.

Al sostener que es condición de posibilidad de la interpretación presuponer que los otros son mayormente consistentes, racionales, creyentes de verdades y que suelen guiar sus acciones por los mismos objetivos que guían las acciones de la intérprete, el principio de caridad implica que la interpretación requiere que asumamos que los otros son suficientemente similares a nosotros como para que podamos proyectar nuestros propios estados mentales en ellos. Ahora bien, Lear concede que el principio de caridad es una condición necesaria *a priori* y que no es opcional, pero dice:

> Aunque el principio es *a priori*, su contenido no lo es. No sabemos *a priori* qué es ser nosotros mismos, y entonces no podemos fijar límites al comportamiento de los otros que podamos reconocer como inteligibles. Ha habido una tendencia en el estudio filosófico de la interpretación a pasar de «interpretar a los otros como nosotros» hacia «interpretar a los otros como racionales» (1991, p. 191).

Lear sostiene dos tesis aquí. Primero dice que ya que no sabemos *a priori* qué es ser nosotros mismos, no podemos saber cuáles son los límites de la inteligibilidad del comportamiento de los otros. Pero, ¿qué significa que no sepamos qué es ser nosotros mismos? Hay dos posibilidades. En una lectura, Lear querría decir que no podemos conocer todos los estados mentales que causan o causarán nuestras acciones y, ya que no tenemos un conocimiento cabal de esos estados mentales, no podemos saber qué tan similares o diferentes de nosotros deben ser los otros para que podamos interpretarlos. Ciertamente, convengo en que no podemos tener un conocimiento cabal de nuestros estados mentales, precisamente porque muchos de ellos son inconscientes, pero esto no es requerido por el principio de caridad. Tendemos a asumir que la gente se comporta de cierta manera bajo circunstancias particulares, aunque no tengamos plena conciencia de las creencias que constituyen tal supuesto. La otra posible lectura de la cita de Lear es que no resulta claro qué cuenta como ser similar a uno, ya que la similitud es una noción vaga y relativa a un contexto dado. Luego, ya que no es claro qué cuenta como ser similar a uno, tampoco es claro qué cuenta como ser inteligible para uno. Ciertamente estaría de acuerdo con esta afirmación, que es perfectamente compatible con mi versión del principio de caridad,

ya que la similitud tanto como la inteligibilidad son cuestiones de grado. Pero regresaré sobre este tema en un momento.

La siguiente afirmación de Lear es que es un error suponer que interpretar a los otros como semejantes es considerarlos racionales. De hecho, comenzamos asumiendo que los otros son racionales, pero esto no impide que pasemos a hacer atribuciones de irracionalidad si eso colabora en dar sentido al agente. Como he sostenido previamente, interpretar a alguien como racional es encontrar un marco de consistencia entre sus estados mentales y acciones, es decir, reconocer que él tiende a actuar de acuerdo con lo que él desea y cree que es apropiado. Una acción no puede ser inconsistente con una creencia aislada, solo puede serlo con un sistema de estados mentales. Las atribuciones de irracionalidad requieren una explicación precisamente porque, a los ojos de la intérprete, deben ser excepcionales.

Para que la intérprete pueda atribuir un estado mental al agente ella tiene que atribuir un sistema de ellos y para que algo sea un sistema este tiene que ser básicamente consistente. Una lista de creencias inconexas o incompatibles no constituye un sistema. De aquí se sigue que si la intérprete ve al agente como básicamente irracional, resultará imposible para ella atribuirle un sistema de estados mentales intencionales y, por tanto, no podrá siquiera atribuir uno solo. Como dice Davidson, «ver demasiada sinrazón de parte de los otros es simplemente minar nuestra habilidad para comprender sobre qué es que son tan poco razonables» (1974a, p. 153). Además, si la intérprete no puede atribuir al agente ningún estado mental, ella tampoco podrá encontrar acciones intencionales, dado que estos son eventos físicos que pueden ser descritos como siendo causados, por lo menos parcialmente, por los estados mentales de los agentes. Una posible explicación de la irracionalidad puede ir en la dirección de la atribución de escisión de la mente en subconjuntos de estados mentales internamente consistentes y superpuestos que, sin embargo, son inconsistentes entre sí. Veremos esto en el próximo capítulo.

11.2. Comprender a quienes son radicalmente diferentes de nosotros

La mayor objeción de Lear es que el principio de caridad no permite el reconocimiento de que el agente sea radicalmente diferente de uno. Pero no resulta claro cómo debería entenderse la expresión «radicalmente diferente de uno». De un lado, debemos asumir que el otro es suficientemente similar a uno para poder considerarlo «otro», sino no lo consideraríamos siquiera un agente intencional, es decir, pasaríamos de largo sin reconocerlo como un agente. Si comprender algo es redescribirlo de una manera más familiar para capturar algo de su subjetividad, solo podemos comprender lo que suponemos que ya es suficientemente familiar como para que

pueda ser redescrito en nuestros propios términos. Lear sostiene que el principio de caridad no permite la diferencia radical, lo que es paradójico si se considera que el principio de caridad surgió, precisamente, a partir del análisis de la interpretación radical. Pero, en efecto, una tesis obvia que se sigue de ese principio es que no podríamos entender algo que sea totalmente diferente de nosotros. Si el comportamiento de una criatura fuese diferente de cualquier cosa que pudiéramos encontrar familiar no estaríamos en condiciones de atribuirle estados mentales ni acciones y, entonces, solo podríamos verlo como un conjunto de eventos físicos. Como dice Michael Root, «describir de una persona aquello que la hace radicalmente otra es ofrecer razones para pensar que no es otra después de todo» (1992, p. 300).

Pero, en todo caso, la diferencia es un asunto de grado. En el caso más extremo, lo que desde otro punto de vista es totalmente diferente no puede ser considerado desde nuestro propio punto de vista «otro». Aunque hay, por supuesto, casos menos extremos. Mientras más difícil sea para la intérprete proyectar sus estados mentales en el agente y simular ser él, más diferente de ella lo reconocerá. En tanto ella pueda proyectar sus estados mentales y pueda simular ser él, lo encontrará suficientemente similar como para considerarlo interpretable. Cuando Colón vio a los primeros nativos de lo que hoy es Centroamérica, pudo reconocerlos como comunicándose mutuamente, poseyendo una sociedad organizada, disfrutando algunas cosas y padeciendo otras y deseando realizar ciertas acciones en función a ciertas creencias. Esto fue suficiente para considerarlos «otros», es decir, suficientemente semejantes para poder ser diferentes. Como es conocido, hubo un interesante debate en la España de la época de la Conquista sobre si los indios eran seres humanos o no, y sobre si tenían alma y podían ser cristianizados. En el fondo, los filósofos españoles estaban debatiendo si los indios eran suficientemente semejantes a ellos como para poder considerarlos «otros» o no (véase Hanke, 1974).

Sin embargo, en tanto la habilidad para reconocer al otro como similar a uno es un asunto de grado y depende de nuestras habilidades como intérpretes, la inteligibilidad de los otros también será un asunto de grado. La historia de la humanidad muestra múltiples ejemplos de cómo algunos grupos de personas no consideran a los otros como suficientemente similares a ellos como para desarrollar lazos de solidaridad o para estar éticamente concernidos por ellos. Los griegos, como es conocido, no creían que los bárbaros fuesen suficientemente similares a ellos como para considerarlos sus iguales, pero fueron transformados por estos (véase Momigliano, 1975 y Kristeva, 1991). No obstante, sí los estimaban suficientemente similares para considerarlos otros.

11.3. En qué sentido el principio de caridad es etnocéntrico

La pregunta ahora es si el principio de caridad involucra algún grado de etnocentrismo o egocentrismo. Toda interpretación está inevitablemente centrada en uno, en tanto siempre involucra una comparación entre el comportamiento y los estados mentales del otro y los propios, de suerte que es la intérprete quien proyecta sus propios estados mentales y simula ser el otro. Es un error pretender que uno comprende al otro solo si logra capturar lo que él «realmente» es, pues esto presupone que hay cierta realidad determinable sobre lo que él es y que solo capturando esa realidad llegará uno a comprenderlo. Esa idea también presupone que uno debería abandonar su propia identidad o perspectiva para adoptar la del otro. Pero si eso pudiera ocurrir no sería un ejemplo de comprensión sino, en todo caso, de difusión de identidad, en la línea de la patología descrita por Erikson (1968). Para comprender al otro requerimos conservar nuestra propia perspectiva y es imprescindible relacionar lo que percibimos de él con esa perspectiva. Por eso es que la comprensión es un fenómeno necesariamente relacional, imperfecto, limitado e interminable.

Pero hay diferentes formas en que se puede ser etnocéntrico. Desde el desarrollo de las ciencias sociales en el siglo XIX, hay una intensa discusión acerca de las maneras de evitar el etnocentrismo en la interpretación de las culturas (Winch, 1964, p. 319). Con frecuencia los participantes en el debate asumen que ellos no están comprometidos con posiciones etnocéntricas, pero sus adversarios sí. Como mencioné en el capítulo anterior, sospecho que hay una forma peligrosa e indeseable de etnocentrismo, así como otra inofensiva e incluso moralmente deseable. De un lado, una sociedad es etnocéntrica cuando se considera axiológica o epistemológicamente superior a las otras. Ese es el sentido indeseable de etnocentrismo. De otro lado, hay un sentido en que todas las culturas y prácticas interpretativas son etnocéntricas y no podrían dejar de serlo: todas las sociedades consideran que las otras son suficientemente similares a ellas para poder ser comprendidas, si pueden llegar a ser consideradas sociedades (véase Rorty, 1991, p. 31). Pero, además, todas las sociedades consideran que sus creencias son las verdaderas y sus valores morales son los correctos. Esto es inevitable y sería contradictorio que uno creyese que sus creencias no son verdaderas y que sus valores no son valiosos. Lo que nos inmuniza de ser etnocéntricos en el sentido indeseable, es aceptar que nuestras creencias podrían ser falsas y que nuestros valores podrían no ser valiosos. Eso es lo que distingue a un falibilista de un fundamentalista, pues este último cree que hay creencias y valores fundamentales que él tiene y que no podrían estar equivocados, mientras el primero asume que cualquiera de sus creencias podría ser falsa —incluso esta misma— y que, por tanto, todas ellas deben ser permanentemente revisadas.

Cuando los primeros conquistadores españoles llegaron a América, su interacción con los pueblos locales fue un juego de múltiples interpretaciones en que ambas culturas proyectaban sus propios sistemas de creencias en la otra, tratando de predecir su comportamiento y modificarlo, asumiéndolo suficientemente similar al de ellos. Ambos, nativos y españoles, proyectaron sus propios sistemas de estados mentales e hicieron ajustes cuando consideraban que sus atribuciones no permitían dar sentido al comportamiento foráneo. No solo tenían estados mentales acerca de los otros sino también acerca de los estados mentales de los otros.

Podría decirse que este caso es precisamente un contraejemplo a la tesis que pretendo defender, pues el encuentro entre nativos y españoles fue un notable caso de incomprensión antes que de comprensión, en tanto las creencias que ambas culturas tenían acerca de la otra no describían lo que los otros «realmente» eran. La respuesta a esta objeción es que podemos juzgar que la interacción entre dos individuos o grupos involucra incomprensión o malentendido solo desde una tercera perspectiva que tiene creencias acerca de los dos sistemas de estados mentales en interacción. Este sería el caso, por ejemplo, si nosotros desde el siglo XXI, y con suficiente información empírica, interpretamos a nativos y españoles, y concluimos que estos se malinterpretaron porque no llegaron a capturar las intenciones y creencias que nosotros creemos que cada uno de ellos tenía. Esto es, por supuesto, posible, pero esa tercera perspectiva tendría las mismas características y limitaciones que cualquier otra práctica interpretativa, con lo cual se vería obligada a aplicar el principio de caridad a las otras dos perspectivas, mostrando el marco general de acuerdo que haría posible el desacuerdo y la malinterpretación. Así, por ejemplo, tendríamos que atribuir a Francisco Pizarro creencias acerca de las creencias de Atahualpa y a este, creencias acerca de las creencias de Pizarro —en ambos casos en varios niveles de intencionalidad—, y a ambos tendríamos que aplicar el principio de caridad. Así, la atribución de incomprensión y malentendido a dos perspectivas solo es posible al interior de un marco mayor de intenciones y creencias verdaderas compartidas, ciertamente desde nuestra propia perspectiva. Dadas las exigencias del principio de caridad, para interpretar a dos interlocutores como malentendiéndose sistemáticamente es necesario atribuirles un marco mayor de intenciones y creencias verdaderas compartidas, con lo cual hacemos el malentendido inteligible para nosotros. La pregunta sobre si Pizarro comprendió a Atahualpa y viceversa debe ser formulada en términos de un tercer manual de interpretación, que en este caso es nuestra perspectiva del siglo XXI.

Así el principio de caridad está comprometido con una débil forma de etnocentrismo que es inevitable y deseable, porque nos permite desarrollar actitudes morales y compromisos para con los otros. La única manera de sentir solidaridad por ellos

es asumir que, en un sentido importante, son semejantes a nosotros. Sería difícil desarrollar actitudes morales y sentir solidaridad por una criatura que consideramos tan diferente de nosotros que ni siquiera podemos comprender su sufrimiento, sus problemas o sus necesidades.

Lo que también nos previene de sucumbir a una forma indeseable de etnocentrismo es reconocer que la interpretación es un proceso dinámico que se desarrolla con la incorporación de elementos del comportamiento extraño dentro de los bordes familiares. Cuando la intérprete proyecte sus estados mentales en el otro probablemente descubrirá que el comportamiento del agente no llega a satisfacer sus expectativas originales, con lo cual se verá obligada a modificar sus atribuciones simulando ser él en circunstancias diferentes. En otras palabras, ella deberá alejarse de sus propios estados mentales para atribuir al agente estados mentales muy diferentes de los suyos. Tendrá que imaginar ser él en circunstancias poco familiares. Las atribuciones exitosas, es decir, aquellas que le permitan hacer inteligible su comportamiento poco familiar, conducirán a la intérprete a producir mejores simulaciones para casos más difíciles. Al tomar conciencia de cuánto tuvo que alejarse de sus originales proyecciones y simulaciones para producir nuevas atribuciones que puedan afrontar exitosamente su nuevo comportamiento, ella percibirá cuán diferente es él de ella y de lo que ella originalmente supuso. De esta manera, la intérprete constatará que no está interpretando a un producto de su imaginación sino a una persona muy diferente. Por tanto, constituiría un error pensar que el principio de caridad no permitiría a la intérprete reconocer las diferencias del agente.

Sería posible objetar a lo que estoy diciendo que el agente podría ser totalmente diferente de nosotros sin que podamos reconocer su diferencia, precisamente porque tendemos a interpretarlo siguiendo nuestros propios estados mentales, es decir, imponiendo nuestros propios rasgos en vez de descubrir los de él. Pero los conceptos de diferencia y similitud son siempre relativos a ciertos estándares de comparación. Así, si el objetor sostuviera que el agente podría ser totalmente diferente de nosotros sin que pudiéramos notarlo, el objetor estaría diciendo que para ciertos estándares —los nuestros, digamos—, él no es radicalmente diferente de nosotros, aunque para otros estándares —los de una tercera perspectiva—, él es muy diferente. Pienso que este sería un típico caso de indeterminación de la interpretación. Sin embargo, ya que la tercera perspectiva tendría que aplicar el principio de caridad como cualquier otro intérprete, esta no podría encontrar al agente totalmente diferente de ella misma.

También podría objetarse a lo que estoy diciendo que el principio de caridad no permite la posibilidad de reconocer que el otro sea tan diferente de la intérprete que él no pueda ser interpretado con los instrumentos que ella tiene, lo que conduciría

a su ininterpretabilidad de principio. Este problema se hace más manejable si nos movemos en el terreno de la traducibilidad lingüística. Así, plantearemos la pregunta de la siguiente manera: ¿Sería posible que las oraciones del hablante sean por principio intraducibles a nuestro lenguaje? Una pregunta asociada es si podría existir una comunidad cuyo lenguaje fuese en principio, por la naturaleza de los pensamientos expresados y no como una dificultad empírica, intraducible a nuestro lenguaje. Algunas formas de relativismo cultural y lingüístico dependen de una tesis de intraducibilidad de principio. Este es el elemento principal, por ejemplo, de la tesis del relativismo lingüístico defendido por Sapir y Whorf, entre otros, y que ya discutí en el tercer capítulo. Ellos sostienen que las lenguas podrían expresar una cosmovisión que a veces no podría ser traducida ni expresada en otra, al punto que las personas que «pertenecen» a cierta lengua «viven», por así decirlo, en diferentes mundos. Si la hipótesis Sapir y Whorf, en su versión clásica, fuese correcta, el principio de caridad sería, en efecto, la forma más desnuda del tipo indeseable de etnocentrismo, en tanto sostendría que todo aquello que no pudiera ser traducible a nuestro lenguaje carecería de contenido. Para poder contestar a estas objeciones debemos separar delicadamente diferentes problemas.

Si la intérprete puede encontrar comportamiento lingüístico en una criatura, su lenguaje podrá en principio ser traducible al de ella. Este es el argumento: la intérprete es capaz de encontrar comportamiento lingüístico si puede reconocer el comportamiento como intencional y no solo como un conjunto de eventos naturales. Una acción intencional es un evento natural descrito como causado —por lo menos parcialmente— por un conjunto de estados mentales atribuidos al agente. Solo es posible atribuir estados mentales al interior de un sistema mayor de estados mentales integrados. De esta manera, si la intérprete ha ido tan lejos como para atribuirle un sistema de estados mentales integrados, es porque de hecho ya está interpretando al agente. Más aún, si la intérprete sabe que las acciones del agente son de tipo lingüístico, esto es, que siguen un patrón de prácticas sociales, es porque ella ya tiene cierta comprensión de las regularidades sociales compartidas que gobiernan tales prácticas sociales o, lo que es lo mismo, ella ya comparte parcialmente esas prácticas sociales y, por tanto, pertenece de alguna manera a esa forma de vida. Podría objetarse que la intérprete puede saber que las acciones del agente han sido causadas por estados mentales sin saber cuáles son estos. Pienso, sin embargo, que si la intérprete cree que está expuesta a acciones y no solo a eventos naturales, es porque ya ha descrito tales eventos como acciones intencionales y, entonces, ya ha atribuido estados mentales específicos al agente. Esto significa que ella no puede saber que las acciones han sido causadas por estados mentales si no tiene alguna idea de cuáles podrían ser estos.

El punto es que interpretar al agente no es descubrir un significado «real» escondido en sus palabras o en sus acciones, sino ser capaz de redescribir tales palabras y acciones en términos de los estados mentales de la intérprete. De esta manera, si ella puede encontrar conexiones entre lo que ella considera que él cree acerca del mundo que ambos comparten, sus deseos, y las prácticas sociales que gobiernan sus acciones, está en camino de interpretarlo. No necesariamente estará ya traduciendo sus proferencias verbales, pero le habrá atribuido un marco de estados mentales y prácticas sociales que serán el contexto necesario para la traducción de sus oraciones.

La intraducibilidad de principio sería posible solo si hubiera tal cosa como «los significados reales» escondidos en el lenguaje del hablante que la intérprete podría llegar a ser incapaz de expresar en su propio lenguaje. Pero no es necesario postular tales significados ocultos para que la intérprete pueda traducir las palabras del hablante; bastará con que ella pueda encontrar conexiones entre las creencias asumidas como verdaderas por el hablante y las creencias asumidas como verdaderas por ella misma. Esto es lo que se suele llamar una teoría del significado veritativo-condicional, en la que el significado de una oración está determinado por sus condiciones de verdad, es decir, por las circunstancias que la hacen verdadera o falsa. Así, conocer el significado de una oración no sería capturar cierto contenido escondido sino saber en qué circunstancias estaríamos dispuestos a proferir esta oración, es decir, a afirmar su verdad o a creer en ella. Dice Ramberg:

> [...] si somos extensionalistas acerca del significado, no podemos imaginar qué es que un lenguaje tenga un «interior» que se mantiene inaccesible incluso si nos las hemos arreglado para representar el «exterior» del lenguaje al emparejar las extensiones de las oraciones en la interpretación radical. Y esto significa que la idea de lenguajes intraducibles carece de sentido (1989, p. 120).

En efecto, que dos oraciones tengan las mismas condiciones de verdad, bajo cierta interpretación, es todo lo que se necesita para sostener que una es una buena traducción de la otra, como ya hemos visto que se encuentra formalizado en las oraciones-T.

