Juan Villacrés Bolaños
EDITOR



Introducción a la FILOSOFÍA ANALÍTICA

VOLUMEN 1

COLECCIÓN FILOSÓFICA ACTUAL

Editorial Universitaria
Universidad Central del Ecuador

Filosófica Editorial

Facultad de Ciencias Sociales y Humanas
Universidad Central del Ecuador

INTRODUCCIÓN A LA FILOSOFÍA ANALÍTICA

La filosofía ocupa un lugar insustituible a la hora de cuestionar los prejuicios y certezas impuestas por el sentido común, las creencias culturales y las costumbres. En lugar de ofrecer respuestas definitivas, expande horizontes y cultiva la duda crítica. Nos libera del dogmatismo y mantiene viva nuestra capacidad de asombro, al estar vinculada al mundo, a la vida, provocando sorpresa y enriqueciendo nuestras percepciones. Además, revela aspectos inadvertidos de lo cotidiano, profundizando nuestra comprensión más allá de lo evidente o asumido como incuestionable.

En un contexto generalizado de fragmentación, desinformación y agotamiento social, el pensamiento filosófico se convierte en un acto de resistencia creativa, permitiendo concebir posibilidades más allá de los límites definidos por las estructuras dominantes de la época.

La filosofía, lejos de ser un artefacto puramente intelectual, es una práctica indispensable y transformadora. La COLECCIÓN FILOSÓFICA ACTUAL surge como una apuesta crucial para impulsar la reflexión crítica desde Ecuador hacia América Latina y otras regiones del mundo. Esta colección busca consolidar un espacio de diálogo interdisciplinario, convirtiendo la filosofía en un recurso activo ante las crisis contemporáneas. Así, la filosofía asume su tarea esencial: liberarnos de certezas que limitan nuestra comprensión, permitiéndonos no solo habitar el presente con mayor lucidez, sino también imaginar y construir futuros posibles de forma reflexiva.

COLECCIÓN FILOSÓFICA ACTUAL
Facultad de Ciencias Sociales y Humanas
Universidad Central del Ecuador

Prof. Santiago M. Zarria Prof. Martín Aulestia Calero

Juan Villacrés Bolaños
(Editor)

INTRODUCCIÓN A LA FILOSOFÍA ANALÍTICA
Volumen I

Alejandro Tomasini
Ricardo Mena
Mario Gómez Torrente
Raymundo R. Meza
Siddhant Khamkar
Miguel Ángel Fernández-Vargas
Santiago Echeverri
Fernando Rudy Hiller
Luis Francisco Estrada Pérez

COLECCIÓN FILOSÓFICA ACTUAL

FILOSÓFICA EDITORIAL
UNIVERSIDAD CENTRAL DEL ECUADOR FACULTAD DE CIENCIAS SOCIALES Y HUMANAS
EDITORIAL UNIVERSITARIA UNIVERSIDAD CENTRAL DEL ECUADOR

COLECCIÓN FILOSÓFICA ACTUAL

La Colección Filosófica Actual constituye un proyecto editorial y de investigación desarrollado entre Filosófica, Fundación de Estudios Filosóficos, Políticos y Culturales, y la Facultad de Ciencias Sociales y Humanas de la Universidad Central del Ecuador.

INTRODUCCIÓN A LA FILOSOFÍA ANALÍTICA
Volumen I

Juan Villacrés Bolaños
Editor

1.ª edición | agosto de 2026
ISBN | 978-9907-840-01-8

EDITORIAL UNIVERSITARIA
Universidad Central del Ecuador
Ciudadela Universitaria, av. América, s. n.
Quito, Ecuador
editorial@uce.edu.ec

FILOSÓFICA EDITORIAL
Filosófica, Fundación de Estudios Filosóficos, Políticos y Culturales
Quito, Ecuador
filosofica@fundacionfilosofica.com
www.fundacionfilosofica.com

FACULTAD DE CIENCIAS SOCIALES Y HUMANAS
Universidad Central del Ecuador
COLECCIÓN FILOSÓFICA ACTUAL
CONSEJO EDITORIAL
Santiago M. Zarria
Goethe-Universität, Frankfurt am Main, Alemania | Universidad Central del Ecuador
Martín Aulestia Calero
Universidad Central del Ecuador

FILOSÓFICA
Fundación de Estudios Filosóficos, Políticos y Culturales

CONSEJO ACADÉMICO
Fernando Barredo
Pontificia Universidad Católica del Ecuador
David Cortez
RWU Hochschule Ravensburg-Weingarten, Alemania
Fernando Wirtz
Universität Tübingen, Alemania | Universidad de Kioto
Micaela A. Díaz,
Lee University, Cleveland, TN. EE.UU.

Publicación apoyada por:
- ICALA, Stipendienwerk Lateinamerika-Deutschland e.V.
- Facultad de Ciencias Sociales y Humanas de la Universidad Central del Ecuador

DERECHOS Y ACCESO
Esta publicación se encuentra disponible para consulta y descarga gratuita. Los contenidos pueden utilizarse para fines no comerciales, siempre que se reconozca adecuadamente la autoría y se cite la fuente original. En caso de realizar modificaciones, adaptaciones o transformaciones, deberá preservarse el principio de libre acceso al contenido.
La autorización de acceso abierto no implica renuncia a los derechos morales de autor ni permite atribuir a los autores, editores o instituciones participantes modificaciones realizadas por terceros.

Índice

INTRODUCCIÓN.....11

PRIMERA SECCIÓN
FILOSOFÍA DEL LENGUAJE Y SU INFLUENCIA EN LA CORRIENTE

CAPÍTULO I. LÓGICA Y EMPIRISMO: ALGUNAS IMPLICACIONES
DE LA TEORÍA DE DESCRIPCIONES.....31
Alejandro Tomasini

CAPÍTULO II. CONVENCIONES LINGÜISTICAS.....53
Ricardo Mena

CAPÍTULO III. EL CONCEPTO DE VERDAD EN LA TRADICIÓN ANALÍTICA.
UNA MIRADA A LAS TEORÍAS PRINCIPALES.....89
Mario Gómez Torrente

CAPÍTULO IV. TEORÍA DE MODELOS.....127
Raymundo R. Meza

CAPÍTULO V. REVISITING THE CARNAP-QUINE DEBATE ON ANALYTICITY.....173
Siddhant Khamkar

SEGUNDA SECCIÓN
EPISTEMOLOGÍA ANALÍTICA

CAPÍTULO VI. TEORÍAS DE LA JUSTIFICACIÓN EPISTÉMICA.....211
Miguel Ángel Fernández-Vargas

CAPÍTULO VII. LAS PARADOJAS ESCÉPTICAS.....261
Santiago Echeverri

TERCERA SECCIÓN
ÉTICA ANALÍTICA

CAPÍTULO VIII. TRES TEORÍAS DE LA AGENCIA RESPONSABLE.....323
Fernando Rudy Hiller