Si la intérprete está en condiciones de encontrar correlaciones entre las oraciones del hablante y lo que ella considera que es el mundo compartido por ambos, entonces ella podrá encontrar correlaciones entre las condiciones de verdad de las oraciones del hablante y las suyas propias, es decir, oraciones-T. Esto es todo lo que la traducción requiere y mientras la intérprete pueda encontrar tales correlaciones no habrá lugar para la intraducibilidad. La única evidencia posible de intraducibilidad sería el encontrar patrones de conexiones regulares entre las proferencias del hablante y el mundo y, sin embargo, ser incapaces de encontrar conexiones entre las proferencias del hablante y las nuestras. Pero al encontrar conexiones entre el lenguaje

del hablante y el mundo, ya estamos atribuyendo valores de verdad a sus oraciones y, entonces, estamos encontrando conexiones entre su lenguaje y el nuestro. Por eso, la única evidencia posible de intraducibilidad sería al mismo tiempo evidencia de traducibilidad.

Pero, ¿qué pasaría si el lenguaje del hablante fuese tan sofisticado y complejo, y sus órganos sensoriales tan diferentes de los nuestros que, por principio y no *de facto*, no pudiéramos llegar a interpretarlo, dada la simplicidad de nuestro aparato interpretativo? Nicholas Rescher sugiere el siguiente experimento mental.

> Imagínese criaturas inteligentes y activamente inquisidoras (animales, digamos, o seres del espacio exterior) cuyos modos de experiencia son muy diferentes a los nuestros. Sus sentidos son altamente susceptibles a diferentes parámetros físicos, mientras que son relativamente insensibles, digamos, al calor o a la luz, pero sustancialmente sensibles a varios fenómenos electromagnéticos. Tales criaturas inteligentes [...], sería plausible suponer, operarían al interior de una estructura de conceptos y categorías muy diferentes (1980. p. 323).

El argumento de Rescher es que, en este caso, estas criaturas tendrían un lenguaje que sería intraducible al nuestro, no como un problema de dificultad empírica sino como un asunto de principio, dadas las diferencias en conceptos, categorías y aparatos sensoriales. Esto constituiría un contraejemplo al modelo de la interpretación basado en el principio de caridad. Contestaremos a esta objeción recordando que, si podemos encontrar suficientes patrones de comportamiento intencional social y compartido en la comunidad en cuestión como para saber que estamos expuestos a un tipo de lenguaje, estaremos ya atribuyendo a tales criaturas creencias, propósitos y otros estados mentales. La frontera entre lo que consideramos ajeno y lo que consideramos familiar está siempre moviéndose en la medida en que nos las arreglemos para redescribir aquello que consideramos ajeno y, así, lo incorporamos dentro de los límites de lo que consideramos familiar. Malpas lo pone de esta manera:

> No podemos confiar en el supuesto de que la frontera entre lo ajeno y lo familiar, entre lo que podemos y lo que no podemos entender, es clara y rápida. Esta cambia como cambia la interpretación. Así, la situación en que intentamos traducir de una lengua extraña —incluso de la lengua de criaturas con capacidades sensoriales diferentes a las nuestras— presenta los mismos problemas de interpretación radical que nuestros esfuerzos de traducción e interpretación más cercanos a casa. Es el mismo problema ya sea si interpretamos a extraños o a nosotros mismos (1989, p. 260).

Es comprensible, sin embargo, que el principio de caridad haya sido acusado varias veces de estar comprometido con un etnocentrismo indeseable, ya que algunos de sus defensores suelen presentarlo de maneras que pueden confundir. Véase por ejemplo:

> Si no podemos encontrar una manera de interpretar las proferencias y el comportamiento de una criatura como revelando un conjunto de creencias básicamente consistente y verdadero según nuestros propios estándares, no tenemos ninguna razón para considerar a esa criatura como racional, como teniendo creencias o como diciendo cualquier cosa en absoluto (Davidson, 1984e [1973], p. 137).

El principio de caridad no debe implicar que la intérprete confiere, de manera condescendiente, el atributo de ser racional o de hablar una lengua a una criatura si ella puede hacerlo inteligible, como erróneamente podría sugerir la misma palabra «caridad». Lo que este principio dice es que si podemos reconocer suficiente comportamiento intencional en una criatura es porque ya le estamos atribuyendo parte de nuestro sistema de estados mentales y, así, estamos en camino de entenderla. De aquí se sigue que hemos encontrado que tal criatura es suficientemente parecida a nosotros como para que podamos decir que es racional.

Sería imposible reconocer comportamiento intencional en una criatura y luego decir que no podemos atribuirle estados mentales. También sería imposible atribuirle estados mentales y luego afirmar que es totalmente irracional. Tampoco sería posible encontrarla plenamente irracional y señalar que la comprendemos. Dice Malpas: «La idea de "un grupo de gente al que no podemos dar sentido" es simplemente la idea de un grupo de "gente" al que no podemos dar sentido como personas» (1989, p. 240).

Esta no es una forma indeseable de etnocentrismo, es simplemente una consecuencia de describir las características de algunos de nuestros conceptos. Nuestro concepto de acción involucra intencionalidad, racionalidad, estados mentales, etcétera, de una forma integrada en la que si hay uno de estos elementos los otros también deben estar presentes.

Como resultará claro, no estoy sugiriendo necesariamente que todos los lenguajes compartan la misma estructura lógica o esquema conceptual que los haría traducibles. Eso podría ser cierto. Es más, es altamente probable, dadas las características de la evolución de la facultad del lenguaje en nuestra especie, pero no es una tesis con la que el modelo de interpretación que aquí defiendo deba estar comprometido. Sostengo, más bien, que podemos traducir una lengua dada a la nuestra porque podemos proyectar la estructura lógica y categorial de la nuestra en la del otro. Al usar nuestro esquema conceptual para interpretar un lenguaje extraño, estamos ya encontrando tal esquema en él. Así no es que las lenguas sean traducibles porque haya ciertos contenidos significativos universales determinados que pueden ser capturados

y expresados en todas ellas. Es, más bien, que el significado es atribuido a un hablante o lengua ajena si somos capaces de encontrar condiciones de verdad en sus oraciones. Eso también ocurre en toda situación interpretativa, tanto en el nivel intercultural como interpersonal.

Al interpretar al otro, la intérprete no es una observadora no comprometida, sino una activa participante que asume que comparte significados, estados mentales y el mundo objetivo con sus interlocutores. En la interpretación continuamos creando un espacio compartido que asumimos tener: una comunidad de creencias, deseos, significados y los objetos del mundo. Como hemos visto, este es un punto importante en el que convergen la tradición hermenéutica alemana y la anglosajona: para Gadamer (1977b, parte II, capítulo 4) el paradigma de la comprensión es el juego, en tanto este existe solo porque hay individuos comprometidos en él. Para Davidson (2005a [1986]) el significado se constituye en la interacción de las diversas hipótesis interpretativas de los interlocutores. Para Wittgenstein (1988) la interpretación se da en una forma de vida constituida intersubjetivamente, aunque previa a cada individuo en particular. Al adscribirle algunos de nuestros estados mentales al agente y al simular los de él participamos de su perspectiva y compartimos parte de su espacio personal. Si él hace lo mismo con nosotros, logramos crear un espacio compartido. La noción clave es la de un espacio común que construimos y que no es necesariamente previo a la interpretación, sino producido y desarrollado durante la interpretación. Un ejemplo específico de la creación de este espacio común es el uso de palabras y de otras acciones intencionales de maneras que ambos interlocutores consideran serán comprendidas por el otro de manera correcta, aunque claramente se aleje de la norma.

Al interpretar al otro, la intérprete no solo mejora su comprensión de él sino también su comprensión de sí misma, su autodescripción. Este proceso debería prevenir cualquier elemento de dogmático etnocentrismo. Al tratar de hacer inteligible al otro, ella comparará sus propias creencias y deseos con aquellos atribuidos al agente, proceso que la obligará a poner en cuestión su propio sistema de creencias y deseos. Si ella ya es suficientemente amplia de mente como para intentar simular ser el otro, incluso si sus estados mentales y circunstancias son muy diferentes, probablemente sentirá la necesidad de poner en cuestión sus propios estados mentales. Esto dará lugar a una exigencia de adaptación y cambio, que podría ayudar a crear un territorio común entre la intérprete y el agente. Con suerte y buena voluntad, allí la comunicación podrá discurrir.

Pero, por supuesto, la irracionalidad existe, aunque solo podamos notarla frente a un marco mayor de racionalidad. Este es el momento, por tanto, de analizar cómo es posible la irracionalidad y qué características tiene.

Capítulo doce
¿Qué es la irracionalidad?

12.1. Algunas confusiones sobre la irracionalidad

La irracionalidad es un fenómeno peculiar. Por una parte, decir que alguien es irracional es casi lo mismo que afirmar que no lo podemos comprender, pues si lo comprendiéramos sabríamos por qué hace lo que hace o por qué tiene los estados mentales que tiene, de manera que no nos parecería sorprendente.

De otro lado, sostener que alguien es racional equivale a decir que nos parece razonable, esto es, que conocemos las razones que gobiernan su vida mental y sus acciones. Por eso, en un importante sentido, llamarlo irracional o poco razonable es solo expresar que no conocemos esas razones. Pero, podría ocurrir que el agente sí tenga razones para actuar como actúa o para tener los estados mentales que tiene, solo que nosotros no las conocemos. Es más, sería imposible que esas razones no existieran. Así como es condición de posibilidad para explicar el universo físico asumir que este es uniforme o altamente probable y no aleatorio, también es condición de posibilidad para interpretar a un agente intencional asumir que sus estados mentales y acciones tienen causas y que, en general, tales causas también son razones que las justifican. La idea de que, en general, las causas de nuestro comportamiento también son las razones que las justifican pertenece a Davidson (1980a [1963], 1980d [1970], 2004c [1982]) y, tal como lo discutimos en el primer capítulo, es fundamental para explicar y comprender causalmente el comportamiento de los agentes intencionales. ¿Es que la irracionalidad no existe, entonces? Sí existe, pero solo cuando el agente actúa o tiene estados mentales en contra de sus propias razones sin saber por qué lo hace. Eso ocurre, por ejemplo, cuando uno tiene la compulsión a actuar boicoteándose o cuando cree lo que sabe que es falso. Sobre esto volveremos en un momento.

Para que haya irracionalidad, el agente irracional debe actuar en contra de su principio de segundo grado —o su metadeseo— según el cual uno debe actuar de acuerdo con lo que considera que es lo mejor, según sus propios criterios respecto de lo que es lo mejor. Por tanto, para que una acción sea considerada irracional, es necesario que esta sea juzgada de esa manera por la intérprete y también por el agente, lo que ocurre solo si él actúa en contra de su mejor juicio, es decir, en contra de su metadeseo de actuar siempre según lo que él considera que es lo mejor. En estos casos, cuando lo irracional es una acción hablamos de irracionalidad práctica, pero si lo irracional es una creencia, lo que tenemos es irracionalidad teórica.

Por otra parte, incluso cuando atribuimos elementos de irracionalidad a alguien lo hacemos con el objetivo mayor de comprenderlo, es decir, de entender que si se comportó de esa manera es porque a veces es irracional, y adscribimos esa irracionalidad excepcional y localizada para poder seguir viéndolo como un agente dotado de intencionalidad y no solo como un conjunto de eventos naturales.

Recordemos que para reconocer acciones intencionales en el comportamiento de alguien —y no solo eventos físicos— es necesario ver ese comportamiento como, en general, causado y justificado por sus propios estados mentales. Pero si notamos irracionalidad en ese individuo —de manera localizada, como es inevitable— no encontraremos conexión entre algunos de sus estados mentales y sus acciones, o entre algunos de sus estados mentales entre sí. ¿Cómo sabremos, entonces, si lo que ocurre es que tiene momentos en que es poco razonable o es más bien que nosotros no lo podemos comprender porque somos malos intérpretes? ¿Cómo saber si el problema es él o nosotros?

La irracionalidad es paradójica porque la atribuimos precisamente para ampliar el ámbito de la racionalidad, es decir, para poder explicar racionalmente, y por tanto para poder comprender lo que antes no podíamos ni explicar ni comprender. Así pues, decimos que alguien se comportó localmente de manera irracional —o que algunos de sus estados mentales son irracionales— para seguir encontrándolo *grosso modo* racional.

El maestro en el arte de atribuir irracionalidad para explicar y comprender de manera racional fue Freud (2002), por eso es que amplió el ámbito de lo racional. Lo hizo en dos sentidos diferentes pero complementarios. Por una parte, logró explicar racionalmente fenómenos que hasta entonces eran considerados ininteligibles y que, por tanto, eran confusamente asumidos como irracionales, como gran parte de la sintomatología neurótica. Al explicar esos fenómenos racionalmente, es decir sobre la base de razones y causas, dejaron de ser considerados irracionales y fueron comprendidos como parte de la complejidad del comportamiento humano. De otro lado, Freud fue parte de un proceso histórico en el que participaron muchos filósofos de los siglos XIX y XX, que implicó superar una concepción demasiado

estrecha de lo racional. Al hacerlo resignificó este término y mostró cómo la racionalidad tiene muchas más facetas de las que solía ver la tradición filosófica moderna. Lejos de implicar un abandono de la racionalidad, el proyecto freudiano nos permitió tener una concepción más amplia y completa de ella.

En las discusiones teóricas contemporáneas, sin embargo, tanto en filosofía como en psicoanálisis, hay pocos conceptos que soporten más diversidad de interpretaciones y malentendidos que los relacionados con la racionalidad y la irracionalidad. Me propongo en este capítulo colaborar en su explicación para tratar de reducir las confusiones en las que estos conceptos se hallan sumidos. Comenzaré con una breve y panorámica contextualización histórica para luego concentrarme en el problema mismo.

En primer lugar, es claro que las palabras «racional» e «irracional» tienen muchos significados y que nadie debería pretender que alguno de ellos sea el correcto o el privilegiado. Pero uno puede argumentar que algunos de estos usos son preferibles a otros en tanto pueden resultar más explicativos o esclarecedores, así como también podría afirmar que otros usos contienen presupuestos injustificados, confusos o simplemente engañosos. Así, intentaré analizar cuáles sentidos de los empleados pueden ser explicativos y cuáles no. Para ser más claro, me detendré brevemente en mostrar las que considero son las concepciones más importantes de racionalidad en la tradición filosófica, a las cuales llamaré la concepción moderna, la culturalista y la formal.

Como hemos visto en el capítulo diez, el pensamiento moderno, heredero de la tradición clásica griega, entiende la razón en dos sentidos principales: uno predicativo y el otro sustantivo. El primero se entiende como la justificación a partir de razones de una creencia o una acción —o de un estado mental en general—, con lo cual podemos decir que es racional creer cierta proposición, desear algo, tener cierta emoción[1] o realizar determinada acción. El segundo sentido es entendido como la capacidad humana que nos permite hacer esas justificaciones y da, a su vez, razón de ellas. Según este modelo, esa supuesta facultad de dar razones —conformada por un conjunto de reglas inferenciales y contenidos cognitivos— sería universal, probablemente innata y constituiría propiamente el rasgo esencial del ser humano.

[1] Puede resultar extraño afirmar que las emociones pueden ser racionales o irracionales, pero no lo es. Las emociones están estructuradas sobre la base de creencias, de la misma manera como las creencias suelen incorporar un contenido afectivo. Una emoción irracional sería aquella que no está justificada sobre la base de las creencias que incorpora. Así, por ejemplo, si yo estoy indignado con una persona porque creo que me ha hecho cierto daño, pero se me demuestra —y lo creo— que ella es inocente de ese cargo, sería irracional que mantenga mi indignación. Si la mantengo, sin embargo, es porque la causa de esa emoción es distinta de la que yo creo que es. Ese sería un caso de irracionalidad. Por otra parte, sería perfectamente racional que uno se indigne frente a un daño real y deliberado de alguien, lo sorprendente —y hasta quizá irracional— sería que eso no ocurriera.

Por eso, en gran medida, el proyecto de la modernidad y de muchos autores contemporáneos consistió, y aun consiste, en la búsqueda de esa capacidad que, a su vez, se convirtió en criterio de demarcación entre lo humano y lo no humano.

Es importante aclarar que, aunque esta noción de racionalidad característica de la modernidad presente algunos problemas, no hay que asumir acríticamente que deba ser abandonada sino, en todo caso, podemos suponer que debería ser modificada. Podría ser que, en efecto, haya principios inferenciales y cognitivos universales e innatos, y que tenga sentido denominarlos constitutivos de la racionalidad; pero seguramente hoy los explicaríamos como producto de la adaptación del cerebro humano al medio —a lo largo de varios millones de años de evolución— y no como una esencia intemporal de lo humano. En todo caso, según esa concepción moderna, lo irracional sería lo injustificable sobre la base de esos criterios objetivos y universales de justificación. Eventualmente esa concepción permitió que se considerara irracional las creencias y los comportamientos de culturas distintas a la europea que no podían ser explicados fácilmente según los patrones occidentales, con lo cual se llamó también irracional a lo diferente y lo otro, así como a lo incomprensible e impredecible. Precisamente por ello, lo irracional terminó siendo asociado a lo amenazador e incontrolable, eventualmente incluso lo considerado inhumano y monstruoso.

Esta concepción de racionalidad comenzó a ser cuestionada desde que algunos filósofos recientes se preguntaron si esa supuesta universalidad no sería más bien una estrategia de justificación de formas de imposición cultural y política. De ser ese cuestionamiento correcto, afirmaron, lejos de haber una racionalidad universal lo que habría es una racionalidad histórica dominante que surgió en Europa hacia el siglo XVI, interesada en imponerse sobre las otras racionalidades y justificada bajo el argumento de una supuesta pero inexistente universalidad. Así surgió la pregunta de si realmente hay una racionalidad universal o si lo que hay es una pluralidad de racionalidades; y si la pretensión de que haya una racionalidad universal —es decir, contenidos cognitivos y criterios universales de justificación gracias a una facultad que todos compartiríamos—, no sería sino la pretensión de una de ellas por alzarse sobre las otras para someterlas. En efecto, con frecuencia Occidente se impuso sobre otras sociedades amparado en el falaz argumento de que llevaba una justificación racional que tendría como consecuencia la capacidad de civilizar a pueblos y culturas que no habían accedido, todavía, al pleno uso de esta racionalidad.

Con el idealismo lingüístico del siglo XIX y comienzos del XX, se sentó las bases para que posteriormente algunos autores sostuvieran que no hay una racionalidad sino varias: esta es la concepción culturalista. Así, se empezó a hablar de una racionalidad occidental diferente de otras racionalidades como, por ejemplo, orientales, africanas, andinas, amazónicas, etcétera. La irracionalidad siguió siendo entendida

como lo otro y lo diferente, pero se asumió que lo que es irracional para nosotros puede ser racional para otros y viceversa. La idea, en esta concepción, es que quienes pertenecen a otra cultura podrían «razonar» de una manera diferente, en el sentido en que les podrían parecer válidas ciertas inferencias que nos parecen inválidas a nosotros, o que ciertas creencias les podrían parecer «razonables» por ser parte del sentido común de su sociedad, aunque esas mismas creencias podrían ser consideradas absurdas por nosotros.

Es importante notar que aquí se está jugando con dos sentidos diferentes de racionalidad, uno que remite a la validez lógica de las inferencias —razonar— y otro que apela a los contenidos mismos de las creencias —ser razonable—. Respecto de lo primero habría que discutir si la validez de las inferencias lógicas es un asunto cultural, es decir, si el *modus ponens* es occidental y, por tanto, podría haber culturas que no lo empleen o que lo consideren inválido. Pero eso sería muy extraño; más factible sería que algunas reglas lógicas de inferencia y algunos principios lógicos —como, por ejemplo, el de no contradicción o el del tercio excluido— sean parte de las características formales del pensamiento humano y, por tanto, que hayan evolucionado con el cerebro para permitirnos una mejor adaptación al medio. Dicho de otra manera, los individuos que razonaban según el principio de no contradicción y el *modus ponens* sobrevivieron porque sus formas de razonamiento eran más exitosas para adaptarse a un mundo complejo, variado, cambiante y hostil. Los que no lo hicieron no dejaron descendencia. Los individuos que razonaban según alguna inferencia bastante más compleja perecieron antes de reproducirse, de manera que esa inferencia no se convirtió en una forma de razonamiento automática e intuitiva, como si lo hicieron el *modus ponens* y el principio de no contradicción. Sin embargo, como esa inferencia más compleja se deduce formalmente de los principios que sí se conformaron con la evolución del cerebro, son actualmente parte de nuestra lógica clásica. De lo que estoy diciendo podría seguirse que los principios lógicos pudieron haber evolucionado de una manera diferente, de haber evolucionado de forma diferente el cerebro humano y que, por tanto, son empíricos y contingentes. Esa sería una afirmación sumamente arriesgada y controversial sobre la que no entraré aquí, por no ser el lugar indicado, pero sin duda es un interesante tema de discusión.