CAPÍTULO IX. LA MÁQUINA DE EXPERIENCIAS DE NOZICK:
ALCANCE DE LA CRÍTICA AL HEDONÍSMO ÉTICO.....355
Luis Francisco Estrada Pérez

INTRODUCCIÓN

Hacia finales del siglo XIX, el idealismo era la corriente dominante entre los filósofos de habla inglesa. Como bien señaló Quinton (1971), su ingreso formal en la tradición británica puede fecharse al año 1876 y no fue hasta aproximadamente 1945 que los departamentos de filosofía dejaron de tener un profesorado que, mayoritariamente, se identificaba con el idealismo. En el ambiente intelectual de esa época se daba por hecho que la realidad era fundamentalmente inmaterial. Sin embargo, fue en 1903, año en el que Russell publicó The Principles of Mathematics y Moore su Refutation of Idealism cuando los inmaterialistas británicos sintieron la primera embestida. Según Candlish (2009), la batalla decisiva se daría entre Russell y Bradley en torno a la naturaleza de las relaciones, lo que a su vez enfrentaría el pluralismo de Russell (asumido en la obra mencionada) con el monismo de Bradley.

The Principles of Mathematics fue una obra de profunda influencia donde se integra a un argumento metafísico técnicas formales. Para darnos una idea de su relevancia consideremos la opinión que Popper compartió a Magee acerca de esta obra:

> The Principles of Mathematics, a mi parecer, es uno de los libros más grandiosos y asombrosos jamás escritos… Su logro no tiene parangón. Russell redescubrió la lógica y la teoría de números de Frege, y sentó las bases de la aritmética y el análisis. Ofreció la primera definición clara y sencilla de los números reales sobre esta base (una mejora con respecto a las de Cantor, Dedekind y Peano). Y no solo aportó una teoría de la geometría, sino también un nuevo enfoque a la mecánica… Fue el primero en ofrecer una teoría satisfactoria de los números irracionales, un problema que había intrigado a la humanidad durante 2400 años. Y dio una solución verdaderamente brillante al antiguo problema del movimiento[1]. (Magee, 1971, pp. 143-145)

1 Todas las traducciones de la introducción son mías.

En esta obra, Russell asume que cada proposición está conformada por las entidades invocadas por sus palabras (como si usara un lenguaje Lagadoniano). Donde, la proposición es en sí misma una unidad en la que las entidades que la conforman estarían como fundidas en ésta. Así, la proposición: <Pablo Palacio escribió Débora> es una entidad distinta al agregado de entidades: Pablo Palacio, escribir, y Débora. Esto se debe a que en su posición pluralista, los bloques fundamentales de la realidad son una diversidad de entidades que al relacionarse entre sí producen nuevas entidades; donde cada una de éstas en vez de ser un agregado, es en sí misma una unidad. Según Russell estas unidades pueden ser invocadas a través de proposiciones relacionales. Así, por ejemplo, en una proposición, el verbo es el que encarnaría la mencionada unificación. Sin embargo, Russell reconoce que la característica unificadora del verbo no abarca su naturaleza.

> El verbo, cuando se usa como verbo, encarna la unidad de la proposición y, por lo tanto, se distingue del verbo considerado como término, aunque no sé cómo explicar con claridad la naturaleza precisa de esa distinción. (Russell, 1903/1937, p. 50)

Precisamente, Bradley muestra la incompatibilidad entre la característica unificadora del verbo, y su rol como *término*. Esto ya que, dicho rol, sería el que tienen las relaciones externas. Una relación externa es aquella cuyo ser y existencia es independiente del relata (i.e. aquello que está siendo relacionado por la relación). Por ejemplo, si Juan ama a Joselyn, la relación *amor* sería externa si ésta puede existir y ser lo que es independientemente de si Juan y Josselyn existen. En Appearance and Reality, Bradley (1893) proponía que para que una relación externa R (una relación no instanciada) vincule a dos entidades A y B, ésta necesitaría de otras dos relaciones R_1 y R_2 que vinculen R con A, y R con B; respectivamente. Es decir, puesto que las naturalezas tanto de A, B como R carecen de algo que vincule estas entidades entre sí, entonces R necesitaría un par de relaciones extra para que A, B y R se vinculen. Sin embargo, si las relaciones son externas, cada relación extra necesitaría de un par

adicional de relaciones, y así sucesivamente ad infinitum. Siendo así, una relación externa *R* nunca podría terminar vinculando *A* y *B*.

En otras palabras, Bradley argumenta que no es posible construir a través de relaciones externas una entidad donde sus componentes estén fundidas formando una unidad. Así, Bradley manifiesta:

> La posición fundamental del señor Russell se me ha mantenido incomprensible. Por una parte, me veo llevado a pensar que defiende un pluralismo estricto, que no admite nada por fuera de términos simples y relaciones externas. Por otra parte, el señor Russell parece afirmar de modo enfático… ideas que un pluralismo semejante seguramente ha de repudiar. Él se apoya de manera constante en unidades que son complejas y que no se pueden analizar en términos y relaciones. A mi entender, estas dos posiciones son irreconciliables, pues la segunda, según mi comprensión, contradice de lleno a la primera. Si existen tales unidades, y, más aún, si tales unidades son fundamentales, entonces, sin duda, el pluralismo queda en principio abandonado por falso. (Bradley, 1910/1914, p. 281)

Russell por su parte contraatacó apuntando a las relaciones concebidas como internas—i.e. la postura en que la relación depende del relata--, puesto que consideró que el monismo de Bradley estaba soportado por la existencia de éstas. Según Russell, la posición del monista que supone que las relaciones son internas es la siguiente. Si una relación interna *R* está instanciada sobre un relata *a*,*b*, estos tres elementos forman un todo unificado, donde la proposición que apunta a esta totalidad no manifiesta algo acerca de *a* o *b*, sino que exhibe dicha totalidad. En sus propias palabras:

> La teoría monista sostiene que toda proposición relacional 'aRb' debe resolverse en una proposición referente al todo que a y b componen… Se nos dice, por parte de quienes defienden esta opinión [la referencia incluye a Bradley], que el todo contiene diversidad en su interior, que sintetiza diferencias… Por mi parte, me es imposible atribuir a estas frases ningún significado preciso. (1903/1937, pp. 219-220)

Luego de esta afirmación, Russell considera que la relación asimétrica 'mayor que' en la proposición <a es mayor que b>, y expresa que, para los monistas mencionados, dicha proposición en realidad se entendería como <La totalidad formada por *a* y *b* contiene diversidad de magnitud>. Esta segunda proposición, sin embargo, es incapaz de reflejar la asimetría que, por otro lado, la primera sí exhibe. Notemos que, en una relación asimétrica no se puede cambiar el orden del relata sin modificar su valor de verdad. De manera que: si *a*>*b* es verdadero, entonces *b*>*a* será falso. En este sentido las relaciones internas no podrían garantizar la existencia de relaciones asimétricas. Por lo que, sería erróneo considerar que las relaciones internas son esenciales para la naturaleza de la realidad.