En el caso de los animales es también debatible afirmar que razonen y que, de hacerlo, lo hagan según principios lógicos. Pero sí me parece claro que por momentos podemos interpretarlos como si lo hicieran. Si pensamos, por ejemplo, que un animal «cree» que su presa es vulnerable y que «desea» capturarla, esperaremos que «actúe» en consecuencia. Si no lo hiciera no lo entenderíamos o incluso lo consideraríamos «poco razonable». Seguramente pensaríamos que hay otra variable —desconocía por nosotros— que ese animal está «considerando». O que tiene una enfermedad

o una herida que no vemos. O que no tiene hambre. O alguna otra cosa. Solo en el caso extremo que hayamos repasado todas las variables y escenarios posibles podríamos concluir, con cierto humor y extravagancia, que el animal es «irracional». Pero, como en este caso será obvio, atribuirle a un animal irracionalidad es solo una manera de decir que no sabemos cómo explicar su conducta. Ya he señalado anteriormente que no tengo ningún inconveniente en aceptar que los animales —y los bebes muy pequeños— tienen estados mentales dotados de experiencia fenoménica —por ejemplo sensaciones, deseos y afectos—, pero tengo algunos reparos para aceptar que tengan estados mentales dotados de contenido proposicional como, por ejemplo, creencias. La razón de ello es que para tener creencias se necesita saber que las creencias podrían ser falsas, lo que implica tener estados mentales de por lo menos dos órdenes de intencionalidad; esto no está categóricamente probado en otras especies, aunque se cree que muchos primates no humanos sí podrían tenerlos y quizá también los perros, quienes han coevolucionado con los humanos por entre 15 000 y 35 000 años.

Pero volvamos al segundo sentido de racionalidad. Con frecuencia se confunde razonabilidad con verdad. Es decir, se confunde el que en algunas culturas una creencia resulte razonable con que los miembros de esa cultura tiendan a considerarla verdadera. Es importante separar aquí la forma del contenido. El no hacerlo podría conducir a trivializar el concepto mismo de racionalidad, porque el identificarlo con una visión del mundo o con un sistema de creencias puede conducir a aceptar que haya tantas racionalidades cuantas cosmovisiones sean posibles. Esto llevaría, a su vez, a que podamos seguir distinguiendo hasta el infinito entre interminables racionalidades.

Por ello, el tercer modelo de racionalidad podría ser denominado formal y está inspirado en Davidson (2004), aunque quizá él no aceptaría algunos de los detalles que defenderé aquí. Según esta concepción, lo que es racional o no —en una circunstancia dada— es un estado mental con contenido proposicional o una acción en particular, no un individuo en su totalidad. Así, una creencia, un deseo, una emoción o una acción son racionales si son compatibles con las otras creencias, deseos, afectos y acciones del agente, desde el punto de vista de la intérprete que hace las atribuciones.

La irracionalidad de un estado mental o acción sería, precisamente, su incompatibilidad con el marco mayor de estados mentales y acciones al que pertenece, desde el punto de vista de una determinada interpretación. Por ejemplo, si yo creo que p y también creo que p implica q, sería irracional si además creyera que no q, siempre que sea consciente de las dos proposiciones que creo y del *modus ponens* involucrado. O, si creo que la acción r es inconveniente y me resulta dañina, y no deseo realizar r, estaría siendo irracional si hiciera r, siempre que sea consciente de lo que creo y deseo.

Es claro, sin embargo, que con frecuencia actuamos de esa manera, con lo cual la irracionalidad requiere de una explicación. En general, es irracional actuar en contra de nuestro mejor juicio —sea este el que fuere— en contra de los que creemos son nuestras prioridades, o en contra de nuestros propios deseos, cualesquiera que estos sean. Así pues, la irracionalidad es una desconexión entre nuestros estados mentales y nuestras acciones conscientes; es una fractura al interior de nosotros mismos. Pero, como ya se ha visto, solo se puede ser irracional si uno es básicamente racional. En efecto, solo una criatura racional puede tener un estado mental o realizar una acción irracional, la cual lo es en relación al sistema de estados mentales y acciones más amplio del propio sujeto, el que es atribuido por una intérprete y que está en relación con el propio sistema de estados mentales de esa intérprete. Se verá, por tanto, que esta es una noción fuertemente relacional de racionalidad que emerge siempre en una situación comunicativa donde hay por lo menos un agente, una intérprete y el mundo objetivo que ambos comparten.

12.2. Explicando racionalmente la irracionalidad

Si consideramos irracional la inconsistencia que no podemos explicar dentro de nosotros mismos —por ejemplo, nuestra capacidad para autoengañarnos, nuestra compulsión a la repetición, nuestra tendencia a boicotear nuestros propios logros o nuestra forma de arreglárnoslas para hacer lo que no queremos hacer y además justificarnos por ello—, en el momento en que explicamos esta irracionalidad —por ejemplo, al reconocer que es el producto de un pasado reprimido que pugna por salir a la luz o de un aspecto de nosotros mismos que no conocemos— deja de ser irracional para ser racionalmente explicable, aunque no nos guste la explicación. Como muestra Davidson (2004c [1982]), la irracionalidad es paradójica porque al explicarla racionalmente encontramos las razones y causas que la hicieron posible, con lo que se disuelve.

Ahora bien, racionalidad no es lo mismo que conocimiento verdadero ni que acción moralmente correcta. Una creencia puede ser verdadera pero irracional o racional pero falsa. En el siglo VI a.C. era irracional, aunque verdadero, creer que la Tierra girara alrededor del Sol y era razonable aunque falso creer que la Tierra era el centro del universo. De igual manera, una acción puede ser irracional, aunque moralmente valiosa o malvada pero racional. Sin embargo, es importante atribuir racionalidad o irracionalidad en un espacio de razones suficientemente amplio. Por ejemplo, alguien podría tener razones para desear exterminar a una etnia y a primera vista parecería racional —aunque inmoral— que lo hiciera. Pero no es así, pues uno siempre podría preguntarle si también tendría razones para desear ese exterminio si él perteneciera a esa etnia. Como probablemente la respuesta sería que no, el agente evidenciaría su irracionalidad.

También es importante distinguir entre creencias mal justificadas e irracionales. Lo primero ocurre cuando uno acepta una creencia sin suficiente evidencia pero no tiene una mejor opción. Lo segundo tiene lugar cuando uno tiene una creencia que es inconsistente con las otras creencias de su sistema de creencias, de manera que sabe que está mal justificada pero por alguna razón que es importante buscar, sigue creyendo en ella.

Consideremos dos ejemplos: el sujeto A cree que p y tiene evidencia para creer que p. Pero la creencia en p le resulta muy dolorosa, de manera que la reprime —en un sentido psicodinámico—, con lo cual cree que ~p. En este caso, p causó ~p, pero obviamente no la justifica. Así, A cree que ~p, cree que cree que ~p y se comporta como si ~p, pero porque también cree que p. Entonces, cree que p y cree que ~p, aunque no cree la contradicción p & ~p.

De esta manera, A tiene razones para creer que ~p, pero como esta creencia es dolorosa o inconveniente —por ejemplo, no es aceptada por su sociedad—, cree que cree que p, que le resulta más conveniente. A cree que tiene razones para creer p, pero no es así; A está engañado y el autoengaño es un tipo de irracionalidad. Cuando A se da cuenta de que cree que p y que esa creencia causa, pero no justifica que ~p, se desvanece la irracionalidad y, además, logramos explicar racionalmente la irracionalidad. El punto es que siempre actuamos sobre la base de razones, incluso cuando somos irracionales, porque esas razones que causan —pero no justifican conscientemente— nuestras acciones, son inconscientes.

El segundo ejemplo es más familiar. Los seres humanos —por ejemplo, los agentes económicos— no siempre actúan sobre la base de lo que ellos mismos creen que maximiza su interés o bienestar ni tampoco actúan siempre maximizando su función de utilidad. ¿Por qué no lo hacen? Porque con frecuencia son irracionales, es decir, tienen estados mentales que causan pero no justifican sus acciones. En otras palabras, siempre actuamos tratando de maximizar nuestra lista de prioridades, pero podríamos tener prioridades inconscientes que causan y no justifican nuestras elecciones. No obstante, nuevamente, esto solo es posible desde el punto de vista de la intérprete que atribuye las creencias y reconoce las inconsistencias del agente —a partir de lo que ella cree que son los estándares del agente—, pues de otra manera solo habría discrepancia entre las creencias del agente y las de la intérprete.

Como hemos visto, lo que es racional o no es un estado mental o una acción, frente al marco de los otros estados mentales del agente. Sin embargo, de una manera laxa y coloquial, se puede decir de un agente que es racional si encontramos consistencia formal entre sus creencias, deseos y acciones, tal como son descritos por una intérprete. Pero esto podría descomponerse en los siguientes puntos. Un agente es racional desde el punto de vista de una intérprete si se cumplen las siguientes condiciones:

(1) Hay cierta coherencia básica entre sus estados mentales y acciones, lo que implica que actúa según sus creencias y deseos, es decir, según su mejor juicio.

(2) Actúa según lo que cree que es beneficioso para él y para quienes él desea beneficiar, según la lista de prioridades que tenga.

(3) Si prefiere A a B y B a C, entonces también prefiere A a C.

Pero solemos decir que un individuo es racional o irracional si tiene muchos estados mentales o acciones que son catalogados de una manera u otra. Y una acción o un estado mental es considerado racional si puede ser explicado mediante razones, es decir, si es posible encontrar las justificaciones que tuvo el agente para adoptar el estado mental que adoptó o para realizar la acción que realizó. El punto es que un estado mental y una acción son eventos que una intérprete puede adscribir a un agente solo si estos pueden ser interconectados en una red que sigue un patrón intencional coherente. Sin embargo, para que la intérprete pueda adscribir razones y racionalidad al interpretado, debe compartir con él un gran número de creencias y debe poder imaginar los estados mentales que ella cree que tendría si fuera el agente. De esta manera, cuando ella interpreta a alguien tiene que hacer tres cosas: debe atribuirle estados mentales, debe atribuir a sus proferencias significados y debe ser capaz de identificar su comportamiento como una sucesión de acciones intencionales. Estos tres elementos constituyen una red holista, por eso, el proceso de interpretación puede comenzar por cualquiera de los tres tipos de atribuciones, aunque la manera habitual de empezar el proceso es vía la identificación de las acciones del agente ya que estos son los fenómenos intersubjetivos más perspicuos.

Como ya hemos visto, las acciones son eventos físicos descritos intencionalmente. Pero ser intencional no es una propiedad de los eventos en sí mismos, sino una propiedad de los eventos tal como son descritos por una intérprete en particular en cierto vocabulario, que incluye un propósito por realizar en el mundo. Bajo esta descripción podemos mostrar que las acciones son eventos físicos realizados por los agentes siguiendo razones. De esta manera, podemos relacionar las acciones con las creencias y los deseos del agente, y entonces atribuiremos al agente razones para realizar las acciones en cuestión. Así, habremos explicado causalmente las acciones por medio de razones.

Ahora volvamos a la irracionalidad. Como se ha visto, esta es una particular desconexión entre los estados mentales y las acciones del agente o, mejor, entre los estados mentales y las acciones que una intérprete atribuye a un agente desde cierta interpretación que ella hace de él. Por tanto es una particular desconexión entre el sistema de creencias que la intérprete atribuye al agente y el comportamiento del agente tal como ella lo interpreta.

Creemos que alguien es irracional cuando actúa de manera contraria a sus propios estados mentales, es decir, a los que le hemos atribuido. Pero aquí surge un problema: si creemos que las acciones de los individuos han sido causadas por sus estados mentales —de hecho, para poder siquiera comenzar a interpretar a alguien debemos encontrar consistencia entre sus estados mentales y sus acciones—, ¿cómo podríamos explicar que en algunos casos haya acciones no causadas por los estados mentales que le hemos atribuido? La propuesta de Davidson (2004b [1985], 2004a [1986], 2004e [1997b]) es que la irracionalidad se explica como un tipo de división de la mente en que ciertos estados mentales —inconscientes— son causa, pero no razón consciente para la acción, aunque puedan ser la razón inconsciente de la acción. En otras palabras, en el caso del comportamiento irracional, nuestros estados mentales inconscientes causan, pero no justifican mediante razones conscientes, nuestras acciones. Esto ocurre hasta que encontramos las razones inconscientes que causan las acciones, así como las razones por las que esos estados mentales son inconscientes, con lo cual logramos explicar las acciones previamente irracionales mediante razones, de manera que la irracionalidad termina por desvanecerse y ceder paso a la racionalidad.

Es importante subrayar nuevamente que atribuimos irracionalidad a los agentes cuando esta es la única manera de poder hacer inteligible su comportamiento, de manera que lo hacemos para mantener la racionalidad. Pero aquí parecería que la irracionalidad es un tipo de ininteligibilidad, porque llamamos a alguien irracional cuando no nos explicamos cómo, considerando los estados mentales que nosotros creemos que él tiene, pudo haber actuado de la manera como lo hizo o pudo tener los otros estados mentales que creemos que tiene. En el caso extremo en que nos resulte imposible encontrar alguna conexión entre sus estados mentales y sus acciones, progresivamente veremos que el contenido de tales estados mentales y acciones empieza a desvanecerse, porque no sabríamos qué estados mentales o acciones atribuir. En ese caso, la atribución de irracionalidad desaparece para dar lugar a la simple ininteligibilidad. Ya no es que sepamos que está comportándose irracionalmente, es que ni siquiera sabríamos si se está comportando de alguna manera.

Por eso, solo podemos atribuir irracionalidad contra un fondo mayor de racionalidad, allí donde la acción irracional es el efecto de una causa que no es una razón consciente. Los estados mentales que son causa, pero no razón consciente, del comportamiento del agente podrían ser, por ejemplo, creencias inconscientes, es decir, disposiciones para actuar cuyo agente desconoce. Estas gobernarían la conducta del agente transitoriamente por razones que el propio agente tanto como la intérprete desconocen.

Es imposible querer siempre ser inconsistente, porque esto haría imposible la acción misma. Más aún, es imposible no buscar consistencia en el agente para poder

comenzar a interpretarlo. Además, no es que la inconsistencia en sí misma sea irracional, es que llamamos irracional a un estado mental o a una acción que tenemos razones para atribuir a un agente pero que resulta inconsistente con el sistema mayor al que pertenece y que, por tanto, nos sorprende con su ininteligibilidad. Como hemos visto, podría ocurrir que el error sea nuestro al haber atribuido estados mentales o acciones inapropiadas al agente, que no solo no explican su comportamiento sino que lo distorsionan. De hecho, una intérprete creativa deberá estar siempre dispuesta a hacer modificaciones en las atribuciones realizadas al interpretado para hacerlo más inteligible. Como resultará claro, hay un sentido en que los estados mentales que la intérprete atribuye al interpretado son los estados mentales del interpretado desde el punto de vista de la intérprete. Esto es inevitable, pero no implica que los estados mentales del interpretado no sean reales o que no sean los de él, simplemente muestra que están subdeterminados por las diversas interpretaciones que se puede hacer de él.

Las creencias, deseos y acciones de un individuo son inconsistentes cuando parecen dirigirse a finalidades incompatibles entre sí. Sin embargo, si bien ser capaz de encontrar estados mentales consistentes en el comportamiento del agente es condición de posibilidad de la interpretación y es un objetivo valioso, la plena consistencia es imposible y no siempre es una virtud. Para poder cambiar, es decir, para estar dispuestos a mejorar algunos aspectos de nosotros mismos, es necesario reconocer cierto grado de inconsistencia, por lo menos en el momento del tránsito. La creatividad, que es el paradigma del cambio, es también una forma de inconsistencia, porque ser creativo es estar dispuesto a transgredir ciertas regularidades para sustituirlas por otras. Finalmente, ser absolutamente consistente puede conducir a la incapacidad de entender la diferencia, es decir, de aceptar la posibilidad de que haya un estado mental o una acción que desafíe nuestros prejuicios. En esos casos, el comportamiento diferente nos parece en un primer momento ininteligible o absurdo, pero en un segundo momento, después de cierta reflexión y cambio de nuestra parte, no solo puede llegar a parecernos razonable sino incluso iluminador. Las relaciones causales entre subsistemas pueden originar conducta acrática, pero también pueden permitir la autocrítica y el autocuestionamiento. Cuando uno considera necesario examinar algunos de sus estados mentales, los elementos que impulsan el cuestionamiento no suelen provenir del mismo subsistema que se quiere cuestionar. Normalmente provienen de otro. Así pues, si tenemos motivaciones para desear cambiar algunas de nuestras creencias y deseos, esas motivaciones suelen proceder de otros subsistemas. Dicho de otra manera: los cambios en un sistema de estados mentales suelen requerir de algún tipo de inconsistencia interna que sea causa del cambio. Un sistema plenamente coherente podría ser incapaz de cambiar, podría incluso ser incapaz de imaginar la posibilidad de necesitar un cambio. El agente puede tener razones para

cambiar sus propios hábitos y carácter, pero tales razones provendrán de un dominio de valores extrínseco al contenido del punto de vista o los valores que requieren cambio. Por eso, Davidson termina «Paradoxes of Irrationality» con la afirmación de que «una teoría que no pudiera explicar la irracionalidad sería una que tampoco podría explicar nuestros saludables esfuerzos, y ocasionales éxitos, en la autocrítica, y el autoperfeccionamiento» (2004c [1982], p. 187).

Un principio fundamental del holismo de lo mental es que solo es posible atribuir a un individuo un estado mental o una acción si estos son consistentes con el sistema al que se atribuirá, dado que su contenido procede de sus relaciones con los otros estados mentales. Por ello, sería imposible tener un estado mental desconectado del sistema al que es atribuido. Pero, si esto es así, ¿cómo sería posible la atribución y la existencia misma de un estado mental o de una acción irracional?

La explicación de Davidson y la de Freud nos remiten a la tesis de la división de la mente: se trata de subconjuntos de sistemas de estados mentales que internamente son básicamente consistentes pero que son inconsistentes entre sí. Dicho de otra manera, con frecuencia no nos queda más remedio que atribuir diversos subconjuntos de estados mentales, inconsistentes entre sí, pero básicamente consistentes internamente, para hacer de alguna manera inteligible al agente, es decir, para hacer inteligible su misma irracionalidad y, con ello, su comportamiento.

La idea aquí es que la mente contiene un número de estructuras semiindependientes. Estas subestructuras están dotadas de estados mentales e interactúan entre sí para producir eventos internos y externos a la mente. Algunas de las disposiciones que caracterizan a las diversas subestructuras de la mente pueden ser vistas bajo el modelo de disposiciones físicas y fuerzas, cuando afectan o son afectadas por otras subestructuras de la mente. ¿Pero de qué naturaleza serían estos eventos que, siendo causa de acciones, no son razones conscientes? Podrían ser de diversos tipos: podrían ser creencias inconscientes —disposiciones para actuar que uno no sabe que tiene— o incluso podrían ser disposiciones físicas. Podríamos, entonces, ponerlo de esta manera: nuestra conducta está inevitablemente moldeada por causas. Sin embargo, es posible describir estas relaciones causales de múltiples formas: podemos describirlas utilizando un vocabulario intencional, con lo cual estas relaciones causales van a ser vistas como estados mentales. O podemos describirlas en términos estrictamente físicos, con lo cual tendremos procesos fisiológicos, interconexiones neuronales, pulsiones, instintos, etcétera. Solo los eventos descritos intencionalmente pueden llegar a ser razones —no todos, sin embargo—, pero los eventos descritos físicamente no pueden jamás ser razones. Entonces, la irracionalidad se produce cuando las causas de nuestras acciones son eventos intencionales que no son razones para actuar sino más bien contradicen nuestras razones conscientes.

Por supuesto, hay muchas otras formas en que eventos mentales pueden causar otros eventos del mismo tipo al interior de la mente de uno mismo, pero el caso paradigmático es la interacción social. La intersubjetividad y la comunicación no son otra cosa que el fenómeno de producir en otra persona ciertos efectos deseados, así como en algunos casos, también efectos indeseados.

Con frecuencia se confunde irracionalidad con no racionalidad. Lo primero, como hemos visto, es una fractura al interior de un modelo de interpretación que una intérprete hace de un agente. Lo segundo, por el contrario, alude a eventos de la naturaleza que carecen de objetivo y finalidad, con lo cual no son ni racionales ni irracionales. Todo comportamiento intencional y toda interpretación se asientan sobre bases no racionales, las que han sido estudiadas por distintos filósofos y denominadas de diversas maneras. Han sido llamadas precomprensión, conocimiento tácito, formas de vida y prácticas sociales compartidas. Lo importante es que se trata de formas de comportamiento muy básico, incluso instintivo, que no llegan a ser conscientes ni conceptuales, pero que son condición de posibilidad y están en la base de toda conciencia y conceptualización. Sin duda, todo comportamiento humano hunde sus raíces en esas bases no racionales, pero sería un grave error llamar a eso irracionalidad. No solo sería una confusión conceptual muy básica, sino además iría en contra del uso habitual de los conceptos.