Sin embargo, concuerdo con Candlish (2009) en que Bradley no defendía la existencia de relaciones internas. El, más bien, consideraba que las relaciones son irreales. Es decir que, en principio sería inadecuado partir al Absoluto (la unidad que es lo único que realmente existe en su monismo) en distintas partes y erróneamente categorizarlas como fundamentales, de manera que estas al relacionarse produzcan situaciones o hechos. En palabras de Bradley:

> En resumidas cuentas, lejos de admitir que el Monismo requiere que todas las verdades puedan interpretarse como la predicación de cualidades del Todo, el Monismo, tal como yo lo concibo, sostiene que toda predicación, sin importar cuál sea, es, en última instancia, falsa e irreal. (Bradley, 1935, p. 672)

Aquí, vale notar que, Russell (1903) consideraba que toda proposición expresa la existencia de una relación. Por lo que, en la cita anterior, Bradley al responder a Russell y manifestar la irrealidad de la predicación (a través de proposiciones) está expresando la irrealidad de las relaciones. En este sentido, la postura de Bradley no estaría siendo atacada por los argumentos de Russell en contra de las relaciones internas. Tomando en cuenta esto, concuerdo con Candlish (2009) en que para explicar el salto del idealismo a la corriente analítica en la filosofía anglo-parlante no es suficiente el relato estereotípico (como él le llama) en el que, Rusell ganó la disputa acerca

de la naturaleza de las relaciones, y al hacerlo supera al idealismo consolidando la corriente analítica. En este sentido prefiero tomar una posición cauta en la línea de Lewis cuando afirma:

> Las teorías filosóficas nunca son refutadas de manera concluyente. La teoría sobrevive a su refutación a un precio… Pero, al final del día, una vez descubiertos todos los argumentos astutos, las distinciones y los contraejemplos, es de prever que aún nos enfrentaremos a la pregunta de qué precios vale la pena pagar. (Lewis, 1983, p. x)

Según Candlish (2009) los factores que van más allá de los argumentos esgrimidos en los debates mencionados incluyen, por ejemplo, la posición de Russell al abandonar el cristianismo y la adopción momentánea del idealismo. Los aspectos místicos de dicha corriente pudieron tentar a Russell en su dimensión espiritual. En el ensayo 'Seems, ¿Madam? Nay, it is' presentado a la sociedad Cambridge Apostles en 1899, Russell sugiere que no pudo encontrar en la metafísica idealista el consuelo que de otra forma hubiese buscado en la religión, y aquello fue parte del desencanto con dicha corriente.

> Se supone que McTaggart aún atribuye a la filosofía el poder de brindar consuelo y alivio —el último poder de los impotentes. Es esta última posesión de la que, esta noche, deseo despojar a la decrépita progenitora de nuestros dioses modernos (Russell, 1899/2004, p. 49)

En particular, la armonía de la única entidad realmente existente según el idealismo monista, con el que Russell coqueteó brevemente, no se alcanzaría ni después de morir, ni en el presente. Esto ya que, en tal postura idealista, nuestra cognición solo nos da acceso a la apariencia de lo plural, mientras lo real estaría fuera de nuestro alcance en la vida mundana. Así, Russell menciona:

> Incluso el puro interés intelectual, del cual surge la metafísica, es un interés por explicar el mundo de la apariencia. Pero, en lugar de explicar de verdad este mundo sensible, palpable y actual, la metafísica construye otro mundo fundamentalmente diferente; tan distinto, tan desconectado de la experiencia real, que el mundo de la vida cotidiana permanece

> completamente inafectado por él y sigue su curso como si no existiera mundo de la Realidad en absoluto. (Russell, 1899/2004, p. 53)

Adicionalmente, Candlish (2009) identifica algunos factores de orden socio-político. Señala, por ejemplo, que el impacto de la primera guerra mundial desmotivó a la gente a tener una aproximación inmaterial o espiritual a la realidad. Así como también apunta al declive del imperio británico cuya dirección moral y espiritual estaba soportada por algunos idealistas hegelianos, y señala también el recelo, postguerra, a la influencia de la filosofía idealista prusiana en el pensamiento anglófono. Estos factores se añaden a la fuerte influencia que causó Russell con su aprecio a las matemáticas, y el ya mencionado distanciamiento del idealismo.

Sin embargo, más allá de todos los factores mencionados hasta aquí, fue la aproximación Russelliana a los problemas filosóficos donde se valora y adopta el lenguaje, la metodología, y la claridad de la ciencia lo que marcó la dirección de la corriente analítica. Russell visionó una nueva forma de hacer filosofía con una precisión y claridad de pensamiento y expresión sin paralelo. Tomemos, por ejemplo, el clásico On Denoting, donde Russell arroja luz sobre aspectos ante los que el lenguaje hegeliano no ofrecía respuesta. El siguiente extracto ilustra el tipo de claridad que busca la corriente analítica, donde la teoría de la denotación resuelve el problema sugerido a continuación:

> En virtud del principio del tercero excluido, o bien "A es B" o bien "A no es B" debe ser verdadero. Por lo tanto, o bien "el actual rey de Francia es calvo" o bien "el actual rey de Francia no es calvo" debe ser verdadero. Sin embargo, si enumeráramos las cosas que son calvas, y luego las cosas que no son calvas, no encontraríamos al actual rey de Francia en ninguna de las dos listas. Los hegelianos, que aman las síntesis, probablemente concluirán que lleva peluca. (Russell, 1905, p. 485)

El desarrollo de la lógica proposicional por parte de Frege, junto con los aportes que, sobre ésta, realizaron tanto Russell como Whitehead dotaron de un poder expresivo y deductivo a la corriente

analítica, que no estaba disponible previamente. Estos desarrollos han sido seminales para campos como la metalógica—la cual por e.g. permitió el desarrollo de los teoremas de incompletitud de Gödel--, las ciencias cognitivas, o la inteligencia artificial.

En la actualidad, la corriente analítica es fundamental en las facultades de filosofía en el mundo anglófono. Por ejemplo, según Philosophical Gourmet Report (2025) editado por Brian Leiter, todas las facultades de filosofía de las universidades Ivy League se autodenominan como analíticas. Los intereses de la corriente se han ampliado al punto de que ahora ya se han trabajado sobre temas que en los primeros años de la corriente eran abordados casi exclusivamente por otras corrientes filosóficas. Por ejemplo, ciertos aspectos de la filosofía budista han sido tratados formalmente por dialeteistas como Graham Priest (2014), lo inefable ha sido trabajado con gran claridad por Silvia Jonas (2016), o el monismo ha sido defendido por Schaffer (2010).