12.3. Algunos tipos de irracionalidad

La discusión en filosofía de la mente ha delimitado por lo menos cuatro tipos de formas de irracionalidad, las cuales se superponen entre sí dando lugar a muchas variedades diferentes:

(1) La acracia o debilidad de la voluntad —*weakness of the will*— se produce cuando el agente actúa contra lo que él mismo considera que es su mejor opción, por lo que produce una acción irracional. Algunos traductores al castellano usan la expresión «incontinencia», pero esta es una desafortunada traducción por dos razones. La primera es que sugiere que el individuo acrático obra irracionalmente porque no se puede contener, lo que no es propiamente un comportamiento irracional. Supongamos, por ejemplo, que uno sabe que determinado tipo de alimento le resulta perjudicial para su salud y que no debe ingerirlo, pero en un momento determinado no puede resistirse a la tentación y lo prueba. En sentido estricto, eso no es irracional, porque, aunque el agente sabe que le resulta dañino probablemente también piense que probarlo unas pocas veces no sería tan grave. Es decir, el agente está justificando racionalmente su comportamiento y, aunque uno podría

no estar de acuerdo con la justificación, esta no es contradictoria con sus otras creencias. Un verdadero caso de irracionalidad se produce cuando el agente actúa en contra de su mejor juicio sin saber por qué lo hace o, incluso, sin saber que lo está haciendo. Imagínese, por ejemplo, a un individuo que tiene una tendencia irrefrenable a verse involucrado en situaciones en las que termina siendo maltratado. Él sabe que esas situaciones le resultan dañinas y cree que debe evitarlas. También desea evitarlas, pero no puede, porque no sabe cómo hacerlo. Es claro que su comportamiento tiene causas y que estas no son sus estados mentales conscientes. Será necesario, por tanto, buscar causas para ese tipo de comportamiento que probablemente serán estados mentales inconscientes. El punto, sin embargo, es que, en caso que encontremos tales causas, el comportamiento será explicado racionalmente, con lo cual su ininteligibilidad se desvanecerá. En el caso que el propio agente tome conciencia de tales estados mentales inconscientes, probablemente el comportamiento irracional ya no vuelva a aparecer.

(2) Davidson (2004c [1982]) denomina «principio de Medea» a la tesis según la cual la acracia se produce porque el agente es invadido por una fuerza extraña que lo obliga a actuar en contra de su mejor juicio. Esta fuerza extraña puede ser la pasión, la ira, la tentación o cualquier otra causa que, siendo interna, actúa como si fuera externa al individuo. Pero explicar el comportamiento de alguien de esa manera no es atribuirle irracionalidad, porque, aunque el agente es consciente de las posibilidades que se abren ante él, no se le está atribuyendo plena responsabilidad a su comportamiento al sostener que está siendo gobernado por una fuerza superior a su voluntad. El caso que nos interesa es este: el agente tiene dos opciones; tiene mejores razones para adoptar una de las dos; después de reflexión cuidadosa opta por una, pero actúa según la otra, es decir, actúa contra su mejor juicio. Eso no está contemplado en el principio de Medea.

(3) El autoengaño —*self-deception*— ocurre cuando el agente tiene creencias contradictorias y lo sabe. Pero, además, tiene una de esas creencias precisamente porque tiene la otra. No es que él tenga creencias falsas, sino que tiene una creencia —inconsciente— que causa y mantiene una creencia opuesta. En este caso, lo que es irracional no es una acción sino una creencia. Imaginemos, por ejemplo, a un individuo que cree tener una grave enfermedad y además está muy atemorizado por ello. Esa creencia y la emoción asociada podrían causar en él la creencia de que tiene una salud de hierro y que nada podría afectarle, lo que sería una creencia irracional porque es incompatible con las otras

creencias que él también tiene. Una formulación más precisa del autoengaño es esta: el sujeto A cree tener evidencia de que la proposición p es verdadera. El pensamiento de que p es verdadera conduce al agente a comportarse de tal manera que se autoinduce la creencia que ~p. Pero lo que convierte al autoengaño en un problema interesante para los filósofos es que la creencia en p no solo ha causado la creencia en ~p, sino además la sostiene.

(4) La debilidad de la justificación —*weakness of the warrant*— ocurre cuando un agente cree lo opuesto de lo que toda la evidencia reconocida por él sugiere. El fenómeno es el siguiente: un individuo tiene evidencia a favor y en contra de una determinada hipótesis. Reunida y considerada toda la evidencia disponible, parece más razonable aceptar la hipótesis, es decir, toda la evidencia reconocida como tal por el propio agente parece coincidir en justificar la hipótesis. Sin embargo, el agente se niega a aceptarla. La inconsistencia —e irracionalidad— reside nuevamente en que uno actúa en contra de su mejor juicio. Sin embargo, para que este sea un caso de irracionalidad y no simplemente de ignorancia, es necesario que el agente sea consciente de estar actuando en contra de su mejor juicio. Pensemos, por ejemplo, en un sujeto que tiene frente a sí una gran cantidad de información que prueba que su contador le está hurtando dinero. Pero, como esta posibilidad lo sumiría en un profundo pesar, adopta la creencia de que el contador es inocente y que quien le está robando es su administrador, por quien tiene poco aprecio. En este caso es nuevamente una creencia la que es irracional. Consideramos este fenómeno como irracional porque presuponemos lo que suele llamarse el «principio de evidencia total». Este principio sostiene que debemos considerar toda la evidencia disponible para evaluar una hipótesis y que de hecho eso es lo que hacemos en los casos normales. ¿Pero qué diríamos si una persona decide no aceptar ese presupuesto? ¿También diríamos que es irracional? Supongamos que alguien dijera que actúa según sus impulsos más inmediatos, que lo hace deliberadamente y que no ve ningún problema en ello. ¿Hasta qué punto esa sería una práctica irracional? Solo lo sería si esa persona es consciente de que esa práctica es inconsistente con otras creencias suyas que tienen como objetivo obtener su propio bien o lograr la mayor consistencia posible. De un lado, la búsqueda del propio bien y el deseo de maximizar el autointerés es natural y no requiere de justificación. De otro lado, la búsqueda de la mayor consistencia es condición necesaria para la acción, porque si uno aceptara deliberadamente estados mentales incompatibles entre sí, su conducta se haría inviable. En efecto, eso es exactamente lo que ocurre en los casos de irracionalidad, pues la conducta suele volverse errática y el sujeto tiende a boicotearse.

(5) Por otra parte, ¿qué pasaría si uno no creyese que siempre deba actuar según su mejor juicio? ¿Qué pasaría si uno creyera que a veces debe actuar siguiendo su intuición? Eso puede perfectamente ocurrir, pero el agente deberá preguntarse cómo debe decidir cuándo actuar según su mejor juicio y cuándo según su intuición. Obviamente seguirá siendo su mejor juicio lo que haga que él decida actuar según su mejor juicio o según su intuición. Dicho de otra manera, aún si uno decide seguir su intuición está actuando según su mejor juicio, que es —según él— su intuición. Este razonamiento demostraría que uno no podría evitar creer que siempre debe actuar según su mejor juicio. Pero esto no resuelve el problema de la acracia, por el contrario, lo hace posible. Porque, entonces, la pregunta nuevamente es: ¿Cómo es que a veces, a pesar de creer que uno debe actuar según su mejor juicio, uno no actúa según su mejor juicio? La inconsistencia —y la irracionalidad— radican precisamente en eso, en que uno hace aquello que cree que no debería hacer, que no desea hacer y que sabe que podría no hacer.

(6) El forjarse ilusiones —*wishful thinking*— ocurre cuando un sujeto cree en algo porque desea que sea verdadero, aunque sabe que no lo es. Consideremos, por ejemplo, el caso de un hombre que es básicamente racional y no cree en los horóscopos, excepto cuando estos predicen situaciones que desea con mucha fuerza. En estos casos también es una creencia la que es irracional.

Debe notarse, sin embargo, que autoengaño, debilidad de la justificación y el forjarse ilusiones son casos parecidos que con frecuencia se superponen. A veces es imposible, y sobre todo innecesario, determinar si un caso pertenece a una categoría u otra. Lo que tienen en común los diversos casos de irracionalidad es que el agente actúa —o tiene estados mentales— en contra de sus estados mentales conscientes. Y dado que un evento natural es considerado una acción solo si puede ser descrito como causado por los estados mentales del agente, nos enfrentamos a dos opciones: o decimos que el supuesto evento natural no es una acción y que el supuesto estado mental en realidad no es un estado mental, o encontramos otros estados mentales —probablemente inconscientes— que los han causado. Pero la pregunta de fondo es cómo podría uno actuar en contra de lo que uno mismo cree y desea, en contra de su mejor juicio, en contra de su mejor opción o en contra de su propio interés. ¿Por qué alguien buscaría boicotearse si sabe que eso es lo que está ocurriendo? Nuevamente es esencial que el agente sea consciente de ello, porque de otra manera no sería un caso de irracionalidad, sino simplemente de error o ignorancia.

En la tradición filosófica, Platón se inclinó por afirmar que la acracia es imposible porque uno nunca actuaría en contra de lo que uno mismo considera su propio

bien, por eso afirmó que nadie hace el mal a sabiendas y que, por tanto, el obrar mal o la supuesta acracia es siempre el producto de la ignorancia de lo que es bueno. Aristóteles (*Ética Nicomaquea*, VII), que era mucho más realista que Platón, constató que la acracia sí existe, pero intentó explicarla de otra manera. Afirmó que, en efecto, es el producto de la ignorancia, pero de una ignorancia transitoria respecto de lo que es bueno para uno mismo. Aristóteles se refiere a algo que el sujeto sabe, pero que pasajeramente ha olvidado, por lo que se genera un daño a sí mismo. Es de particular importancia notar que este olvido momentáneo también puede ser visto como algo que el sujeto sabe, pero que temporalmente ha ocultado de forma deliberada. Se trata de una suerte de «autoocultamiento» porque la causa de este extraño olvido circunstancial son los otros estados mentales del agente, ya sea sus creencias, deseos o emociones.

En el análisis aristotélico el sujeto no es consciente de la alternativa que se le presenta: actuar según su mejor juicio o no. El sujeto actúa por desconocimiento de que hay una mejor opción que él mismo conoce, pero que en este momento no tiene presente ante la mente. En un momento posterior puede reconstruir su acción —y la interpretación que en ese momento tuvo de su acción— y también puede darse cuenta de que hubo una mejor opción, que él la conoció, pero que en ese momento no reparó en ella. La acción, sin embargo, es plenamente intencional. Aristóteles está a un paso de proponer algún tipo de división de la mente en la que hay estados mentales conscientes y otros no conscientes, siendo los últimos reprimidos por los primeros, pero ciertamente nunca dio ese paso y hubo que esperar hasta fines del siglo XIX con autores como Schopenhauer, Nietzsche y Freud para que se diera.

Desde un punto de vista lógico uno de los casos paradigmáticos de irracionalidad, que es el autoengaño, se puede formular de la siguiente manera: una persona cree que ~p precisamente porque cree que p. Es decir, se resiste a admitir que p —que es lo que en un importante sentido cree— y por ello cree que cree que ~p. Dicho de otra manera: el autoengañado no es simplemente una persona que tiene una creencia falsa. Es una persona que cree proposiciones contradictorias.

Es necesario preguntarse qué características debe tener un sistema de creencias para que una creencia termine siendo causa de su negación. En este caso, por supuesto, la creencia en que p no puede justificar conscientemente la creencia en que ~p solo la causa. Así, volvemos nuevamente a la distinción entre justificar y causar. Algo debe ocurrir en el sistema de creencias del autoengañado para que, en algunos casos, las relaciones entre creencias no sean solo de justificación sino también de causalidad. Además, es necesario explicar cómo puede uno creer creencias incompatibles.

Hay que notar que, si atribuimos a alguien autoengaño, afirmamos que cree una proposición y que también cree su opuesta, pero no que tiene una creencia contradictoria. En otras palabras, A creería que p y también creería que ~p, pero no podría ocurrir que crea al mismo tiempo que p y ~p. Si lo formulamos lógicamente, esto es posible:

(1) (A ◊ p) & (A ◊ ~p)

[A cree que p & A cree que ~p]

Pero esto no:

(2) A ◊ (p & ~p)

[A cree que p & ~p]

Algún lector familiarizado con la lógica clásica pensará que de (1) se infiere (2), por simple introducción del operador, con lo cual esta sería una inferencia válida:

(3) [(A ◊ p) & (A ◊ ~p)] → [A ◊ (p & ~p)]

Pero no es así, porque en contextos epistémicos uno no siempre cree lo que se deduce de sus propias creencias. Esta, por ejemplo, es una inferencia válida:

(4) p → q

p

q

[Si p implica q, y p, entonces q]

Pero esta no es válida:

(5) A ◊ (p → q)

A ◊ p

A ◊ q

[Si A cree que p implica q, y A cree que p, entonces A cree que q]

Aunque uno no puede creer una contradicción, sí puede creer simultáneamente proposiciones que se contradicen mutuamente, siempre que estas proposiciones pertenezcan a subsistemas diferentes de creencias al interior del mismo sistema global de creencias del agente. Hay varias razones por las que uno no podría creer una contradicción. En primer lugar, sería imposible que una intérprete atribuya a un agente la creencia en una contradicción, porque ningún tipo de comportamiento podría ser evidencia de ello. Aunque sí sería posible que un comportamiento errático

sea evidencia de que el agente cree al mismo tiempo dos proposiciones que son contradictorias entre sí. Dicho de otra manera, podemos imaginar que una persona tenga conductas incompatibles entre sí —en distintos momentos de su vida, incluso muy cercanos entre sí— pero no podemos imaginar que la misma acción refleje simultáneamente una contradicción. En segundo lugar, sería imposible que alguien se represente al mismo tiempo un hecho como ocurriendo y como no ocurriendo, aunque sí podría representarse ambas cosas en momentos diferentes o en subsistemas distintos de su sistema de creencias.

Así pues, para que una persona crea que p y crea que ~p es necesario que mantenga convenientemente separadas ambas creencias. Es decir, es necesario atribuirle al agente que consideramos autoengañado una división en su sistema de creencias, un fraccionamiento o escisión de su yo. Por ello en última instancia la irracionalidad es un caso de división de la mente. En otras palabras, al interpretar al agente como una criatura *grosso modo* racional, en algunos casos la intérprete necesita atribuirle un número de estructuras semiindependientes que causan y justifican su comportamiento. Estas subestructuras están —como si fueran agentes mismos— dotadas de creencias, deseos y afectos. El comportamiento global del agente —tal como es interpretado por una intérprete— será el producto de la compleja dinámica entre estas subestructuras que interactúan y se obstaculizan entre sí. Solo podemos explicar las inconsistencias al interior de la propia mente —y, en consecuencia, de las acciones— si postulamos subsistemas de creencias internamente consistentes, pero inconsistentes entre sí, que dirigen —siendo causa de— la conducta del individuo de manera alternativa.

En este punto se podría sostener que la división de la mente es inconsistente con el principio de caridad, pues este nos exige asumir que el otro es coherente para poder interpretarlo. Pero ese es precisamente el tema de fondo. La intérprete comienza el proceso de interpretación asumiendo que el otro es coherente y, *grosso modo*, semejante a ella, pero en algunos casos y a medida que el proceso se desarrolla, ella notará que no le queda más remedio que atribuirle a él sistemas de creencias inconsistentes. Ahí es donde empieza a atribuirle irracionalidad. Pero lo central es que aún si ella le atribuye irracionalidad, lo hace con el fin de seguir viéndolo *grosso modo* inteligible y racional. Por eso, la atribución de irracionalidad debe ser excepcional al interior de una mayor coherencia, que es precisamente lo que lo hace inteligible. Si las inconsistencias fuesen —desde el punto de vista de la intérprete— tantas como para hacer imposible cierta coherencia fundamental, ella no tendría frente a sus ojos a una persona comportándose según creencias y deseos, sino a un conjunto de eventos físicos desconectados entre sí. En ese momento, ella habría dejado de interpretarlo y él habría dejado de ser un agente para ella.

Así pues, en general debemos asumir al agente coherente, pero en ocasiones tenemos que atribuirle inconsistencias. ¿Qué tanta inconsistencia debemos atribuirle para que deje de tener sentido decir que lo estamos interpretando? Es decir, ¿en qué momento la atribución de irracionalidad se convierte en ininteligibilidad? Es un asunto de grado y ciertamente no hay una frontera clara. En general, cuando hemos atribuido demasiada inconsistencia resulta poco claro que podamos atribuir incluso una creencia u otro estado mental, porque un estado mental se identifica por sus interconexiones con los demás estados del sistema.

Por otra parte, ¿cómo podemos llegar a saber si el agente es irracional o es que nosotros somos malos intérpretes? Tampoco hay una respuesta clara a esa pregunta. El principio de caridad nos exige buscar toda la consistencia que sea posible y atribuir irracionalidad solo cuándo no quede otra opción. Pero es claro que podría ser nuestra poca habilidad como intérpretes lo que nos condujo a atribuir irracionalidad en un caso en que no era necesario.

Así pues, la comprensión es más un asunto de habilidad, técnica e ingenio que de seguir reglas preestablecidas. Algunas reglas nos pueden dar pistas sobre qué camino seguir, pero dependerá de nuestra experiencia y sabiduría como intérpretes el poder dar más sentido a más comportamiento del agente.

***

Este libro se propone describir, analizar e integrar los diversos procesos que tienen lugar en el fenómeno de la comprensión, es decir, lo que ocurre cuando las personas o las comunidades se comprenden o malentienden mutuamente, incluso si no son conscientes de ello. Esto exige abordar fenómenos asociados, como la explicación, la interpretación, la atribución psicológica, el significado, la metáfora, las formas de vida, la racionalidad y la irracionalidad. Aunque estos temas están en la intersección entre disciplinas diversas como la psicología, las ciencias cognitivas y las ciencias sociales, en este libro es una perspectiva filosófica la que los hilvana y estructura. Es imprescindible plantearse estas preguntas ahora, cuando personas de muy distinta procedencia, identidad y objetivos vivimos en un pequeño y abigarrado mundo, con la obligación moral y real de entendernos mutuamente.

Bibliografía

Anscombe, Gertrude Elizabeth (2000). *Intention.* Massachusetts: Harvard University Press.

Aristóteles (1968). *Poética*. Traducción de Valentín García Yebra. Madrid: Gredos.

Aristóteles (1981). *Posterior Analytics.* Iowa: Peripatetic Press.

Aristóteles (1982). *Metafísica.* Traducción de Valentín García Yebra. Madrid: Gredos.

Aristóteles (1983). *Acerca del alma.* Traducción de Tomás Calvo Martínez. Madrid: Gredos.

Aristóteles (1984). *The Complete Works of Aristotle: The Revised Oxford Translation.* Edición de Jonathan Barnes. Princeton: Princeton University Press.

Aristóteles (1985). *Ética Nicomaquea. Ética Eudemia.* Madrid: Gredos.

Aristóteles (1988). *Política.* Madrid: Gredos.

Aristóteles (1996). *Física.* Traducción de José Luis Calvo Martínez. Madrid: Consejo Superior de Investigaciones Científicas.

Aristóteles (2002). *Retórica*. Traducción de Alberto Bernabé. Madrid: Alianza.

Aristóteles (2015) *De Interpretatione.* Traducción de E.M. Edghill. Adelaida: Universidad de Adelaida. https://ebooks.adelaide.edu.au/a/aristotle/interpretation/

Armstrong, David (1966). *La percepción y el mundo físico.* Madrid: Tecnos.

Astington, Janet (1996). What is Theoretical about The Child's Theory of Mind? A Vygotskian View of Its Development. En Peter Carruthers y Peter Smith (eds.). *Theories of Theories of Mind* (pp. 184-200). Cambridge: Cambridge University Press.

Astington, Janet; Paul Harris & David Olson (1988). *Developing Theories of Mind*. Cambridge: Cambridge University Press.

Atran, Scott (1998). Folk Biology and the Anthropology of Science: Cognitive Universals and Cultural Particulars. *Behavioral and Brain Sciences,* 21, 547-609.

Ayer, Alfred (1952 [1936]). *Language, Truth and Logic.* Nueva York: Dover Publications.

Ayer, Alfred (1986). *El positivismo lógico*. Ciudad de México: Fondo de Cultura Económica [publicado originalmente en *Proceedings of the Aristotelian Society,* Vol. 37, 1936-1937].

Bachelard, Gaston (1985 [1934]). *The New Scientific Spirit.* Traducido por A. Goldhammer. Boston: Beacon Press.

Bacon, Francis (1961) [1620]. *Novum organum. Interpretación de la naturaleza y dominio del hombre.* Madrid: Aguilar.

Bain, Alexander (1859). *The Emotions and the Will*. Londres: John W. Parker.

Barnes, Barry (1986). *T.S. Kuhn y las ciencias sociales*. México: Fondo de Cultura Económica.

Baron-Cohen, Simon; Alan Leslie & Uta Frith (1985). Does the Autistic Children Have a Theory of Mind? *Cognition*, 21, 37-46.

Begby, Endre & Bjorn Ramberg (2016). Davidson's Derangement Revisited. Guest's Editor's Introduction. *Inquiry*, *59*(1), 1-5.

Bennett, Jonathan (1971). *Locke, Berkeley, Hume: Central Themes*. Oxford: Clarendon Press.

Berkeley, George (1974 [1710]). *Principios del conocimiento humano.* Buenos Aires: Aguilar.

Bernstein, Richard (1983). *Beyond Objectivism and Relativism.* Pensilvania: University of Pennsylvania Press.

Bettocchi, Bárbara & Raúl Fatule (eds.) (2014). *Una visión binocular. Psicoanálisis y filosofía* (pp. 123-137). Lima: Fondo Editorial PUCP.

Bickerton, Derek (1990). *Language and Species.* Chicago: University of Chicago Press.

Bickerton, Derek (2009). *Adam's Tongue. How Humans Made Language, How Language Made Humans.* Nueva York: Hill and Wang.

Bickerton, Derek (2014). *More than Nature Needs: Language, Mind, and Evolution.* Massachusetts: Harvard University Press.

Bieri, Peter; Lorenz Krüger & Rolf Peter Horstmann (eds.) (1979). *Transcendental Arguments and Science*. Dordrecht: Reidel.

Bilgrami, Akeel (2006). *Self-Knowledge and Resentment.* Massachusetts: Harvard University Press.

Bion, Wilfred (1962). *Learning from Experience*. Londres: Heinemann.

Bion, Wilfred (1963). *Elements of Psycho-Analysis*. Londres: Heinemann.

Bion, Wilfred (1965). *Transformations: Change from Learning to Growth*. Londres: Heinemann.

Biro John Ivan & Robert W. Shahan (eds.) (1982). *Mind, Brain and Function*. Brighton: Harvester Press.

Black, Max (1977). More about Metaphor. *Dialectica*, *31*(3-4), 431-457.

Black, Max (1979). How Metaphors Work: A Reply to Donald Davidson. *Critical Inquiry*, *6*(1), 131-143.

Black, Max (1981). Metaphor. En Mark Johnson (ed.), *Philosophical Perspectives on Metaphor*: Mineápolis: University of Minnesota Press.

Blackburn, Simon (1995). Theory, Observation and Drama. En Martin Davies y Tony Stone (eds.), *Folk Psychology: The Theory of Mind Debate* (pp. 274-290). Oxford: Blackwell.

Borges, Jorge Luis (2018). *Obras completas.* Buenos Aires: Emecé.

Boroditsky, Lera (2001). Does language Shape Thought? English and Mandarin Speakers' Conceptions of Time. *Cognitive Psychology*, *43*(1), 1-22.

Boroditsky, Lera (2003). Linguistic Relativity. En L. Nadel (ed.), *Encyclopedia of Cognitive Science* (pp. 917-922). Londres: Macmillan.

Bouma, Hanni K. (2006). Radical Interpretation and High Functioning Autistic Speakers: A Defense of Davidson on Thought and Language. *Philosophical Psychology,* 19, 639-662.

Brandom, Robert (1994). *Making it Explicit. Reasoning, Representing, and Discursive Commitment*. Massachusetts: Harvard University Press.

Brentano, Franz (1995). *Psychology from an Empirical Standpoint*. Editado por Linda L. McAlister. Londres: Routledge.

Brown, Harold (1984). *La nueva filosofía de la ciencia.* Madrid: Tecnos.

Brunsteins, Patricia (2010). *La psicología folk. Teorías, prácticas, perspectivas,* Buenos Aires: Ediciones del Signo.

Brunsteins, Patricia (2011). El rol de la empatía en la atribución mental. *Revista Argentina de Ciencias del Comportamiento*, *3*(1), 75-84.

Brunsteins, Patricia (2018). El carácter emotivo de la experiencia empática. En Diana Inés Pérez y Diego Lawler (2017). *La segunda persona y las emociones* (pp. 227-249). Buenos Aires: SADAF.

Burling, Robbins (2005). *The Talking Ape. How Language Evolved*. Oxford: Oxford University Press.

Butterfield, Jeremy (ed.) (1986). *Language, Mind and Logic*. Cambridge: Cambridge University Press.

Byrne, Alex (2011). Knowing that I Am Thinking. En Anthony Hatzimoysis (ed.), *Self-Knowledge* (pp. 105-224). Oxford: Oxford University Press.

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

Campbell, Norman (1957). *Foundations of Science*. Nueva York: Dover.

Caorsi, Carlos (ed.) (1999). *Ensayos sobre Davidson.* Montevideo: Universidad de la República de Uruguay.

Caorsi, Carlos (2001). *De una teoría del lenguaje a una teoría de la acción intencional. Una introducción a la filosofía de Donald Davidson.* Salamanca: Factotum.

Caorsi, Carlos & Waldomiro Silva Filho (eds.). (2008). *Razones e interpretaciones. La filosofía después de Donald Davidson.* Buenos Aires: Ediciones del Signo.

Campbell, Keith (1987). *Cuerpo y mente.* Ciudad de México: Universidad Nacional Autónoma de México.

Carnap, Rudolf (1931). Die Physikalische Sprache als Universal Sprache der Wissenschaft. *Erkenninis*, *11*, 432.

Carnap, Rudolf (1953). Testability and Meaning. En H. Feigl y M. Brodbeck (eds.), *Readings in the Philosophy of Science* (pp. 47-92). Nueva York: Appleton-Century-Crofts.

Carnap, Rudolf (1956). The Methodological Character of Theoretical Concepts. En H. Feigl y M. Scriven (eds.), *Minnesota Studies in the Philosophy of Science* (vol. 1, pp. 38-76). Mineápolis: University of Minneapolis Press.

Carpenter, A. (2002). Davidson's Transcendental Arguments. En Jeffrey Malpas (ed.), *From Kant to Davidson: Philosophy and the Idea of the Transcendental* (pp. 219-237). Londres: Routledge.

Carroll, John Bissell (ed.) (1956). *Language, Thought and Reality: Selected Writings of Benjamin Lee Whorf*. Massachusetts: The MIT Press.

Carruthers, Peter (1996a). Autism as Mindblindness. En Peter Carruthers y Peter Smith (eds.). *Theories of Theories of Mind* (pp. 257-276). Cambridge: Cambridge University Press.

Carruthers, Peter (2006). *The Architecture of the Mind: Massive Modularity and the Flexibility of Thought*. Oxford: Oxford University Press.

Carruthers, Peter (2009). How We Know Our Own Mind. The Relationship between Mindreading and Metacognition. *Behavioral and Brain Sciences*, 32, 121-182.

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

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

Cassam, Quassim (2014). *Self-Knowledge for Humans.* Oxford: Oxford University Press.

Cavell, Marcia (1993). *The Psychoanalytic Mind. From Freud to Philosophy*. Massachusetts: Harvard University Press.

Cavell, Marcia (1998). Triangulation, One's Own Mind and Objectivity. *International Journal of Psychoanalysis*, 79, 449-467.

Cavell, Marcia (2006). *Becoming a Subject. Reflections in Philosophy and Psychoanalysis*. Oxford: Clarendon Press.

Chalmers, David (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press.

Chalmers, David (2010). *The Character of Consciousness*. Oxford: Oxford University Press.

Chalmers, David (2012). *Constructing the World*. Oxford: Oxford University Press.

Chalmers, David (2015). Panpsychism and Panprotopsychism. En Torin Alter y Yujin Nagasawa (eds.), *Consciousness in the Physical World: Perspectives on Russellian Monism* (pp. 246-276). Nueva York: Oxford University Press.

Cheney, Dorothy & Robert Seyfarth (2007). *Baboon Metaphysics. The Evolution of a Social Mind.* Chicago: Chicago University Press.

Cheng, Yawei; Po-Lei Lee, Chia-Yen Yang, Ching-Po Ling, Jean Decety (2008). Gender Differences in the Mu Rhythm of the Human Mirror-Neuron System. *PLOS ONE*, *3*(5), e 2113.

Child, William (1994). *Causality, Interpretation and the Mind.* Oxford: Clarendon Press.

Chirinos, Eduardo (ed.) (1992). *Infame Turba*. *Poesía en la Universidad Católica, 1917-1997*. Lima: Fondo Editorial PUCP.

Chomsky, Noam (1965). *Aspects of the Theory of Syntax.* Massachusetts: The MIT Press.

Chomsky, Noam (1988). *Language and Problems of Knowledge*. Massachusetts: The MIT Press.

Churchland, Patricia (1986). *Neurophilosophy: Toward a Unified Science of the Mind-Brain*. Massachusetts: The MIT Press.

Churchland, Patricia (2002). *Brain-Wise: Studies in Neurophilosophy*. Massachusetts: The MIT Press.

Churchland, Paul (1979a). *Matter and Consciousness.* Massachusetts: The MIT Press.

Churchland, Paul (1979b). *Scientific Realism and the Plasticity of Mind*. Nueva York: Cambridge University Press.

Cicerón (1944). *Cuestiones académicas.* Ciudad de México: El Colegio de México.

Cisneros, Antonio (1992). Domingo en Santa Cristina de Budapest y frutería al lado. En Eduardo Chirinos (ed.), *Infame Turba*. *Poesía en la Universidad Católica, 1917-1997* (p. 127). Lima: Fondo Editorial PUCP.

Clark, Andy (1999). *Cerebro, cuerpo y mundo en la nueva ciencia cognitiva,* Buenos Aires: Paidós.

Clark, Andy (2008). *Supersizing the Mind. Embodiment, Action and Cognitive Extension.* Oxford: Oxford University Press.

Clark, Andy (2016). *Surfing Uncertainty. Prediction, Action and the Embodied Mind.* Oxford: Oxford University Press.

Collingwood, Robin George (1946). *The Idea of History.* Oxford: Oxford University Press.

Comte, Auguste (2002 [1830, 1844]). *Curso de filosofía positiva. Discurso sobre el espíritu positivo.* Barcelona: Folio.

Currie, Gregory (1995). Imagination and Simulation: Aesthetics Meets Cognitive Science. En Martin Davis y Tony Stone (eds.), *Mental Simulation: Evaluations and Applications* (pp. 151-170). Oxford: Blackwell.

Cusa, Nicolás de (1957 [1440]). *De docta ignorantia.* Buenos Aires: Aguilar.

Damasio, Antonio (2001). *La sensación de lo que ocurre. Cuerpo y emoción en la construcción de la conciencia*. Madrid: Debate.

Darwin, Charles (1994 [1971]). *El origen del hombre.* Bogotá: Panamericana.

Darwin, Charles (2009 [1959]). *El origen de las especies.* Ciudad de México: Universidad Nacional Autónoma de México.

Dasenbrok, Reed Way (ed.) (1993). *Literary Theory after Davidson*. Pensilvania: Pennsylvania State University Press.

Davidson, Donald (1980a [1963]). Actions, Reasons and Causes. En Donald Davidson, *Essays on Actions and Events* (pp. 3-21). Oxford: Clarendon Press.

Davidson, Donald (1980b). *Essays on Actions and Events*. Oxford: Clarendon Press.

Davidson, Donald (1980c [1971]). Eternal vs. Ephemeral Events. En *Essays on Actions and Events* (pp. 189-205). Oxford: Clarendon Press.

Davidson, Donald (1980d [1970]). Mental Events. En *Essays on Actions and Events* (pp. 207-225). Oxford: Clarendon Press.

Davidson, Donald (1984a [1974a]). Belief and the Basis of Meaning. En *Inquiries into Truth and Interpretation* (pp. 141-155). Oxford: Clarendon Press.

Davidson, Donald (1984b [1982]). Communication and Convention. En *Inquiries into Truth and Interpretation* (pp. 265-281). Oxford: Clarendon Press.

Davidson, Donald (1984c). *Inquiries into Truth and Interpretation*. Oxford: Clarendon Press.

Davidson, Donald (1984d [1974b]). On the Very Idea of a Conceptual Scheme. En *Inquiries into Truth and Interpretation* (pp. 183-189). Oxford: Clarendon Press.

Davidson, Donald (1984e [1973]). Radical Interpretation. En *Inquiries into Truth and Interpretation* (pp. 125-141). Oxford: Clarendon Press.

Davidson, Donald (1984f [1979]).The Inscrutability of Reference. En *Inquiries into Truth and Interpretation* (pp. 215-227). Oxford: Clarendon Press.

Davidson, Donald (1984g [1967]). Truth and Meaning. En *Inquiries into Truth and Interpretation* (pp. 17-37). Oxford: Clarendon Press.

Davidson, Donald (1984h [1978]). What Metaphors Mean. En *Inquiries into Truth and Interpretation* (pp. 245-265). Oxford: Clarendon Press.

Davidson, Donald (1993). Thinking Causes. En John Heil y Alfred Mele (eds.), *Mental Causation* (pp. 3-17). Oxford: Clarendon Press.

Davidson, Donald (1997). Indeterminacy and Anti-Realism. En Christopher Kulp (ed.), *Realism/Antirealism and Epistemology* (pp. 109-122). Lanham: Rowman and Littlefield.

Davidson, Donald (2001a [1984]). First Person Authority. En *Subjective, Intersubjective, Objective* (pp. 3-15). Oxford: Clarendon Press.

Davidson, Donald (2001b). *Subjective, Intersubjective, Objective.* Oxford: Clarendon Press.

Davidson, Donald (2001c [1997]). The Emergence of Thought. En *Subjective, Intersubjective, Objective* (pp. 123-135). Oxford: Clarendon Press.

Davidson, Donald (2001d [1992]). The Second Person. En *Subjective, Intersubjective, Objective* (pp. 107-123). Oxford: Clarendon Press.

Davidson, Donald (2001e [1991]). Three Varieties of Knowledge. En *Subjective, Intersubjective, Objective* (pp. 205-221). Oxford: Clarendon Press.

Davidson, Donald (2004a [1986]). Deception and Division. En *Problems of Rationality* (pp. 199-2012). Oxford: Clarendon Press.

Davidson, Donald (2004b [1985]). Incoherence and Irrationality. En *Problems of Rationality* (pp. 189-198). Oxford: Clarendon Press.

Davidson, Donald (2004c [1982]). Paradoxes of Irrationality. En *Problems of Rationality* (pp. 169-187). Oxford: Clarendon Press.

Davidson, Donald (2004d). *Problems of Rationality.* Oxford: Clarendon Press.

Davidson, Donald (2004e [1997]). Who is Fooled? En *Problems of Rationality* (pp. 213-230). Oxford: Clarendon Press.

Davidson, Donald (2005a [1986]). A Nice Derangement of Epitaphs. En *Truth, Language and History* (pp. 89-109). Oxford: Clarendon Press.

Davidson, Donald (2005b). *Truth, Language and History.* Oxford: Clarendon Press.

Davies, Martin & Tony Stone (1995a). *Folk Psychology: The Theory of Mind Debate.* Oxford: Blackwell.

Davies, Martin & Tony Stone (eds.) (1995b). *Mental Simulation: Evaluations and Applications*. Oxford: Blackwell.

Deacon, Terrence W. (1997). *The Symbolic Species. The Co-Evolution of Language and the Brain.* Nueva York: Norton & Company.

Dennett, Daniel (1998a [1996]). Making Sense of Ourselves. En *The Intentional Stance* (pp. 103-116). Cambridge, MA: The MIT Press.

Dennett, Daniel (1998b [1987]). Three Kinds of Intentional Psychology. En *The Intentional Stance* (pp. 43-69). Cambridge, MA: The MIT Press.

Dennett, Daniel (2017). *From Bacteria to Bach and Back: The Evolution of Minds.* Londres: Allen Lane.

Denyer, Nicholas (1991). *Language, Thought and Falsehood in Ancient Greek Philosophy.* Londres: Routledge.

Descartes, René (1936 [1626-1628]). *Reglas para la dirección del espíritu*. Santiago de Chile: Ercilla.

Descartes, René (1965 [1649]). *Las pasiones del alma.* Madrid: Aguilar.

Descartes, René (1968 [1641]). *Meditaciones metafísicas.* Madrid: Espasa Calpe.

Descartes, René (2011 [1641]). *Meditaciones metafísicas.* Madrid: Alianza.

Deutscher, Guy (2010). *Through the Language Glass. Why the World Looks Different in Other Languages.* Nueva York: Metropolitan Books.

Devitt, Michael (1981). *Designation.* Nueva York: Columbia University Press.

Devitt, Michael (1984). *Realism and Truth.* Oxford: Basil Blackwell.

Devitt, Michael (2004). *Intuitions*. En Proceedings of the VI International Ontology Congress «From the Gene to Language: The State of the Art». San Sebastián, setiembre-octubre.

De Waal, Frans (1997). *Bien natural. Los orígenes del bien y el mal en los humanos y en otros animales*. Barcelona: Herder.

De Waal, Frans (2013). *The Bonobo and the Atheist. In Search of Humanism among the Primates.* Nueva York y Londres: Norton & Company.

Dewey, John (1910). *The Influence of Darwin on Philosophy and Other Essays.* Nueva York: Henry Holt & Company.

Dilthey, Wilhelm (1949 [1910]). *Introducción a las ciencias del Espíritu*. Ciudad de México: Fondo de Cultura Económica.

Dilthey, Wilhelm (1989). *Introduction to the Human Sciences*. Editado por Rudolf Makkreel y Rodi Frithjof. Princeton: Princeton University Press.

Doppelt, Gerald (1978). Kuhn's Epistemological Relativism: An Interpretation and Defence. *Inquiry*, 21, 33-86.

Dreyfus, Hubert (1996). *Ser en el mundo. Comentarios a la división I de Ser y Tiempo.* Santiago de Chile: Cuatro Vientos.

Droysen, Johann Gustav (1983 [1937]). *Histórica: lecciones sobre la enciclopedia y metodología de la historia.* Barcelona: Alfa.

Duhem, Pierre (1962). *The Aim and Structure of Physical Theory*. Traducción de Philip Weiner. Nueva York: Atheneum.

Duica, William (2014). *Conocer sin representar. El realismo epistemológico de Donald Davidson*. Bogotá: Universidad Nacional de Colombia.

Dummett, Michael (1975). What Is a Theory of Meaning. En Samuel Guttenplan (ed.), *Mind and Language* (pp. 97-138). Oxford: Oxford University Press.

Dunbar, Robin (2003). The Social Brain: Mind, Language and Society in Evolutionary Perspective. *Annual Review of Anthropology*, 32, 163-81.

Dunbar, Robin (2009). Why Only Humans Have Language. En Rudolf Botha y Chris Knight (eds.), *The Prehistory of Language* (pp. 12-35). Oxford University Press.

Dunn, Judy (1991). Understanding Others: Evidence from Naturalistic Studies of Children. En Andrew Whiten (ed.), *Natural Theories of Mind. Evolution, Development and Simulation of Everyday Mindreading* (pp. 51-61). Oxford: Basil Blackwell.

Dwyer, Susan (1999). Moral Competence. En Kumiko Murasugi y Robert Stainton (eds.), *Philosophy and Linguistics* (pp. 169-190). Colorado: Westview Press.

Eco, Umberto (1981). *Lector in fabula. La cooperación interpretativa en el texto narrativo.* Barcelona: Lumen.

Edgley, Roy (1969). *Reason in Theory and Practice.* Londres: Hutchison University Library.

Eisenberg, Nancy & Janet Strayer (eds.) (1987). *Empathy and its Development*. Cambridge: Cambridge University Press.

Engel, Pascal (1994). *Davidson et la Philosophy du Langage.* París: PUF.

Erikson, Erik (1968) *Identity: Youth and Crisis*. Nueva York: Norton Company.

Escajadillo, César (2018). La importancia filosófica de la perspectiva de la segunda persona: agencia y explicación intencional. En Cecilia Monteagudo y Pablo Quintanilla (eds.), *Los caminos de la filosofía. Diálogo y método* (pp. 261-275). Lima: Fondo Editorial PUCP.

Evans, Garreth & John McDowell (1976). *Truth and Meaning*. Oxford: Oxford University Press.

Evans, Garreth & John McDowell (1982). *The Varieties of Reference.* Oxford: Oxford University Press.

Everett, Daniel (2017). *How Language Began. The Story of Humanity's Greatest Invention.* Nueva York: W.W. Norton & Co.

Evnine, Simon (1991). *Donald Davidson.* Stanford: Stanford University Press.

Feigl, Herbert & May Brodbeck (eds.) (1953). *Readings in the Philosophy of Science*. Nueva York: Appleton-Century-Crofts.

Feigl, Herbert & Michael Scriven (eds.) (1956). *Minnesota Studies in the Philosophy of Science*. Volumen 1. Mineápolis: University of Minneapolis Press.