La constitución de estos dos volúmenes es parte de un esfuerzo por despertar interés en la filosofía analítica en las personas de habla hispana que no están aún familiarizadas con esta corriente. Sobre todo, se busca dotar de algunas herramientas agudas de pensamiento a sus lectoras y lectores; las cuales, sin duda, sirven como bases para profundizar en la forma analítica de tratar los problemas filosóficos. Si bien es cierto, la constante amenaza de los conflictos sociales, ambientales y políticos que se cierne sobre Latinoamérica sugiere a la corriente continental como el campo teórico sobre el que se piensen estos conflictos, sin embargo, la filosofía analítica provee de una precisión que beneficiaría la argumentación y análisis de éstos. De hecho, cuando propuse la conformación de esta obra, tenía en mente más que nada el entregarse con la mente clara a los distintos problemas de la realidad y a la amplitud de la experiencia humana, la cual no se limita a lo social o político.

El libro tiene la vocación de ser integral, los primeros cinco capítulos abordan principalmente aspectos fundamentales de la filosofía del lenguaje y cómo ésta influyó en el desarrollo de la corriente. En esta primera parte se encuentran los capítulos con más

formalismo, en particular al tratar el tema de la verdad y la teoría de modelos. Los capítulos sexto y séptimo tratan temas epistemológicos, mientras que el octavo y noveno abordan aspectos de la filosofía de la acción y ética.

Este volumen empieza con el capítulo de Alejandro Tomasini quien examina los presupuestos e implicaciones en la teoría de las descripciones de Russell, la cual es presentada en el ya mencionado y célebre artículo On Denoting. En esta obra se exhibe un proceso (o intento) de clarificación del lenguaje natural por medio del lenguaje lógico. Éste consiste en presentar en términos de propiedades aquello que el lenguaje natural exhibe como objetos. Según Russell las personas que participan de la comunicación tendrían conocimiento directo de dichas propiedades puesto que éstas corresponderían al *sense-data*. Esto a su vez, según Tomasini, implicaría una posición empirista radical, en la que el conocimiento humano se funda en la experiencia sensible. Sin embargo, el proceso de reducción de objetos a propiedades puede hacernos pensar que el lenguaje natural es incapaz de darnos a conocer aquello de lo que hablamos. Posición que, de acuerdo a Alejandro, estaría motivada por un énfasis en el tratamiento formal del lenguaje que puede perder de vista cómo en efecto funciona el lenguaje natural en el día a día. En sus palabras "lógica sin contextualización es filosóficamente equívoca" (pp. 50-51). El capítulo de Tomasini nos invita a considerar los puntos ciegos que podría tener una aproximación a los problemas de la filosofía del lenguaje abordada exclusivamente en términos formales.

En el segundo capítulo, Ricardo Mena presenta un análisis del lenguaje en el que éste adquiere su significado en virtud de actividades comunales. Para esto, Mena, retoma la observación de Quine en la que sería imposible que el significado del lenguaje sea establecido únicamente por acuerdos. Esto ya que, en principio, se necesitaría ya de un lenguaje para establecer dichos acuerdos. Esta objeción es superada por la propuesta de Lewis quien sostiene que el significado se establece por medio de convenciones. Aquí, una convención no es un acuerdo explícito, sino más bien una

solución—auto perpetuada y arbitraria—a problemas de coordinación. Donde dichos problemas se darían cuando el emisor y la audiencia interpretan la emisión lingüística de manera distinta. Mena, además, examina cómo la teoría de Lewis enfrenta algunas posibles objeciones como la de la innovación léxica planteada por Davidson; en la que una palabra que aún no cuenta con significado convencional es usada dentro de una comunicación exitosa. En este sentido, Mena nos muestra un análisis del lenguaje donde el énfasis está en la función que éste desempeña al concretarse a través de la coordinación de acciones. Esto en contraste a la aproximación al lenguaje examinada en el primer capítulo donde lo relevante para el significado es a qué apuntan las palabras.

Con el texto de Mario Gómez Torrente, en el tercer capítulo, entramos en una cuestión fundamental y transversal a la tradición analítica: la verdad. Aquí, se exhibe con maestría una cartografía de aproximaciones metafísicas como la teoría correspondentista—trazada hasta al *dictum* Aristotélico donde expresar algo verdadero es manifestar lo que es el caso--, así como también perspectivas epistemológicas como la coherentista en la que una proposición es verdadera en la medida que sea coherente con otras proposiciones. También se examinan las teorías pragmatistas, redundantistas, y pro-oracionales. Estas dos últimas consideran que expresiones como 'es verdad' no manifiestan una propiedad sustantiva (o metafísica) de las oraciones; sino más bien exhiben una propiedad insustantiva. Dentro de las teorías insustantivas, Mario Gómez, enfatiza en la teoría semántica de la verdad de Tarsky, y en la teoría de Kripke. Éstas superan los retos que plantea la paradoja del mentiroso—i.e. aquellos que se presentan al tratar de evaluar el valor de verdad de proposiciones como <Esta oración no es verdadera>. La preeminencia de estas teorías tanto en la lógica matemática como en la teoría de modelos se explica, en parte, por la caracterización de la verdad en términos puramente matemáticos. Así, por ejemplo, la teoría semántica mencionada plantea un método de construcción de un predicado basado en la noción de satisfacción de una expresión matemática; donde dicho predicado es

coextensional (i.e. aplica a lo mismo) al predicado 'es verdadera en un lenguaje formal L.'

El capítulo cuarto de Raymundo Meza enfatiza en un aspecto del espíritu analítico para abordar distintos problemas filosóficos, aquel que valora los análisis formales. En particular, Meza hace un breve recorrido por un campo que se encuentra en la intersección entre la semántica y la lógica matemática; a saber: la teoría de modelos. Así, sienta las bases conceptuales y formales para la exposición del problema de la inescrutabilidad de la referencia desde el punto de vista de dicha teoría. De la misma forma que es problemático conocer a qué se está refiriendo el término 'gavagai' al emitirse cuando se ve un conejo mientras se apunta al sector donde éste se encuentra—ya que 'gavagai' podría referirse e.g. a una colección de partes (cola, orejas, etc.) en contraste a un único individuo--, así las teorías (en particular las expresadas en lenguajes donde los elementos básicos son individuos y no predicados) presentan dificultades para informar de manera exclusiva y exhaustiva acerca de los referentes y relaciones que satisfacen dichas teorías. Al respecto Meza afirma: "si la relación entre el lenguaje y el mundo no es una correspondencia directa y nuestras descripciones del mundo no pueden capturar una única estructura ontológica subyacente, entonces la existencia de múltiples modelos nos ofrece diversas posibilidades de conceptualizar la realidad" (p. 169). En la filosofía de las matemáticas la inescrutabilidad de la referencia puede evidenciarse e.g. en la dificultad de determinar la identidad de los números. En su recorrido por la teoría de modelos, Meza, exhibe como la teoría de modelos brinda un fundamento riguroso a nociones tan centrales como satisfacibilidad o consecuencia lógica, donde los procesos de definición rigurosa son inspirados en la teoría semántica de la verdad de Tarsky ya examinada en el tercer capítulo.