Fernández, Jordi (2013). *Transparent Minds: A Study of Self-Knowledge.* Oxford: Oxford University Press.

Feyerabend, Paul (1965). On the Meaning of Scientific Terms. *The Journal of Philosophy*, 62, 266-274.

Flavell, John H.; Barbara A. Everett, Karen Croft, Eleanor R. Flavell (1981). Young Children's Knowledge about Visual Perception: Further Evidence for the Level 1-Level 2 Distinction. *Developmental Psychology, 17*(1), 99-103.

Fleck, David (2006). Complement Clause Type and Complementation Strategies in Matses. En Robert M. W. Dixon y Alexandra Y. Aikhenvald (eds.), *Complementation: A Cross-Linguistic Typology* (pp. 224-244). Oxford: Oxford University Press.

Flórez, Alfonso; Raúl Meléndez & Magdalena Holguín (eds.) (2003). *Del espejo a las herramientas. Ensayos sobre el pensamiento de Wittgenstein*. Bogotá: Siglo del Hombre, Universidad Nacional de Colombia y Pontificia Universidad Javeriana.

Fodor, Jerry (1986). *La modularidad de la mente*. Madrid: Moratta.

Fodor, Jerry (1987). *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Massachusetts: The MIT Press.

Fodor, Jerry (1990). *A Theory of Content and Other Essays*. Massachusetts: The MIT Press.

Fodor, Jerry (1994). *The Elm and the Expert: Mentalese and Its Semantics*. The 1993 Jean Nicod Lectures. Massachusetts: The MIT Press.

Fodor, Jerry (1995). A Theory of the Child's Theory of Mind. En Martin Davies y Tony Stone (eds.), *Folk Psychology: The Theory of Mind Debate* (pp. 283-296). Oxford: Blackwell.

Fodor, Jerry (1998). *Concepts: Where Cognitive Science Went Wrong*. The 1996 John Locke Lectures. Oxford: Oxford University Press.

Fodor, Jerry & Ernie Lepore (eds.) (1993). Holism: A Consumer Update. *Grazer Philosophische Studien*, *46*, 303-322.

Fonagy, Peter; Gyorgy Gergely, Elliot Jurist & Mary Target (eds.) (2002). *Affect Regulation, Mentalization and the Development of the Self.* Londres: Other Press.

Frege, Göttlob (1984). *Estudios sobre semántica*. Barcelona: Ariel.

Freud, Sigmund (2002 [1856-1939]). *Obras completas*. Ordenamiento, comentarios y notas de James Strachey con la colaboración de Anna Freud, asistidos por Alix Strachey, Alan Tyson y Angela Richards. Traducción de José L. Etcheverry. Buenos Aires: Amorrortu.

Gadamer, Hans-Georg (1977a). *Die Aktualität des Schönen*. Stuttgart: Philip Reclam.

Gadamer, Hans-Georg (1977b). *Verdad y método. Fundamentos de una hermenéutica filosófica.* Salamanca: Sígueme.

Gadamer, Hans-Georg (1992). *Verdad y método II*. Salamanca: Sígueme.

Gallese, Vittorio & Alvin Goldman (1998). Mirror Neurons and the Simulation Theory of Mind-Reading. *Trends in Cognitive Sciences*, *2*(12), 493-501.

Galparsoro, José Ignacio & Alberto Cordero-Lecca (2013). *Reflections on Naturalism.* Ámsterdam: Sense Publishers.

García Bacca, Juan David (2009). *Los presocráticos.* Ciudad de México: Fondo de Cultura Económica.

Gertler, Brie (2011). *Self-Knowledge.* Nueva York: Routledge.

Gettier, Edmund (1963). Is Justified True Belief Knowledge? *Analysis*, 23, 121-123.

Glock, Hans-Johann (2003). Q*uine and Davidson on Language, Thought, and Reality*. Cambridge: Cambridge University Press.

Glüer, Kathrin (2006). Triangulation. En Ernest Lepore y Bary Smith (eds.), *The Oxford Handbook of Philosophy of Language* (pp. 1006-1019). Oxford: Oxford University Press.

Godfrey-Smith, Peter (2017). *Other Minds.* Londres: Penguin Random House.

Goldberg, Nathaniel (2008). Tension with Triangulation. *The Southern Journal of Philosophy*, *XLVI*(3), 363-383.

Goldman, Alvin (1992a). Empathy, Mind and Morals. *Proceedings and Addresses of the American Philosophical Association*, *66*(3), 17-41.

Goldman, Alvin (1992b). *Liaisons. Philosophy meets the cognitive and social sciences.* Londres: Bradford Book.

Goldman, Alvin (1993). The Psychology of Folk Psychology. *Behavioral and Brain Sciences, 16*(1), 15-28.

Goldman, Alvin (1995a [1993]) In Defense of the Simulation Theory. En Martin Davies y Tony Stone (eds.), *Folk Psychology: The Theory of Mind Debate* (pp. 191-206). Oxford: Blackwell.

Goldman, Alvin (1995b [1993]). Interpretation Psychologized. En Martin Davies y Tony Stone (eds.), *Folk Psychology: The Theory of Mind Debate* (pp. 74-99). Oxford: Blackwell.

Goldman, Alvin (2006). *Simulating Minds. The Philosophy, Psychology and Neuroscience of Mindreading.* Oxford: Oxford University Press.

Goldman, Alvin (2007). Philosophical Intuitions: Their Target, their Source, and their Epistemic Status. *Grazer Philosophische Studien*, *74*(1), 1-26.

Goldman, Alvin (2013). *Joint Ventures: Mindreading, Mirroring, and Embodied Cognition*. Oxford: Oxford University Press.

Gómez Lobo, Alfonso (1976). Sobre «lo que es en cuanto es» en Aristóteles. *Revista Latinoamericana de Filosofía*, *II*(1), 19-26.

Gomila, Antoni (2001). La perspectiva de segunda persona: mecanismos mentales de la intersubjetividad. *Contrastes*, suplemento *VI*, 65-86.

Gomila, Antoni (2002). La perspectiva de segunda persona de la atribución mental: mecanismos mentales de la intersubjetividad. *Azafea, IV*, 123-138.

Gomila, Antoni (2003). La perspectiva de segunda persona. En Eduardo Rabossi y Aníbal Duarte (eds.), *Psicología cognitiva y filosofía de la mente* (pp. 195-218). Buenos Aires: Alianza.

Gomila, Antoni (2008). La relevancia moral de la perspectiva de segunda persona. En Diana Pérez y Luis Fernández (eds.), *Cuestiones filosóficas: ensayos en honor de Eduardo Rabossi* (pp. 155-173). Buenos Aires: Catálogos.

Gomila, Antoni (2012). *Verbal Minds: Language and the Architecture of the Mind.* Ámsterdam: Elsevier.

Gomila, Antoni & Diana Pérez (2017). Lo que la segunda persona no es. En Diana Inés Pérez y Diego Lawler, *La segunda persona y las emociones.* Buenos Aires: SADAF.

González, Wenceslao (coord.) (2002). *Diversidad de la explicación científica.* Barcelona: Ariel.

Goodman, Nelson (1955). *Fact, Fiction and Forecast.* Cambridge, MA: Harvard University Press.

Gopnik, Alison & Henry Wellman (1995). Why the Child's Theory of Mind Really Is a Theory. En Martin Davies y Tony Stone (eds.), *Folk Psychology: The Theory of Mind debate* (pp. 232-258). Oxford: Blackwell.

Gordon, Robert (1986). Folk Psychology as Simulation. *Mind and Language, 1*(2), 158-171.

Gordon, Robert (1992). The Simulation Theory: Objections and Misconceptions. *Mind and Language, 7*(1-2), 11-34.

Gordon, Robert (1995a). Folk Psychology as Simulation. En Martin Davies y Tony Stone (eds.), *Folk Psychology: The Theory of Mind Debate* (pp. 60-73). Oxford: Blackwell.

Gordon, Robert (1995b). Reply to Perner and Howes. En Martin Davies y Tony Stone (eds.), *Folk Psychology: The Theory of Mind Debate* (pp. 185-190). Oxford: Blackwell.

Gordon, Robert (1995c). Reply to Stich and Nichols. En Martin Davies y Tony Stone (eds.), *Folk Psychology: The Theory of Mind Debate* (pp. 174-184). Oxford: Blackwell.

Gordon, Robert (1995d). Simulation without Introspection or Inference from Me to You. En Martin Davies y Tony Stone (eds.), *Mental Simulation. Evaluations and Applications* (pp. 53-67). Oxford: Blackwell.

Graesser, Andreas (1977). On Language, Thought and Reality in Ancient Greek Philosophy. *Dialectica*, *31*(3-4), 359-388.

Grandy, Richard (1973). Reference, Meaning and Belief. *The Journal of Philosophy*, *70*(14), 439-452.

Green, Mitch (2017). *Know Thyself. The Value and Limits of Self-Knowledge*. Nueva York: Routledge.

Grimm, Stephen (2006). Is Understanding a Species of Knowledge? *British Journal for the Philosophy of Science*, *57*(3), 515-553.

Grimm, Stephen (2016). How Understanding People Differs From Understanding the Natural World. *Philosophical Issues* (suplemento *Noûs*), 26, 209-225.

Grimm, Stephen; Christopher Baumberger, Sabine Ammon (2017). *Explaining Understanding. New Perspectives in Epistemology and Philosophy of Science.* Nueva York: Routledge.

Guidens, Anthony (ed.) (1995 [1993]). *Habermas y la modernidad.* Ciudad de México: REI.

Guignon, Charles (1990). Philosophy after Wittgenstein and Heidegger. *Philosophy and Phenomenological Research*, *L*(4), 649-672.

Guttenplan, Samuel (ed.) (1975). *Mind and Language*. Oxford: Oxford University Press.

Habermas, Jürgen (1995 [1993]). Cuestiones y contracuestiones. En Anthony Guidens (ed.), *Habermas y la modernidad* (pp. 305-343). Ciudad de México: REI.

Hacking, Ian (1975). *Why does Language Matter to Philosophy*. Cambridge: Cambridge University Press.

Hacking, Ian (1981). *Scientific Revolutions*. Oxford: Oxford University Press.

Hahn, Lewis Edwin (ed.) (1997). *The Philosophy of Hans-Georg Gadamer.* The Library of Living Philosophers, Vol. XXIV. Chicago: Open Court.

Hampshire, Stuart (1971). *Freedom of Mind.* Princeton: Princeton University Press.

Hanke, Lewis (1974). *All Mankind is One; a Study of the Disputation between Bartolomé de las Casas and Juan Ginés de Sepúlveda in 1550 on the Intellectual and Religious Capacity of the American Indians*. Illinois: Northern Illinois University Press.

Happé, Francesca (1995). Understanding Minds and Metaphors: Insights from the Study of Figurative Language in Autism. *Metaphor and Symbolic Activity*, *10*(4), 275-295.

Harman, Gilbert (1999). Moral Philosophy meets Social Psychology: Virtue Ethics and the Fundamental Attribution Error. *Proceedings of the Aristotelian Society,* 99, 315-31.

Harré, Rom (1986). *Varieties of Reference.* Oxford: Basil Blackwell.

Harris, Paul (1995). Imagining and Pretending. En Martin Davies y Tony Stone (eds.), *Mental Simulation. Evaluations and Applications* (pp. 170-184). Oxford: Blackwell.

Heal, Jane (1986). Replication and Functionalism. En Jeremy Butterfield (ed.), *Language, Mind and Logic* (pp. 135-150). Cambridge: Cambridge University Press.

Hegel, George W. F. (1966 [1807]). *Fenomenología del espíritu.* Ciudad de México: Fondo de Cultura Económica.

Hegel, George W. F. (1980). La visión racional de la historia universal. En *Lecciones sobre filosofía de la historia universal* (pp. 43-59). Madrid: Alianza Universidad.

Heidegger, Martin (1986 [1927]). *Ser y tiempo*. Ciudad de México: Fondo de Cultura Económica.

Heidegger, Martin (1987). *De camino al habla*. Barcelona: Odós.

Hempel, Carl (1970). *Aspects of Scientific Explanation and other Essays in the Philosophy of Science.* Nueva York: The Free Press.

Hempel, Carl (1973). *Filosofía de la ciencia natural*. Madrid: Alianza.

Hempel, Carl & Paul Oppenheim (1948). Studies in the Logic of Explanation. *Philosophy of Science,* 15, 135-175.

Henderson, David (1993). *Interpretation and Explanation in the Human sciences.* Nueva York: State University of New York Press.

Henzi, Peter; L. de Souza Pereira, D. Hawker-Bond, J. Stiller, R. Dunbar & L. Barrett (2007). Look Who's Talking: Developmental Trends in the Size of Conversational Clique. *Evolution and Human Behavior, 28*(4), 66-74.

Hernández Iglesias, Manuel (2003). *El tercer dogma. Interpretación, metáfora e inconmensurabilidad.* Madrid: Antonio Machado Libros.

Higginbotham, James (2003). *Meaning.* Oxford: Blackwell.

Hiley, David R.; James F. Bohman & Richard Shusterman (eds.) (1991). *The Interpretive Turn. Philosophy, Science, Culture*. Ithaca: Cornell.

Hintikka, Jaakko (1972). Transcendental Arguments: Genuine and Spurius. *Noûs, 6*(3), 274-281.

Hirsch, Eric Donald (1967). *Validity in Interpretation*. New Haven: Yale University Press.

Hirsch, Eric Donald (1978). *The Aims of Interpretation*. Chicago: University of Chicago Press.

Hirshfield, Lawrence & Susan Gelman (1994) (eds.). *Mapping the Mind: Domain Specificity in Cognition and Culture.* Nueva York: Cambridge University Press.

Hollander, Eric y otros (2007). Oxytocin Increases Retention of Social Cognition in Autism. *Biological Psyquiatry,* 61, 498-503.

Honderlich, Ted (1982). The Argument for Anomalous Monism. *Analysis*, XLII, 59-64.

Hopkins, Eduardo (ed.) (2002). *Homenaje a Luis Jaime Cisneros*. Lima: Fondo Editorial PUCP.

Hoyos, Luis Eduardo (ed.) (2005). *Relativismo y racionalidad*. Bogotá: Universidad Nacional de Colombia.

Hume, David (1968) [1739]. *A Treatise of Human Nature*. Edición de L.A. Selby-Bigge. Oxford: Clarendon Press.

Hume, David (1975). *Inquiries Concerning Human Understanding and Concerning the Principles of Morals.* Reimpreso de la edición póstuma de 1777 y editado con introducción, tabla comparativa de contenidos e índice analítico por L.A. Selby-Bigge, revisada por P.H. Nidditch. Oxford: Oxford University Press y Clarendon Press.

Hume, David (2002 [1748]). *Investigación sobre el conocimiento humano.* Madrid: Biblioteca Nueva.

Hume, David (2006 [1751]). *Investigación sobre los principios de la moral.* Madrid: Alianza.

Husserl, Edmund (1973a). *Experience and Judgment*. Edición de Ludwig Landgrebe. Evanston: Northwestern University Press.

Husserl, Edmund (1973b [1936]). *The Crisis of European Sciences and Transcendental Phenomenology*. Evanston: Northwestern University Press.

Ivask, Ivar & Astrid Ivask (1975). Odysseus Elytis on His Poetry. *Books Abroad, 49*(4), 631-645.

James, William (1904). A World of Pure Experience. *The Journal of Philosophy, Psychology and Scientific Methods*, *1*(20-21), 533-543, 561-570.

James, William (1983 [1890]). *The Principles of Psychology.* Massachusetts: Harvard University Press.

James, William (1994 [1902]). *Las variedades de la experiencia religiosa. Estudio de la naturaleza humana.* Barcelona: Península.

James, William (1996 [1912]). *Essays in Radical Empiricism*. Nebraska: University of Nebraska Press.

Jauss, Hans Robert (1992), *Experiencia estética y hermenéutica literaria*. Madrid: Taurus.

Johnson, Mark (ed.) (1981). *Philosophical Perspectives on Metaphor*: Minnesota: University of Minnesota Press.

Kant, Immanuel (1987). La historia universal en un sentido cosmopolita. En *Filosofía de la historia* (pp. 39-67). Traducción de E. Imaz. Ciudad de México: Fondo de Cultura Económica.

Kant, Immanuel (2009 [1781]). *Crítica de la razón pura.* Buenos Aires: Colihue.

Kauppinen, Antti (2007). The Rise and Fall of Experimental Philosophy. *Philosophical Explorations*, *10*(2), 95-118.

Kerferd, G.B. (1985). The Presocratics and the Meanings of Words. En *Language and Reality in Greek Philosophy* (pp. 16-21). Atenas: Greek Philosophical Society.

Keynes, John Maynard (1943). *A Treatise on Probability*. Londres: Macmillan.

Kim, Jaegwon (1984). Epiphenomenal and Supervenient Causation. *Midwest Studies in Philosophy,* IX, 257-270.

Kim, Jaegwon (1989). The Myth of Non-reductive Materialism. *Proceedings and Addresses of the American Philosophical Association,* LXIII, 31-47.

Kim, Jaegwon (2000). *Mind in a Physical World.* Massachusetts: The MIT Press.

Kim, Jaegwon (2007). *Physicalism, or Something Near Enough.* Princeton: Princeton University Press.

Knobe, Joshua; Wesley Buckwalter, Shaun Nichols, Philip Robbins, Hagop Sarkissian & Tamler Sommers, (2012). Experimental Philosophy. *Annual Review of Psychology, 63*(1), 81-99.

Kordig, Carl R. (1971). *The Justification of Scientific Change*. Dordrecht: Reidel.

Koyré, Alexandre (1979 [1957]). *Del mundo cerrado al universo infinito*. Traducción de Carlos Solís. Madrid: Siglo XXI.

Krausz, Michael (ed.) (1975). *Relativism: Interpretation and Confrontation*. Indiana: University of Notre Dame Press.

Kripke, Saul (1980). *Naming and Necessity.* Massachusetts: Harvard University Press.

Kristeva, Julia (1991). *Strangers to Ourselves*. Nueva York: Columbia University Press.

Kuhn, Thomas (1971 [1962]). *La estructura de las revoluciones científicas.* Ciudad de México: FCE.

Kuhn, Thomas (1979 [1957]). *La revolución copernicana.* Barcelona: Ariel.

Kuhn, Thomas (1982) [1978]. *La tensión esencial. Estudios selectos sobre la tradición y el cambio en el ámbito de la ciencia.* Ciudad de México: Fondo de Cultura Económica.

Kuhn, Thomas (1991). The Natural and the Human Sciences. En David Hiley, James Bohman y Richard Shusterman (eds.), *The Interpretive Turn* (pp. 17-25) Ithaca: Cornell University Press.

Kulp, Christopher (ed.) (1997). *Realism/Antirealism and Epistemology*. Maryland: Rowman and Littlefield.

Lakatos, Imre & Alan Musgrave (eds.) (1970). *Criticism and the Growth of Knowledge*. Cambridge: Cambridge University Press.

Lasonen, Maria & Marvan, Tomasz (2004). Davidson's Triangulation: Content Endowing Causes and Circularity. *International Journal of Philosophical Studies,* 12, 177-195.

Lear, Jonathan (1991). *Love and its Place in Nature. A Philosophical Interpretation of Freudian Psychoanalysis.* Nueva York: The Noonday Press.

Lefebvre, Henri (2016). *Metaphilosophy.* Londres: Verso Books.

Leibniz, Gottfried W. (1972 [1686]). *Discurso de metafísica*. Buenos Aires: Aguilar.

Leibniz, Gottfried W. (1992 [1704]). *Nuevos ensayos sobre el entendimiento humano.* Madrid: Alianza.

Lepore, Ernest (1992). *Truth and Interpretation. Perspectives on the Philosophy of Donald Davidson.* Cambridge: Blackwell.

Lepore, Ernest & Kirk Ludwig (2005). *Donald Davidson: Meaning, Truth, Language and Reality*. Oxford: Clarendon Press.

Lepore, Ernest & Kirk Ludwig (eds.) (2013). *A Companion to Donald Davidson.* Oxford: John Wiley & Sons.

Leslie, Alan (1988). Some Implications of Pretense for Mechanisms Underlying the Child's Theory of Mind. En Janet Astington, Paul Harris y David Olson (1988). *Developing Theories of Mind* (pp. 19-46). Cambridge: Cambridge University Press.

Leslie, Alan (1992). *Pretense, Autism and the Theory of Mind Module*. http://ruccs.rutgers.edu/images/personal-alan-leslie/publications/Leslie%201992.pdf

Leslie, Alan (1994). ToMM, ToBy, and Agency: Core Architecture and Domain Specificity. En L. Hirschfeld y S. Gelman (eds.), *Mapping the Mind: Domain Specificity in Cognition and Culture* (pp. 119-148). Massachusetts: Cambridge University Press.