En el capítulo quinto, Siddhant Khamkar explora el debate entre Carnap y Quine acerca de la distinción entre oraciones analíticas y sintéticas. Los juicios analíticos, según Kant, son aquellos en los que el predicado está contenido en el sujeto; mientras que, en los sintéticos el predicado no se encuentra ya en el sujeto.

Khamkar arroja luz sobre los detalles epistemológicos y metodológicos de las propuestas tanto de Carnap como de Quine examinando el debate entre estos filósofos más allá del análisis estereotípico (uno fuertemente influenciado por la lectura de Quine). Para Carnap la noción de analiticidad es substancial para construir lenguajes formales, mientras que para Quine dicha noción ni siquiera es clara. En particular, Quine argumenta contra la inteligibilidad de la noción de analiticidad en su célebre 'Two Dogmas of Empiricism' indicando, por ejemplo, la circularidad que implicaría explicar dicha noción en términos de substitución por sinonimia. Así, el juicio analítico: <Ningún soltero está casado> (aquí, el predicado estaría implicado en el sujeto) no podría entenderse como <Ninguna persona no-casada está casada> (lo cual es lógicamente verdadero independientemente de la interpretación de los términos no lógicos) sin caer en una explicación circular que considera la sinonimia entre 'persona no-casada' y 'soltero'. El análisis de Quine lo lleva a defender una posición holista epistemológica en la que las oraciones que componen nuestras teorías científicas son una red interconectada de creencias; donde, ante una posible evidencia empírica que rete la teoría, se buscaría replantear sus oraciones concebidas dentro de dicha red, y no de manera aislada—como supuestamente estaría planteando la posición epistemológica de Carnap. Khamkar exhibe tanto los puntos comunes en las posiciones epistemológicas de ambos filósofos como sus diferencias, además de clarificar sus motivaciones. En palabras de Siddhant "Quine pretende investigar cómo se lleva a cabo realmente la teorización científica, mientras que Carnap busca analizar la estructura lógica de las teorías y formular conceptos más sistemáticos mediante la construcción de lenguajes artificiales en un intento de aclarar y resolver problemas filosóficos" [mi traducción] (p. 204).

Miguel Ángel Fernández, en el capítulo sexto, nos da un recorrido detallado por aspectos fundamentales de la epistemología analítica, con énfasis en la justificación epistémica de las creencias. Fernández explica con gran claridad distintas respuestas a preguntas como ¿Qué significa que una creencia está justificada?

o ¿Es posible justificar una creencia? A breves rasgos, para la primera pregunta, algunas opciones incluyen afirmar que una creencia tiene justificación en caso: cuente con evidencia, o el método para adquirirla es fiable, o ésta ayuda a maximizar el conjunto de creencias verdaderas. Por otra parte, la segunda pregunta puede surgir al considerar el Trilema de Agripa; donde uno de sus lemas implicaría que una creencia necesitaría justificarse en otra, y ésta en otra, en una cadena de justificaciones que no termina nunca; esto evitaría que exista una creencia primera cuya justificación no necesite descansar en otra creencia ya justificada. Sin dicha creencia inicial, sugiere el lema, no podríamos afirmar a cabalidad que alguna creencia está justificada. Miguel Ángel explica cómo ha sido respondido, el reto planteado por este tipo de problema, por tres posturas distintas fundacionismo, coherentismo e infinitismo. La primera, por ejemplo, rechaza la suposición del lema mencionado, aquella en la que una creencia justificada necesita estar soportada en alguna otra ya justificada. A lo largo del capítulo, Miguel Ángel nos explica distintos aspectos imprescindibles de la justificación sentando las bases para adentrarse en futuros debates. Por ejemplo, muestra la diferencia entre estar acreditado a creer que e.g. el piso va a seguir existiendo en el siguiente instante (creencia que podría argumentarse no descansa en evidencia), y entre estar justificado en que creer que el piso sobre el que estoy parado existe. Así también explica ciertas diferencias conceptuales como las que hay entre conservadurismo y liberalismo epistémico; los liberales, en contraste a los conservadores, aceptan que la memoria y la percepción cuentan como evidencia inmediata de ciertas creencias. En este sentido, el conservadurismo "comparte el espíritu de la posición [escéptica]: sostener que para estar justificado en creer que tengo manos es necesario tener justificación para creer que no soy un [cerebro en una cubeta]" (p. 229); aquí Miguel Ángel se refiere a un cerebro estimulado neuronalmente para reproducir experiencias en primera persona.

El escepticismo en clave analítica lo aborda Santiago Echeverri. En el séptimo capítulo detalla con agudeza un enfoque en el que el

escepticismo toma un rol central en la epistemología. Aquí, éste no es entendido como una posición guiada por una actitud de sospecha que lleva a debates infructuosos. Sino más bien, Echeverri nos muestra cómo el contradecir una posición ampliamente aceptada exhibe la extensión del ecosistema de justificaciones y conocimientos vinculados a nuestro sentido común. En este caso son las paradojas las que ayudan a determinar dicha extensión, donde su estructura, puede observarse en el siguiente ejemplo. Supongamos que te gustaría afirmar que sabes que <estás leyendo esto>, sin embargo, para esto debes ser capaz de descartar que eres <un cerebro estimulado neuronalmente de manera artificial para reproducir la experiencia de leer esto, aunque en realidad no estés leyendo nada> (i.e. el clásico cerebro en una cubeta C.C., mencionado también en el capítulo anterior). Puesto que el escenario propuesto por esta última proposición es difícil de descartar, existe una tensión entre dicha dificultad y aseverar que sabes que <lees esto>. En general, las soluciones a las paradojas escépticas consideran relajar las condiciones necesarias y suficientes para poder aseverar que se posee justificación de lo que se cree o conoce. Así, por ejemplo, una solución plantea que estarías justificado en el contexto de la vida diaria en creer que <leíste esto>, mientras que en otro contexto en el que sea relevante que descartes que no eres un C.C., dicha creencia sería difícil de justificar. Podrías también considerar que, de entre todas las situaciones posibles aquellas donde eres un C.C. no son relevantes para justificar que crees que <leíste esto>. Este, por ejemplo, fuera el caso, si aquellas situaciones posibles donde eres un C.C. difieren notablemente de aquellas donde se tiene que: si crees que no eres un C.C., entonces no eres un C.C. (este tipo de solución se analiza en términos modales, a través de la noción de mundos posibles). Así, Santiago nos lleva por distintas respuestas al reto que plantea el escenario escéptico analizando los problemas asociados con cada una de estas. El autor nos invita a considerar cuáles dudas nos conciernen en distintos contextos, y bajo distintos estándares epistémicos. Así, las paradojas

escépticas no se presentan como obstáculos para el conocimiento, sino como elementos que lo aclaran.

En el capítulo octavo Fernando Hiller ilumina el campo de la responsabilidad moral contemporánea proponiendo una clasificación de tres familias de teorías que explican el rol de la capacidad psicológica de los agentes, las evaluaciones, y reacciones ante sus actos dentro del ámbito moral. Tradicionalmente se ha dado por hecho que una persona que ejecuta una acción negativa amerita un castigo, si ésta pudo haber actuado de otra manera y no lo hizo; es decir, solo si la persona contaba con libre albedrío al momento de actuar. Sin embargo, tal como Hiller señala, a partir del trabajo de Strawson en 'Libertad y resentimiento' los criterios para determinar responsabilidad moral no se basan únicamente en la posesión de libre albedrío, sino en cómo se entienden y gestionan las reacciones ante un acto moralmente evaluable. La primera familia de teorías son las punitivistas, donde en general, la actitud reactiva ante un acto incorrecto es de naturaleza sancionatoria, si--antes de que éste sea ejecutado--el agente poseía capacidad de entender las implicaciones morales de su accionar de manera que podía evitar actuar incorrectamente. Aquí, por otro lado, Hiller argumenta que las reacciones a actos incorrectos no necesariamente se determinan en función de si el agente, antes de actuar, entendía sus obligaciones morales. En este sentido dichas reacciones no necesariamente serían un castigo por realizar una acción que el agente sabía que era incorrecta. La segunda familia corresponde a las teorías conversacionales en las que las actitudes reactivas son parte de una comunicación en el ámbito moral, donde es parte de dicha comunicación la capacidad de entender--posteriormente a ejecutar una acción incorrecta—las razones por las que ésta es incorrecta. Así, tener la disposición a, por ejemplo, pedir disculpas, o experimentar arrepentimiento serían parte del diálogo moral. Aquí, un agente sería apto de ser responsabilizado moralmente, si éste tiene la capacidad de tener una "conversación moral". La tercera familia son las teorías de calibración personal, en las que las actitudes reactivas buscan establecer cierto tipo de interacción

interpersonal, mas que sancionar, o comunicar. Por ejemplo, como menciona Hiller, "no queremos confiar en quienes no son de confianza." (p. 347). De manera que, una reacción de desconfianza por parte de una persona A hacia otra B–cuando la segunda presenta comportamientos abusivos que afectan a la primera–reconfigura la relación entre estas. Así, nuestras reacciones positivas o negativas calibrarían nuestra relación interpersonal. Aquí, Hiller argumenta que es esta tercera familia la que explica de mejor manera nuestras prácticas de atribución de responsabilidad.

Luis Estrada analiza el célebre experimento mental de la máquina de experiencias de Nozick en el noveno capítulo. Se suele considerar que el experimento motiva a refutar el hedonismo ético; postura que plantea que lo bueno–incluyendo la buena vida–se determina en función del tipo y cantidad de placer (en relación a la cantidad de dolor) que experimentamos. Nozick duda que los defensores de esta postura acepten permanecer conectados de por vida a una máquina que recrea artificialmente las mayores experiencias placenteras que imaginemos; donde, éstas no están relacionadas con la realidad. Si en efecto se da el caso en que los hedonistas rechazaran conectarse a la máquina, esto implicaría que en realidad los placeres no necesariamente están conectados con la buena vida, refutando así el hedonismo ético. Aquí, Estrada enriquece el análisis distinguiendo entre hedonismo sensorial y hedonismo actitudinal para luego argumentar que no es el primero él que sería refutado, sino más bien el segundo. En el hedonismo sensorial los placeres están relacionados con algún tipo de sensación, como por ejemplo, el placer sería la experiencia cualitativa a la que se accede al degustar algo que consideremos delicioso. Por otro lado, el hedonismo actitudinal, no asocia necesariamente el placer con una sensación sino con actitudes mentales (algunas actitudes mentales usuales en filosofía de la mente incluyen: creencias, miedos, expectativas, deseos, etc.). Por ejemplo, en cierto momento podríamos carecer de experiencias sensoriales placenteras, pero estar complacidos porque las personas que amamos están sanas. Es decir, el placer actitudinal vendría del hecho de que nuestras ex-

pectativas acerca de la salud de los que amamos son satisfechas. Estrada argumenta que "un hedonista ético que privilegie el acceso a experiencias placenteras y estime a estas como lo intrínsecamente valioso optaría por conectarse." (p. 371), y que son los hedonistas sensoriales aquellos que encontrarán intrínsecamente positivo a la experiencia sensorial placentera independientemente de si esta es recreada por la máquina o no. Mientras que, por el contrario, habrá hedonistas actitudinales que valorarán el placer actitudinal de que nuestras experiencias estén relacionadas auténticamente con la realidad. Así este tipo de hedonistas considerarán que una buena vida va más allá de una que maximice los placeres sensoriales. Por lo que, son los hedonistas actitudinales los que se negarían a conectarse, de lo que se concluiría que el experimento de Nozick refutaría este tipo de hedonismo. Estrada muestra cómo la actitud de los hedonistas actitudinales en la que la mejor forma de vivir trasciende el placer sensorial es una manifestación del perfeccionismo ético que Nozick apoyaba.

Juan Villacrés Bolaños
EDITOR

Referencias

Bradley, F.H. (1983). *Appearance and Reality: A Metaphysical Essay*, Oxford University Press.

Bradley, F. H. (1914). *Essays on truth and reality*. Cambridge University Press. (Trabajo originalmente publicado en 1910).

Bradley, F.H. (1935). Relations. En *Collected Essays*. Clarendon Press (pp. 629-676). (Trabajo originalmente escrito en 1924).

Candlish, S. (2009). *The Russell/Bradley dispute and its significance for twentieth century philosophy*. Springer.

Jonas, S. (2016). *Ineffability and its metaphysics:The unspeakable in art, religion, and philosophy*. Palgrave Macmillan.

Leiter, B. (s. f.). Analytic and Continental Philosophy. *The Philosophical Gourmet Report*. Recuperado 12 de noviembre de 2025 de https://philosophicalgourmet.com/analytic-and-continental-philosophy/

Lewis, D. (1983). *Philosophical PapersVolume I*. Oxford University Press.

Magee, B. (1971). *Modern British Philosophy*. London: Secker & Warburg.

Priest, G. (2014). *One: Being an Investigation into the Unity of Reality and of its Parts, including the Singular Object which is Nothingness*. Oxford University Press (UK).

Quinton, Anthony (1971). Absolute Idealism. *Proceedings of the British Academy* 57, 303–29

Russell, B. (1905). On denoting. *Mind*, 14(56), 479-493.

Russell, B. (1937). *The Principles of Mathematics* (2nd ed.). W. W. Norton & Company. (Trabajo original publicado en 1903).

Rusell, B. (2004). Seems, ¿Madam? Nay, it is. En *Why I am not a Christian and other essays on religion and related subjects* (pp. 48-57). Routledge. (Ensayo presentado en 1899).

Schaffer, J. (2010). Monism:The priority of the whole. *Philosophical review*, 119(1), 31-76.