Leslie, Alan (1999). Modularity, Development and Theory of Mind. *Mind and Language*, *14*(1), 131-153.

Leslie, Alan (2000). «Theory of Mind» as a Mechanism of Selective Attention. En M. Gazzaniga (ed.), *The New Cognitive Neurosciences* (pp. 1235-1247). Segunda edición. Massachusetts: The MIT Press.

Leslie, Alan & Laila Thaiss (1992). Domain Specificity in Conceptual Development: Neuropsychological Evidence from Autism. *Cognition*, 43, 225-251.

Lewis, David (1972). Psychophysical and Theoretical Identifications. *Australasian Journal of Philosophy*, 50, 249-258.

Lewis, David (1983). Radical Interpretation. En *Philosophical Papers* (pp. 108-121). Oxford: Oxford University Press.

Lewis, David (2002). *Convention. A Philosophical Study*. Londres: Blackwell.

Lipps, Theodor (1903). Einfühlung, inner Nachamung, und Organempfindungen. *Archiv für die Gesamte Psychologie*, 2, 185-204.

Lipton, Peter (2004). *Inference to the Best Explanation*. Nueva York: Routledge.

Locke, John (1982 [1690]). *Ensayo sobre el entendimiento humano.* Ciudad de México: Fondo de Cultura Económica.

López Maguiña, Santiago; Gonzalo Portocarrero, Rocío Silva-Santisteban y Víctor Vich (eds.) (2001). *Estudios culturales. Discursos, poderes, pulsiones*. Lima: Red Para el Desarrollo de las Ciencias Sociales en el Perú.

Macdonald, Graham & Philip Pettit (1981). *Semantics and Social Science.* Londres: Routledge and Kegan Paul.

Machery, Edouard (2017). *Philosophy within Its Proper Bounds.* Oxford: Oxford University Press.

MacIntyre, Alasdair (1985). Relativism, Power and Philosophy. *Proceedings and Addresses of the American Philosophical Association*, *59*(1), 5-22.

MacKay, Gilbert & Adrianne Shaw (2004). A Comparative Study of Figurative Language in Children with Autistic Spectrum Disorders. *Child Language, Teaching and Therapy*, *20*(1), 13-32.

Malebranche, Nicolás (1972-1986). *Oeuvres completes.* París: Vrin.

Malpas, Jeffrey (1989). The Intertranslatability of Natural Languages. *Synthese*, 78, 233-264.

Malpas, Jeffrey (1990). Transcendental Arguments and Conceptual Schemes. A Reconsideration of Körner's Uniqueness Argument. *Kant-Studien*, 81, 232-251.

Malpas, Jeffrey (1992). *Donald Davidson and the Mirror of Meaning.* Cambridge: Cambridge University Press.

Malpas, Jeffrey (ed.) (2002). *From Kant to Davidson: Philosophy and the Idea of the Transcendental*. Londres: Routledge.

Malpas, Jeffrey (ed.) (2011). *Dialogues with Davidson. Acting, Interpreting, Understanding.* Cambridge: Cambridge University Press.

Manicas, Peter (1987). *A History and Philosophy of the Social Sciences.* Oxford: Basil Blackwell.

Manicas, Peter (2006). *A Realist Philosophy of Social Science. Explanation and Understanding.* Cambridge: Cambridge University Press.

Mantilla, Carla (2014a). El vínculo de apego como escenario para el desarrollo de la cognición social temprana. En Pablo Quintanilla, Carla Mantilla y Paola Cépeda (eds.). *Cognición social y lenguaje. La intersubjetividad en la evolución de la especie y en el desarrollo del niño* (pp. 351-370). Lima: Fondo Editorial PUCP.

Mantilla, Carla (2014b). Mente y realidad: modos de vinculación no reflexivas en la experiencia subjetividad temprana. En Bárbara Bettocchi y Raúl Fatule (eds.), *Una visión binocular. Psicoanálisis y filosofía* (pp. 137-154). Lima: Fondo Editorial PUCP.

Masterman, Margaret (1970). The Nature of Paradigm. En Imre Lakatos y Alan Musgrave (eds.), *Criticism and the Growth of Knowledge* (pp. 59-90). Cambridge: Cambridge University Press.

Mayoral, José Antonio (ed.) (1987). *Estética de la recepción*. Madrid: Arco.

McCall, Cade & Tania Singer (2012). The Animal and Human Neuroendocrinology of Social Cognition. Motivations and Behavior. *Nature Neuroscience*, 15, 681-688.

McDowell, John (1985). Functionalism and anomalous monism. En Ernest Lepore y Brian McLaughlin (eds.). *Actions and Events. Perspectives in the Philosophy of Donald Davidson* (pp. 387-398). Oxford: Basil Blackwell.

McDowell, John (1998). *Mind, Value and Reality.* Massachusetts: Harvard University Press.

McDowell, John (2003). Subjective, Intersubjective, Objective. *Philosophy and Phenomenological Research, LXVII*(3), 675-681.

McDowell, John (2007). Triangulating with Davidson. *Philosophical Quarterly, 57*(226), 96-103.

McGinn, Colin (1977). Charity, Interpretation and Belief. *Journal of Philosophy*, 74, 521-535.

McGinn, Colin (2013). Triangulation. En Ernest Lepore y Kirk Ludwig (eds.), *A Companion to Donald Davidson* (pp. 456-471). Malden y Oxford: Wiley Blackwell.

Mead, George Herbert (1972 [1934]). *Espíritu, persona y sociedad*. Buenos Aires: Paidós.

Mikhail, John; Cristina Sorrentino & Elizabeth Spelke (1998). Toward a Universal Moral Grammar. En Morton Ann Gernbacher y Sharon Derry (eds.), *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society* (p. 1250). Nueva Jersey: Erlbaum.

Mill, John Stuart (1998 [1874]). *Three Essays on Religion. Nature, The Utility of Religion, Theism.* Nueva York: Prometheus Books.

Mill, John Stuart (2002 [1874]). *A System of Logic Rationative and Inductive: Being a Connected View of the Principles of Evidence, and the Methods of Scientific Investigation.* Honolulu: University Press of the Pacific.

Mill, John Stuart (2017 [1912]). *Nature*. http://www.earlymoderntexts.com/assets/pdfs/mill1873b.pdf

Moll, Henrike & Andrew Meltzoff (2011). Joint Attention as the Fundamental Basis of Understanding Perspectives. En Axel Seemann (ed.), *Joint Attention. New Developments in Psychology, Philosophy of Mind and Social Neuroscience* (pp. 393-414). Massachusetts: The MIT Press.

Momigliano, Arnaldo (1975). *Alien Wisdom; the limits of Hellenization*. Cambridge: Cambridge University Press.

Monteagudo, Cecilia (2018). ¿Verdad o método? En torno al llamado gadameriano de ir «más allá del control de la metodología científica». En Cecilia Monteagudo y Pablo Quintanilla (eds.), *Los caminos de la filosofía. Diálogo y método* (pp. 303-322). Lima: Fondo Editorial PUCP.

Monteagudo, Cecilia & Quintanilla, Pablo (2018). *Los caminos de la filosofía. Diálogo y método.* Lima: Fondo Editorial PUCP.

Moran, Richard (2012). Self-Knowledge, «Transparency» and the Forms of Activity. En Declan Smithies y Daniel Stoljar (eds.), *Introspection and Consciousness* (pp. 211-236). Oxford: Oxford University Press.

Moya, Carlos (1990). *The Philosophy of Action*. Oxford: Basil Blackwell.

Myers, Robert & Claudine Verheggen (2016). *Donald Davidson's Triangulation Argument: A Philosophical Inquiry.* Nueva York: Routledge.

Nagel, Ernest (1961). *The Structure of Science. Problems in the Logic of Scientific Explanation*. Nueva York: Harcourt, Brace and World.

Nagel, Ernest (2006). *La estructura de la ciencia. Problemas de la lógica de la investigación científica.* Barcelona: Paidós.

Nagel, Thomas (1996 [1979]). *Mortal Questions.* Cambridge: Cambridge University Press.

Nagel, Thomas (1998 [1986]). *Una visión desde ningún lugar.* Ciudad de México: Fondo de Cultura Económica.

Neurath, Otto (1932-1933). Protokolsätze. *Erkenntnis,* 3, 15-28.

Newton, Isaac (1982 [1687]). *Principios matemáticos de la filosofía natural y su sistema del mundo*. Edición de Antonio Escohotado Madrid: Editora Nacional.

Newton-Smith, William (1987). *La racionalidad de la ciencia.* Barcelona: Paidós.

Nichols, Shaun & Stephen P. Stich (2003). *Mindreading: An Integrated Account of Pretence, Self-awareness, and Understanding other Minds.* Oxford: Oxford University Press.

Nozick, Robert (1981). *Philosophical Explanations*. Cambridge: The MIT Press.

Nozick, Robert (1993). *The Nature of Rationality*. Princeton: Princeton University Press.

Oberman, Lindsay & Vilayanur S. Ramachandran (2009). Reflections on the Mirror Neuron System: Their Evolutionary Functions Beyond Motor Representation. En Jaime Pineda, Jaime (ed.), *Mirror Neuron Systems: The Role of Mirroring Processes in Social Cognition* (pp. 39-62). Nueva York: Humana Press.

O'Brian, Lucy (2007). *Self-Knowing Agents.* Oxford: Oxford University Press.

Olivé, León (ed.) (1988a). *Racionalidad. Ensayos sobre la racionalidad en ética y política, ciencia y tecnología*. Ciudad de México: Siglo XXI.

Olivé, León (1988b). Racionalidad y relativismo: relativismo moderadamente radical. En León Olivé (ed.), *Racionalidad. Ensayos sobre la racionalidad en ética y política, ciencia y tecnología* (pp. 267-294). Ciudad de México: Siglo XXI.

Onishi, Kristine & Renée Baillargeon (2005). Do 15-Month-Old Infants Understand False Beliefs? *Science,* 308, 255-258.

O'Shaughnessy, Brian (1980). *The Will.* Cambridge: Cambridge University Press.

Pagin, Peter (2001). Semantic triangulation. En Peter Pagin y Gabriel Segal (eds.), *Interpreting Davidson* (pp. 199-212). Stanford: CSLI.

Papineau, David (1978). *For Science in Social Science.* Londres: Macmillan.

Parfait, Derek (1984). *Reasons and Persons.* Oxford: Oxford University Press.

Pedace, Karina (2012). La normatividad de lo mental y el rol de la segunda persona. *Areté, XXIV*(1), 109-152.

Pedace, Karina (2017). *Mente y lenguaje. La filosofía de Donald Davidson, modelo para armar.* Buenos Aires: SADAF.

Pedace, Karina (2018). La atribución psicológica y la normatividad de lo mental. En Tomás Balmaceda y Karina Pedace (comps.), *Temas de filosofía de la mente* (pp. 205-227). Buenos Aires: SADAF.

Peirce, Charles Sanders (1931-1958). *Collected Papers of Charles Sanders Peirce.* Volúmenes 1-6 (1931-1935) editados por Charles Hartshorne y Paul Weiss. Volúmenes 7 y 8 (1958) editados por Arthur W. Burk. Massachusetts: Harvard University Press.

Peirce, Charles Sanders (1982-2000). *Writings of Charles S. Peirce: A Chronological Edition*. Edición de M.H. Fisch y otros. Seis volúmenes. Indiana: Indiana University Press.

Peirce, Charles Sanders (1992-1998). *The Essential Peirce. Selected Philosophical Writings*. Volúmenes 1-2. Edición de Nathan Houser y otros. Indiana: Indiana University Press.

Pérez, Diana Inés (2013). *Sentir, desear, creer: una aproximación filosófica a los conceptos psicológicos.* Buenos Aires: Prometeo.

Pérez, Diana Inés & Diego Lawler (2017). *La segunda persona y las emociones.* Buenos Aires: SADAF.

Perner, Joseph (1991). *Understanding the Representational Mind*. Cambridge: The MIT Press.

Perner, Josef & Heinz Wimmer (1983). Beliefs about Beliefs: Representation and Constraining Function of Wrong Beliefs in Young Children's Understanding of Deception. *Cognition,* 13, 103-128.

Perner, Josef & Ted Ruffman (2005). Infants' Insight into the Mind: How Deep? *Science,* 308, 214-216.

Pineda, Jaime (2009). *Mirror Neuron Systems: The Role of Mirroring Processes in Social Cognition.* Nueva York: Humana Press.

Platón (1981-1999). *Obras completas.* Madrid: Gredos.

Platón (1992). *Diálogos VII*. Madrid: Gredos.

Platón (2010). *Fedón.* Traducción de Conrado Eggers Lan. Buenos Aires: Eudeba.

Polanyi, Michael (1966). *The Tacit Dimension*. Nueva York: Doubleday and Company.

Popp, Jerome (2007). *Evolution First Philosopher. John Dewey and the Continuity of Nature.* Nueva York: SUNY.

Popper, Karl (1982). *La lógica de la investigación científica.* Madrid: Tecnos.

Popper, Karl (1983). *Conjeturas y refutaciones. La lógica del conocimiento científico.* Barcelona: Paidós.

Premack, David & Guy Woodruff (1978). Does the Chimpanzee Have a Theory of Mind? *Behavioural Brain Sciences,* 4, 515-526.

Proclo (2007). *Commentary on Plato's Cratylus*. Londres: Duckworth.

Quine, Willard Van Orman (1960). *Word and Object.* Cambridge: The MIT Press.

Quine, Willard Van Orman (1963). Two Dogmas of Empiricism. En *From a Logical Point of View* (pp. 20-46). Nueva York: Harper Torchbooks.

Quine, Willard Van Orman (1969). *Ontological Relativism and Other Essays*. Nueva York: Columbia University Press.

Quine, Willard Van Orman (1970). On the Reasons for Indeterminacy of Translation. *Journal of Philosophy*, 67, 178-183.

Quine, Willard Van Orman (1974). Comments on Donald Davidson. *Synthese*, 77, 325-329.

Quine, Willard Van Orman (1978). A Postscript on Metaphor. *Critical Inquiry, 5*(1), 161-162.

Quine Willard Van Orman & Joseph Ullian (1970). *The Web of Belief*. Nueva York: Random House.

Quintanilla, Pablo (1990). Lenguaje y pensamiento. Aristóteles y el modelo de la melodía. *Areté*, *III*(1), 23-40.

Quintanilla, Pablo (1993). Teoría de la acción y la racionalidad en Donald Davidson. *Areté*, *V*(1-2), 145-161.

Quintanilla, Pablo (1994). Denyer, Nicholas (1991). *Language, Thought and Falsehood in Ancient Greek Philosophy*. (Issues in Ancient Philosophy), Londres: Routledge, 1991, 222 pp. [reseñas]. *Areté*, *VI*(1), 181-183.

Quintanilla, Pablo (1995). Metáfora e interpretación en Donald Davidson. *Areté*, *VII*(1), 113-129.

Quintanilla, Pablo (1997). Significado y verificación. Las posibilidades de una teoría holista de la interpretación. *Ideas y valores*, *46*(105) 30-50.

Quintanilla, Pablo (1999). La hermenéutica de Davidson: metáfora y creación conceptual. En Carlos Caorsi (ed.), *Ensayos sobre Davidson* (pp. 69-93). Montevideo: Universidad de la República de Uruguay.

Quintanilla, Pablo (2000). John Preston (ed.). *Thought and Language*. Royal Institute of Philosophy. Supplement 42. Cambridge: Cambridge University Press. 1997, 249 pp. [reseña]. *Areté, XII*(1), 153-158.

Quintanilla, Pablo (2001a) El lugar de la racionalidad en la comprensión del otro. En Santiago López Maguiña, Gonzalo Portocarrero, Rocío Silva-Santisteban y Víctor Vich (eds.), *Estudios culturales. Discursos, poderes, pulsiones* (pp. 357-376). Lima: Red Para el Desarrollo de las Ciencias Sociales en el Perú.

Quintanilla, Pablo (2001b). La esfera o la tortuga. Las posibilidades de una teoría holista de la justificación. *Areté, XIV*(1), 121-144.

Quintanilla, Pablo (2002a). La doctrina de los dos puntos de vista. En Eduardo Hopkins (ed.), *Homenaje a Luis jaime Cisneros* (tomo 1). Lima: Fondo Editorial PUCP.

Quintanilla, Pablo (2002b). Ver un mundo diferente. Consecuencias ontológicas de la filosofía de Thomas Kuhn. En VV.AA., *Actas del Primer Simposio de Estudiantes de Filosofía* (pp. 101-118). Lima: PUCP.

Quintanilla, Pablo (2003a). Conocimiento, demarcación y elección de teorías. En Chaparro, Adolfo y Christian Schumacher (eds.). *Racionalidad y discurso mítico.* Bogotá: Universidad del Rosario.

Quintanilla, Pablo (2003b). El lenguaje de la intimidad. Sobre la constitución intersubjetiva de las emociones. En Alfonso Flórez, Raúl Meléndez y Magdalena Holguín (eds.), *Del espejo a las herramientas. Ensayos sobre el pensamiento de Wittgenstein* (pp. 241-260). Bogotá: Siglo del Hombre, Universidad Nacional de Colombia y Pontificia Universidad Javeriana.

Quintanilla, Pablo (2004). Comprender al otro es crear un espacio compartido. Caridad, empatía y triangulación. *Ideas y Valores*, 124, 117-134.

Quintanilla, Pablo (2005). Interpretando al otro: comunicación, racionalidad y relativismo. En Luis Eduardo Hoyos (ed.), *Relativismo y racionalidad* (pp. 19-40). Bogotá: Universidad Nacional de Colombia.

Quintanilla, Pablo (2007). La conquista aristotélica de las emociones. *Revista de Psicoanálisis. Sociedad Peruana de Psicoanálisis*, 5, 139-146.

Quintanilla, Pablo (2008a). Comprensión, imaginación y transformación. *Areté, XX*(1), 111-135.

Quintanilla, Pablo (2008b). *Ensayos de metafilosofía.* Lima: Fondo Editorial PUCP.

Quintanilla, Pablo (2008c). Taylor contra Davidson: intérpretes participantes o no comprometidos. En Carlos Caorsi y Waldomiro Silva-Filho (eds.), *Razones e interpretación. La filosofía después de Donald Davidson* (pp. 145-160). Buenos Aires: Del Signo.

Quintanilla, Pablo (2009a). Consecuencialismo ético, desarrollo y etnocentrismo. En Patricia Ruiz-Bravo, Patricia, Pepi Patrón y Pablo Quintanilla (eds.), *Desarrollo humano y libertades. Una aproximación interdisciplinaria* (pp. 181-198). Lima: Fondo Editorial PUCP.

Quintanilla, Pablo (2009b). Mostrando metáforas. *Lexis, XXXIII*(1), 141-150.

Quintanilla, Pablo (2009c). Pragmatismo y realismo moral. *Cuadernos de Filosofía*, 27, 7-22.

Quintanilla, Pablo (2011). The Evolution of Agency. En Jorge Martínez Contreras y Aura Ponce de León (eds.). *Darwin's Evolving Legacy* (pp. 444-456). Ciudad de México: Siglo XXI.

Quintanilla, Pablo (2013). Naturalism and the Mind. En José Ignacio Galparsoro y Alberto Cordero-Lecca (eds.), *Reflections on Naturalism* (pp. 33-42). Ámsterdam: Sense Publishers.

Quintanilla, Pablo (2014a). Filosofía analítica y filosofía continental. La apropiación de un debate por la filosofía peruana. *Solar. Revista de Filosofía Iberoamericana, 10*(2), 31-42.

Quintanilla, Pablo (2014b). La evolución de la atribución psicológica: lectura de mentes y metacognición. En Pablo Quintanilla, Carla Mantilla y Paola Cépeda (eds.), *Cognición social y lenguaje. La intersubjetividad en la evolución de la especie y en el desarrollo del niño* (pp. 241-260). Lima: Fondo Editorial PUCP.

Quintanilla, Pablo (2014c). La mente como un sistema complejo de propiedades emergentes. *Revista Peruana de Psiquiatría*, *4*(1), 30-40.

Quintanilla, Pablo (2014d). Los límites de la irracionalidad. En Bárbara Bettocchi y Raúl Fatule (eds.), *Una visión binocular. Psicoanálisis y filosofía* (pp. 123-136). Lima: Fondo Editorial PUCP.

Quintanilla, Pablo (2014e). ¿Qué es la agencia? En Fidel Tubino, Catalina Romero, Efraín Gonzales de Olarte (eds.), *Inclusiones y desarrollo humano: relaciones, agencia, poder* (pp. 123-141). Lima: Fondo Editorial PUCP.

Quintanilla, Pablo (2015). Verdad y justificación. Los límites del etnocentrismo. En Pablo Quintanilla y Claudio Viale (eds.), *El pensamiento pragmatista en la actualidad: conocimiento, lenguaje, religión, estética y política* (pp. 295-317). Lima: Fondo Editorial PUCP.