PRIMERA SECCIÓN

FILOSOFÍA DEL LENGUAJE Y SU INFLUENCIA EN LA CORRIENTE

CAPÍTULO I

LÓGICA Y EMPIRISMO: ALGUNAS PRESUPOSICIONES E IMPLICACIONES DE LA TEORÍA DE LAS DESCRIPCIONES

Alejandro Tomasini Bassols[1]

RESUMEN

En este ensayo paso en revista críticamente tres "intuiciones" filosóficas de Bertrand Russell vinculadas con su célebre Teoría de las Descripciones y que pueden ser vistas como presuposiciones o como consecuencias de ella. Está, primero, la convicción de que la distinción entre significado y denotación (o entre sentido y referencia) no puede mantenerse sistemática e indefinidamente. Tiene

1 Alejandro Tomasini Bassols, Instituto de Investigaciones Filosóficas, UNAM y miembro del Sistema Nacional de Investigadores. Autor de múltiples ensayos y libros entre los cuales están *Los Atomismos Lógicos de Russell y Wittgenstein, Enredos Filosóficos y Filosofía Wittgensteiniana, Pecados Capitales y Filosofía, Tópicoas Wittgensteinianos,* ***Pena Capital*** *y otros ensayos, Filosofía de la Religión. Historia y debates, Wittgenstein: del* ***Tractatus*** *a las* ***Investigaciones*** *y Ensayos sobre el Wittgenstein Intermedio.*

que haber un momento en el que la denotación sea el significado de un término. Esta "intuición" automáticamente confronta a Russell con Frege. En segundo lugar, está la lectura que el propio Russell hace de su "principio de conocimiento directo" (*acquaintance*), que él presenta como una consecuencias de su teoría. Este principio compromete a Russell con un fenomenalismo radical y con un programa construccionista en teoría del conocimiento que culminará en un fracaso total. Y, por último, discuto brevemente la posición de Russell concerniente a las relaciones entre la lógica y el lenguaje. Russell entiende su teoría como fundada filosóficamente en un platonismo que el *Tractatus* vendría a destruir por completo.

Palabras clave: Teoría de las Descripciones, principio de conocimiento directo, Russell, Frege.

ABSTRACT

I this essay I critically review three Russellian philosophical "insights" which derive from his famous Theory of Descriptions, either as presuppositions or as consequences of it. There is first the conviction that the distinction between meaning and denotation (or between sense and reference) cannot systematically and indefinitely be maintained. There **must** be a moment in which the denotation is the meaning of a term. This "insight" automatically makes Russell oppose Frege. Secondly, I consider the reading that Russell himself makes of his so called 'principle of acquaintance', which according to him follows as a consequence from his theory. This principles commits Russell with a radical phenomenalism and with a kind of constructivist programme in the theory of knowledge which will culminate in a total failure. And, finally, I rapidly discuss Russell' stance concerning the relation between logic and language. Russell always understood his theory as being philosophically founded on a kind of Platonism that the *Tractatus* would totally destroy a few years later.

Keywords: Theory of Descriptions, principle of acquaintance, Russell, Frege.

1. Presentación

La célebre Teoría de las Descripciones constituye, sin duda alguna, la mayor contribución de Bertrand Russell a la filosofía. Es bien sabido que todas las grandes aportaciones de Russell, desde el logicismo hasta el realismo estructural, han sido cuestionadas de los más variados modos y aunque, como siempre en filosofía, ya sea por Russell ya sea por sus seguidores, se han ofrecido contra-argumentos y se han desarrollado líneas de argumentación en su defensa, de todos modos sigue siendo cierto que han sido puestas en entredicho. Esto incluye propuestas filosóficas tan diversas como la Teoría de los Tipos Lógicos y la teoría de las Construcciones Lógicas, una forma de hacer filosofía que Russell ejerció en los más variados contextos, tanto en metafísica (la construcción del mundo a partir de *sense-data* y universales) como en filosofía de la mente (todo su *Analysis of Mind* es un ejercicio construccionista en este otro contexto)[2]. En cambio, más de un siglo después de haber sido elaborada la Teoría de las Descripciones sigue siendo inatacable, por lo que tal vez podamos ya afirmar con suficiente confianza que de hecho es una teoría irrefutable.

La teoría en sí misma es relativamente simple, pero tiene tanto presuposiciones como implicaciones que no siempre han sido examinadas en concordancia con la importancia que revisten. Hay desde luego multitud de excelentes presentaciones y estudios de la teoría, pero podemos afirmar que quizá no haya un estudio que abarque o contemple absolutamente todos los aspectos de la misma. Obviamente, no me propongo siquiera intentar realizar un trabajo así en este ensayo. Mis objetivos son bastante más limitados y precisos. Esto de algún modo se explica por el hecho de que quisiera deslindarme de lo que parece ser una tendencia ya muy fuerte en relación con la teoría de Russell, a saber, que en la actualidad el enfoque correcto para exponerla parece ser el de volver a

2 Véase a este respecto el excelente artículo de M. Weitz. (1944). Analysis and the Unity of Russell's Philosophy. En *The Philosophy of Bertrand Russell*. The Library of Living Philosophers. Evanston and Chicago. Northwestern University.

presentar el mismo material pero intensificando la labor de formalización, efectuando paráfrasis cada vez más alambicadas de lo que Russell sostiene, concentrándose en las facetas más "técnicas" de la teoría, como las relacionadas con temas como los de la naturaleza de la existencia o el alcance de las descripciones. Discusiones así, sin embargo, en no pocas ocasiones dejan escapar aspectos de la teoría que filosóficamente son más importantes y sutiles, facetas de la teoría en las que lo que está en juego son más que tecnicismos intuiciones filosóficas fundamentales o últimas. En este trabajo quisiera precisamente ocuparme de algunas "intuiciones" russellianas que me parecen fundamentales y que vale la pena que se les revise críticamente. Me propongo, por consiguiente, examinar rápidamente tres de ellas, a saber:

a) la intuición anti-fregeana y anti-*Principios de las Matemáticas* concerniente al sentido y a la referencia (significado y denotación, en terminología de Russell) consistente en negar que dicha distinción pueda mantenerse indefinidamente. **Tiene** que haber un momento en el que el sentido y la referencia (significado y denotación) se funden uno en la otra.
b) La interpretación empirista de lo que Russell presenta como la principal implicación de su teoría, a saber, su principio del conocimiento directo (*acquaintance*) de los componentes de las proposiciones que comprendemos, y
c) la concepción russelliana (claramente 'pre-wittgensteiniana'), implícita en la lectura que Russell mismo hace de ella, de la relación entre el lenguaje y la lógica.

A grandes rasgos, lo que yo deseo defender es que:

a) Russell tiene razón en su crítica a la distinción fregeana y de *Los Principios de las Matemáticas* entre sentido y referencia
b) impone una lectura equivocada sobre la implicación epistemológica de su teoría
c) su visión de las relaciones entre la lógica y el lenguaje es totalmente errada. Examinaré los temas en el orden indicado.