Quintanilla, Pablo (2016a). Interpretation and Intentional Actions. Three Non-Reducible Features. *Analítica,* 9, 113-129.

Quintanilla, Pablo (2016b). Nosotros contra ustedes: evolución y desarrollo de la moral. En Adolfo Chaparro Amaya, Bert Van Roermund y Wilson Herrera Romero (eds.), *¿Quiénes somos nosotros? O cómo (no) hablar en primera persona del plural* (pp. 43-60). Lima: Fondo Editorial PUCP.

Quintanilla, Pablo (2017a). Agencia, voluntad y autoconocimiento. En Ismael Muñoz, Marcial Blondet y Gonzalo Gamio (eds.), *Ética, agencia y desarrollo humano* (pp. 25-38). Lima: Fondo Editorial PUCP.

Quintanilla, Pablo (2017b). Atención compartida, triangulación y la perspectiva de la segunda persona. En Diana Inés Pérez y Diego Lawler (eds.), *La segunda persona y las emociones* (pp. 141-166). Buenos Aires: SADAF.

Quintanilla, Pablo; César Escajadillo & Richard Antonio Orozco, (2009). *Pensamiento y acción. La filosofía peruana a comienzos del siglo XX.* Lima: Fondo Editorial PUCP e Instituto Riva-Agüero (IRA).

Quintanilla, Pablo; Carla Mantilla & Paola Cépeda (eds.) (2014). *Cognición social y lenguaje. La intersubjetividad en la evolución de la especie y en el desarrollo del niño*. Lima: Fondo Editorial PUCP.

Quiroz, Rubén; Pablo Quintanilla & Joel Rojas (comps.). (2015). *Pedro Zulen. Escritos reunidos.* Lima: Fondo Editorial del Congreso del Perú.

Ramberg, Bjorn (1989). *Donald Davidson's Philosophy of Language*. Oxford: Basil Blackwell.

Ramberg, Bjorn (1997). The Source of the Subjective. En Lewis Edwin Hahn (ed.), *The Philosophy of Hans-Georg Gadamer* (pp. 459-471). The Library of Living Philosophers, Vol. XXIV. Chicago: Open Court.

Rawls, John (1971). *A Theory of Justice.* Massachusetts: Harvard University Press.

Renz, Ursula (ed.) (2017). *Self-Knowledge. A History*. Oxford: Oxford University Press.

Rescher, Nicholas (1958). *On Induction*. Cambridge: Cambridge University Press.

Rescher, Nicholas (1980). Conceptual Schemes. En Peter French, Theodor Uehling y Howard Wettstein (eds.), *Midwest Studies in Philosophy,* 5. *Studies in Epistemology* (pp. 323-346). Minneapolis: University of Minnesota Press.

Rescher, Nicholas (2014). *Metaphilosophy. Philosophy in Philosophical Perspective.* Maryland: Lexington Books.

Richards, Ivor A. (1936). *The Philosophy of Rhetoric*. Oxford: Oxford University Press.

Ricœur, Paul (1981). The Metaphorical Process as Cognition, Imagination and Feeling. En Mark Johnson (ed.), *Philosophical Perspectives on Metaphor* (pp. 228-247). Mineápolis: University of Minnesota Press.

Ricœur, Paul (1986 [1975]). *The rule of Metaphor. Multidisciplinary Studies of the Creation of Meaning in Language*. Londres: Routledge and Kegan Paul.

Rivarola, José Luis (1991). *Signos y significados. Ensayos de semántica lingüística*. Lima: Fondo Editorial PUCP.

Rizzolatti, Giacomo & Laila Craighero (2004). The Mirror-neuron System. *Annual Review of Neuroscience, 27*(1), 169-192.

Root, Michael (1992). Davidson and Social Science. En Ernest Lepore (ed.), *Truth and Interpretation. Perspectives on the Philosophy of Donald Davidson* (pp. 272-306). Cambridge: Blackwell

Rorty, Richard (1980). A Reply to Dreyfus and Taylor. *The Review of Metaphysics*, *XXXIV*(1), 39-46.

Rorty, Richard (1989 [1979]). *La filosofía y el espejo de la naturaleza.* Madrid: Cátedra.

Rorty, Richard (1991a). Inquiry as Recontextualization: An Anti-Dualist Account of Interpretation. En *Objectivity, Relativism and Truth*. *Philosophical Papers Volume One* (pp. 93-112). Cambridge: Cambridge University Press.

Rorty, Richard (1991b). *Objectivity, Relativism and Truth*. *Philosophical Papers Volume One.* Cambridge: Cambridge University Press.

Rorty, Richard (1991c). Solidarity or Objectivity? En *Objectivity, Relativism and Truth*. *Philosophical Papers Volume One* (pp. 21-34). Cambridge: Cambridge University Press.

Rorty, Richard (1991d). Unfamiliar Noices: Hesse and Davidson on Metaphor. En *Objectivity, Relativism and Truth*. *Philosophical Papers Volume One* (pp. 162-174). Cambridge: Cambridge University Press.

Ruiz-Bravo, Patricia; Pepi Patrón & Pablo Quintanilla (2009). *Desarrollo humano y libertades. Una aproximación interdisciplinaria*. Lima: Fondo Editorial PUCP.

Russell, Bertrand (1905). On Denoting. *Mind*, 14, 479-493.

Russell, Bertrand (1914a). On Scientific Method in Philosophy. En *Mysticism and Logic* (pp. 33-35). Londres: Allen & Unwin.

Russell, Bertrand (1914b). On the Nature of Acquaintance. *The Monist*, 24, 1-16, 161-187, 435-453.

Russell, Bertrand (1917). *Mysticism and Logic*. Londres: Allen & Unwin.

Russell, Bertrand (1957).The Relations of Sense-Data to Physics. En *Mysticism and Logic* (pp. 108-131). Doubleday: Anchor Books.

Russell, Bertrand (2017 [1919]). *Introduction to Mathematical Philosophy.* s.l.: Create Space Independent Publisher Platform.

Ryle, Gilbert (1949). *The Concept of Mind*. Londres: Hutchinson Hose.

Ryle, Gilbert (1967 [1949]). *El concepto de lo mental*. Buenos Aires: Paidós.

Sacks, Sheldon (ed.) (1979). *On Metaphor*. Chicago: University of Chicago Press.

Salmon, Merrilee (2002). La explicación causal en las ciencias sociales. En Wenceslao González (coord.), *Diversidad de la explicación científica* (pp. 161-179). Barcelona: Ariel.

Salmon, Wesley (ed.) (1971). *Statistical Explanation and Statistical Relevance.* Pittsburgh: University of Pittsburgh Press.

Salmon, Wesley (1988). *Causality and Explanation.* Nueva York: Oxford University Press.

Salmon, Wesley (1990). *Four Decades of Scientific Explanation.* Minneapolis: University of Minnesota Press.

Salmon, Wesley (1998).The Importance of Scientific Understanding. En *Causality and Explanation.* Nueva York: Oxford University Press.

Sapir, Edward (1921). *Language. An Introduction to the Study of Speech*. Nueva York: Harcourt Brace and Company.

Saussure, Ferdinand de (1980 [1916]). *Curso de lingüística general.* Buenos Aires: Losada.

Scanlon, Thomas (1998). *What We Owe to Each Other.* Massachusetts: Harvard University Press.

Schatzki, Theodor (1966). *Social Practices. A Wittgensteinian Approach to Human Activity and the Social*. Cambridge: Cambridge University Press.

Scheffler, Israel (1967). *Science and Subjectivity*. Nueva York: Bobbs Merrill.

Schilbach, Leonhard; Bert Timmermans & Tobias Schlicht (2013). Toward a Second Person Based Neuroscience. *Behavioral and Brain Sciences*, *36*(4), 393-462.

Schleiermacher, Friedrich (1986). *Hermeneutics. The Handwritten Manuscript*. Edición de Heinz Kimmerle. Atlanta: Scholars Press.

Schleiermacher, Friedrich (1996). *Schriften*. Edición de Andreas Arndt. Frankfurt: Deutscher Klassiker Verlag.

Scotto, Carolina (2002). Interacción y atribución mental: la perspectiva de la segunda persona. *Análisis Filosófico, XXII*(2), 135-151.

Scotto, Carolina (2004). La simpatía y la perspectiva de la segunda persona. *Epistemología e historia de la ciencia*, *10*(10), 485-491.

Searle, John (1983). *Intentionality: An Essay in the Philosophy of Mind.* Nueva York: Cambridge University Press.

Searle, John (1992). *The Rediscovery of the Mind.* Massachusetts: The MIT Press.

Searle, John (1995). *The Construction of Social Reality.* Nueva York: Free Press.

Seemann, Axel (ed.) (2011). *Joint attention. New Developments in Psychology, Philosophy of Mind and Social Neuroscience.* Massachusetts: The MIT Press.

Sen, Amartya (2000). *Desarrollo y libertad.* Buenos Aires: Planeta.

Shakespeare, William (1993). Romeo and Juliet. En *The Complete Works.* Londres: Magpie Books.

Shapere, Dudley (1981). Meaning and Scientific Change. En Ian Hacking (1981). *Scientific Revolutions* (pp. 28-59). Oxford: Oxford University Press.

Sherman, Nancy (1998). Empathy and Imagination. *Midwest Studies in Philosophy*, XXII, 82-119.

Simplicio (2012). *On Aristotle Physics 8.1-5.* Bristol: Bristol Classics.

Singer, Peter (1981). *The Expanding Circle. Ethics, Evolution and Moral Progress*. Princeton: Princeton University Press.

Slade, Arietta (2005). Parental Reflective Functioning: An Introduction. *Attachment and Human Development*, *7*(3), 269-281.

Smith, Adam (1976 [1759]). *The Theory of Moral Sentiments*. Indianápolis: Liberty Classics.

Smith, Plinio & Waldomiro Silva Filho (eds.) (2005). *Significado, Verdade, Interpretacao: Davidson e a Filosofia.* São Paulo: Loyola.

Soames, Scott (2012). *Philosophical essays.* Princeton: Princeton University Press.

Solms, Mark & Oliver Turnbull (2002). *El cerebro y el mundo interior. Una introducción a la neurociencia de la experiencia subjetiva.* Bogotá: Fondo de Cultura Económica.

Sosa, Ernest (1984). Mind-Body Interaction and Supervenient Causation. *Midwest Studies in Philosophy,* IX, 271-283.

Spaulding, Shannon (2018). *How We Understand Others: Philosophy and Social Cognition*. Nueva York: Routledge.

Spelke, Elizabeth Shilin (1988). *The Origins of Physical Knowledge.* En Lawrence Weiskrantz (ed.), *Thought without Language* (pp. 168-184). Oxford: Clarendon Press.

Sperber, Daniel (1985). *On Anthropological Knowledge.* Cambridge: Cambridge University Press.

Sperber, Daniel (1996). *Explaining Culture.* Oxford: Blackwell.

Spinoza, Baruch (2002). *Spinoza: Complete Works*. Editado por Michael L. Morgan. Indianápolis y Cambridge: Hackett Publishing Company.

Sripada, Chandra Sekhar (2008). Nativism and Moral Psychology: Three Models of the Innate Structure that Shapes the Contents of Moral Norms. En Walter Sinnott-Armstrong (ed.), *Moral Psychology.* Volumen 1. *The Evolution of Morality: Adaptation and Innateness* (pp. 319-344). Massachusetts: The MIT Press.

Sterelny, Kim (1981). Davidson on Truth and Reference. *Southern Journal of Philosophy*, 19, 95-116.

Stern, Daniel (1991). *El mundo interpersonal del infante. Una perspectiva desde el psicoanálisis y la psicología evolutiva*. Buenos Aires: Paidós.

Stich, Stephen (1993). Moral Philosophy and Mental Representation. En Michael Hechter, Lynn Nadel y Richard Michod (eds.), *The Origin of Values* (215-228). Nueva York: Aldine de Gruyter.

Stich, Stephen & Ian Ravenscroft (1994). What is Folk Psychology. *Cognition,* 50, 457-459.

Stich, Stephen & Shaun Nichols (1995). Folk Psychology: Simulation or Tacit Theory? En Martin Davies y Tony Stone (eds.), *Folk Psychology: The Theory of Mind Debate* (pp. 123-158). Oxford: Blackwell.

Strawson, Peter (1959). *Individuals.* Londres: Methuen.

Strawson, Peter (1985). *Skepticism and Naturalism: Some Varieties*. Nueva York: Columbia University Press.

Strawson, Peter (1995). *Libertad y resentimiento.* Barcelona: Paidós.

Strevens, Michael (2006). Scientific Explanation. En Donald Borchert (ed.), *The Encyclopedia of Philosophy*. Nueva York: Macmillan.

Stüber, Karsten (1993). *Donald Davidsons Theorie sprachlichen Verstehens.* Frankfurt: Anton Hein.

Tallerman, Maggie (2005). *Language Origins. Perspectives on Evolution*. Oxford: Oxford University Press.

Tarski, Alfred (1944). The Semantic Conception of Truth. *Philosophy and Phenomenological Research*, 4, 341-376.

Tarski, Alfred (1956). *Logic, Semantics, Metamathematics*. Oxford: Oxford University Press.

Taylor, Charles (1985). Interpretation and the Sciences of Man. En *Philosophy and the Human Sciences* (capítulo 1). Cambridge: Cambridge University Press.

Taylor, Charles (1995a). *Human Agency and Language*. *Philosophical Papers I*. Cambridge: Cambridge University Press.

Taylor, Charles (1995b). Language and Human Nature. En *Human Agency and Language*. *Philosophical Papers I* (pp. 83-87). Cambridge: Cambridge University Press.

Taylor, Charles (1995c). Theories of Meaning. En *Human Agency and Language*. *Philosophical Papers I* (capítulo 10). Cambridge: Cambridge University Press.

Tecumseh Fitch, William (2010). *The Evolution of Language*. Cambridge: Cambridge University Press.

Titchener, Edward (1909). *Experimental Psychology of the Thought Processes*. Nueva York: Macmillan.

Tomasello, Michael (1995). Language is not an instinct. *Cognitive Development,* 10, 131-156.

Tomasello, Michael (1999). *The Cultural Origins of Human Cognition*. Massachusetts: Harvard University Press.

Tomasello, Michael (2014). *A Natural History of Human Thinking*. Massachusetts: Harvard University Press.

Tomasello, Michael (2016). *A Natural History of Human Morality*. Massachusetts: Harvard University Press.

Verheggen, Claudine (2006). How Social Must Language Be. *Journal for the Theory of Social Behavior,* 36, 203-219.

Verheggen, Claudine (2007). Triangulating with Davidson. *Philosophical Quarterly,* 57, 96-113.

Verheggen, Claudine (ed.) (2017). *Wittgenstein and Davidson on Language, Thought and Action.* Cambridge: Cambridge University Press.

Vericat, José (ed. y trad.) (1988). *Charles S. Peirce. El hombre, un signo (El pragmatismo de Peirce)*. Barcelona: Crítica.

Vinden, Penelope (1996). Junin Quechua Children's Understanding of Mind. *Child development,* 67, 1707-1716.

Wallace, William (1974). *Causality and Scientific Explanation.* Michigan: Ann Arbor University of Michigan Press.

Warning, Rainer (ed.) (1989). *Estética de la recepción.* Madrid: Antonio Machado.

Weber, Max (1964 [1922]). *Economía y sociedad. Esbozo de una sociología comprensiva.* Ciudad de México: FCE.

Weber, Max (2013). *Ensayos sobre metodología sociológica.* Buenos Aires: Amorrortu.

Wellman, Henry; David Cross & Julanne Watson (2001). Meta-Analysis of Theory of Mind Development: The Truth about False-Belief. *Child Development, 72*(3), 655-684.

Wheeler, Samuel (1992). Indeterminacy of French Interpretation: Derrida and Davidson. En Ernest Lepore (ed.), *Truth and Interpretation. Perspectives on the Philosophy of Donald Davidson* (pp. 477-498). Cambridge: Blackwell

Whiten, Andrew (ed.) (1991). *Natural Theories of Mind. Evolution, Development and Simulation of Everyday Mindreading*. Oxford: Basil Blackwell.

Whorf, Benjamin Lee (1956). *Language, Thought and Reality*. Massachusetts: The MIT Press.

Williamson, Timothy (2008). *The Philosophy of Philosophy.* Oxford: Blackwell.

Wilson, Neil L. (1959), Substances without Substrata. *Review of Metaphysics*, 12, 521-539.

Wilson, Neil L. (1970). Grice on Meaning: The Ultimate Counter-Example. *Noûs*, 4970, 295-302.

Wilson, Timothy (2004). *Strangers to Ourselves: Understanding the Adaptive Unconscious.* Massachusetts: Harvard University Press.

Winch, Peter (1958). *The Idea of a Social Science.* Londres: Routledge and Kegan Paul.

Winch, Peter (1964). Understanding a Primitive Society. *American Philosophical Quarterly*, I, 307-324.

Winnicott, Donald (1971). *Playing and Reality.* Nueva York: Basic Books.

Wispé, Lauren (1986).The Distinction between Sympathy and Empathy. To Call Forth a Concept a Word is Needed. *Journal of Personality and Social Psychology, 20*(2), 314-321.

Wispé, Lauren (1987). History of the Concept of Empathy. En Nancy Eisenberg y Janet Strayer (eds.), *Empathy and its Development* (pp. 17-37). Cambridge: Cambridge University Press.

Wispé, Lauren (1991). *The Psychology of Empathy*. Nueva York: Plenum Press.

Wittgenstein, Ludwig (1958). *The Blue and Brown Books. Preliminary Studies for the «Philosophical Investigations».* Oxford: Basil Blackwell.

Wittgenstein, Ludwig (1972 [1969]). *Sobre la certidumbre.* Caracas: Tiempo Nuevo.

Wittgenstein, Ludwig (1975 [1922]). *Tractatus Logico-Philosophicus*. Madrid: Alianza.

Wittgenstein, Ludwig (1976). *Estética, psicoanálisis y religión*. Buenos Aires: Sudamericana.

Wittgenstein, Ludwig (1988 [1953]). *Investigaciones filosóficas*. Barcelona y Ciudad de México: Crítica y Universidad Nacional Autónoma de México.

Wright, Crispin; Barry Smith & Cynthia Macdonald (2006). *Knowing Our Own Minds.* Oxford: Oxford University Press.

Se terminó de imprimir en
los talleres gráficos de
Tarea Asociación Gráfica Educativa
Psje. María Auxiliadora 156, Breña
Correo e.: tareagrafica@tareagrafica.com
Teléfono: 332-3229 Fax: 424-1582
Se utilizaron caracteres
Adobe Garamond Pro en 11 puntos
para el cuerpo del texto
octubre 2019 Lima - Perú

Otras publicaciones del Fondo Editorial PUCP

Lingüística misionera
Aspectos lingüísticos, discursivos, filológicos y pedagógicos
Rodolfo Cerrón-Palomino, Álvaro Ezcurra Rivero, Otto Zwartjes (eds.)

Ni amar ni odiar con firmeza
Cultura y emociones en el Perú posbélico (1885-1925)
Francesca Denegri (ed.)

Nuevas aproximaciones a viejas polémicas: cine/literatura
Giovanna Pollarolo (ed.)

El cine de los maestros
Federico de Cárdenas

Gestión de la innovación empresarial: conceptos, modelos y sistemas
Jean Pierre Seclen Luna y Jon Barrutia Güenaga

Lecciones de derecho tributario: principios generales y código tributario
Sandra Sevillano Chávez

Seguridad internacional
Una introducción crítica
Farid Kahhat

Velasco, el fracaso de una revolución autoritaria
Luis Pásara

Los malentendidos y la incomprensión mutua son causa de muchos de los conflictos humanos. En el pasado, las personas o las comunidades que se malentendían podían alejarse a territorios apartados para evitar la convivencia. Esto es cada vez menos posible. Comprendernos recíprocamente no es, pues, una elección, sino una necesidad. Pero los fenómenos involucrados en la comprensión humana son múltiples y complejos, pues pueden ser procesos neuronales hereditarios que tienen millones de años, habilidades culturales aprendidas en una comunidad o técnicas individuales desarrolladas en la experiencia, entre otros.

Este libro se propone analizar lo que ocurre cuando nos comprendemos o, por el contrario, cuando nos malentendemos. Esto nos obliga a analizar otros fenómenos vinculados con la comprensión, como la explicación, la interpretación, la naturaleza de las comunidades epistémicas, el significado, la racionalidad, la irracionalidad y las formas de vida. Estas cuestiones no son solo relevantes para la filosofía sino también para la psicología, la ética, la política y las ciencias sociales, pues vivimos en un pequeño mundo en el que compartimos proyectos e infortunios con seres humanos que pertenecen a diferentes tipos de comunidades e identidades.

FONDO EDITORIAL
PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

ISBN 978-612-317-474-3