Como es bien sabido, la Teoría de las Descripciones fue presentada por primera vez en 1905, en el famoso artículo "On Denoting"[3]. La teoría, más formalmente presentada, fue empleada en *Principia Mathematica* como parte integral de la solución russelliana al problema general de las paradojas. En esa monumental obra, escrita en conjunción con su ex-maestro, A. N. Whitehead, la teoría se aplica a las expresiones en las que aparecen símbolos de clase mostrando que, al igual que con las descripciones, desaparecen en el análisis, por lo que en el fondo son "símbolos incompletos":

> Los símbolos para clases, como aquellos para descripciones, son, en nuestro sistema, símbolos incompletos: sus usos están definidos, pero no se asume que ellos mismos signifiquen nada en absoluto. Es decir, los usos de dichos símbolos son definidos de tal manera que, cuando el **definiens** es sustituido por el **definiendum**, ya no queda ningún símbolo que pudiera suponerse que representa una clase. (Whitehead y Russell, 1973, pp. 71-72)

La Teoría de las Descripciones no es por sí sola la solución al problema de las paradojas, pero claramente contribuye a que el programa de reducción de números a clases avance. En todo caso, Russell volvió a presentar su teoría, enfatizando distintas consecuencias de la misma, primero en las conferencias sobre la filosofía del atomismo lógico y luego en el capítulo "Descripciones" de su libro *Introducción a la Filosofía Matemática*. Por último, Russell hace una veloz alusión a la teoría en su *Historia de la Filosofía Occidental*. En este trabajo me concentraré exclusivamente en algunas de las afirmaciones que Russell hace en "Sobre el Denotar".

2. Sentido y referencia

En general se asume que hay por lo menos dos grandes modelos explicativos rivales en filosofía del lenguaje, a saber, la teoría de

3 Ver B. Russell . (1972). On Denoting. En *Logic and Knowledge*.Allen and Unwin. Hay varias traducciones al español.

G. Frege del sentido y la referencia y la Teoría de las Descripciones. Ambas teorías tienen tanto críticos acerbos como apasionados partidarios. Las teorías son a menudo equiparadas desde diversos puntos de vista: capacidad y superioridad explicativa, simplicidad teórica, posibilidades de expansión, consecuencias negativas o inaceptables, etc. Yo pienso que la Teoría de las Descripciones proporciona en general mejores explicaciones que la teoría de Frege, pero dado que se trata de teorías filosóficas es poco probable que sobre la base de una mera confrontación entre ellas se logre inclinar la balanza de manera contundente en favor de una teoría sobre la otra. Sin embargo, "On Denoting" contiene un argumento que, deseo sostener, constituye, cuando es debidamente explicado, un ataque decisivo y definitivo en contra de un cierto "punto de vista", que es relevante y que se tiene que especificar. Este "punto de vista" que Russell somete a crítica, deseo sostener, en realidad encarna no en una sino en dos teorías filosóficas, una de las cuales es la de Frege. Naturalmente, para entender y apreciar lo que muy probablemente sea uno de los argumentos más elusivos y escurridizo jamás escrito y del cual aquí presentaremos tan sólo la intuición fundamental que lo anima, tenemos que hacer previamente unos cuantos recordatorios. Esto es importante porque si bien la Teoría de las Descripciones es sumamente original, habría que admitir también que tiene antecesores.

Frege presenta su teoría del sentido y la referencia para resolver un problema concreto, a saber, la naturaleza de la igualdad (¿o de la identidad?)[4]. Dicho en dos palabras, el problema consiste,

4 La duda emerge con la primera palabra ('*Gleichheit*') del celebérrimo ensayo de Frege, *Über Sinn und Bedeutung*. Comenta Frege en la nota de la primera página: *Ich brauche dies Wort im Sinne von Identität und verstehe "a = b" in dem Sinne von "a" ist dasselbe wie "b" oder "a und b fallen zusammen".* O sea, en primer lugar, Frege mismo identifica igualdad con identidad, respecto a lo cual lo menos que podemos decir es que es problemático. En la versión inglesa de M. Black y P. Geach, la cita es vertida de la siguiente manera: "Uso esta palabra estrictamente y entiendo que 'a = b' tiene el mismo sentido de 'a es lo mismo que b' o 'a y b' coinciden'" (*sic*). En las versiones al español, esto es, la de Tomás M. Simpson (*Semántica Filosófica: problemas y discusiones* (Argentina: Siglo XXI, 1973)) y la de U. Moulines (G. Frege. Estudios sobre Semántica (Barcelona: Ariel, 1973)), la nota en cuestión es traducida de exactamente la misma manera salvo por un detalle, que

como Frege lo hace ver, en que cuando emitimos un enunciado de la forma '*a* = *b*', lo que mediante el signo '=' queremos decir no puede ser ni una relación entre objetos ni una relación entre nombres de objetos. Si decimos que Napoleón es Bonaparte (usando 'es' para expresar el significado de '='), no podemos estar hablando de objetos puesto que, si efectivamente Napoleón es Bonaparte, entonces no podría haber ninguna diferencia entre 'Napoleón es Bonaparte' y 'Napoleón es Napoleón'. Es obvio que esta respuesta no funciona (un enunciado es analítico, el otro sintético, uno es *a priori*, el otro *a posteriori*, etc.). Pero la otra opción, a saber, que '=' indica una relación entre nombres de objetos es igualmente inaceptable, por la sencilla razón de que entonces todos los enunciados de identidad verdaderos se volverían estipulaciones lingüísticas: se tendría que decir que **en español** o **en francés** o … etc., la palabra 'Napoleón' designa lo mismo o tiene el mismo significado que la palabra 'Bonaparte'. Se supone, sin embargo, que con 'Napoleón es Bonaparte' lo que se hace es una afirmación de orden **histórico**, no una de carácter lingüístico. Pero ahora sí estamos en un problema serio, porque si la igualdad (identidad) no es ni una relación entre objetos ni una relación entre nombres de objetos, entonces ¿qué es? Es para responder a este interrogante que Frege ofrece una explicación completamente diferente en tér-

---

es el siguiente: "Empleo esta palabra en el sentido de identidad y entiendo 'a = b' **con** (Simpson) **en** (Moulines) el sentido de 'a es lo mismo que b' o 'a y b coinciden'". A mi modo de ver, las traducciones citadas son todas, por una u otra razón, fallidas. La versión al inglés está en *Translations of the Philosophical Writings of Gottlob Frege* (Oxford: Basil Blackwell, 1952) y, aparte de un evidente error (probablemente tipográfico) en el uso de las comillas, en ella tranquilamente se suprime la aclaración de Frege concerniente a la identidad. Las versiones al español, por su parte, le son infieles a Frege en el uso de las comillas. Según mi leal saber y entender, lo que Frege afirma es lo siguiente: *Uso esta palabra en el sentido de identidad y entiendo 'a = b' en el sentido de* **'a' es lo mismo que 'b'** (énfasis mío. ATB) *o 'a y b coinciden'.* Obviamente, "'a" es lo mismo que "b'" no es lo mismo que 'a es lo mismo que b'. En el primer caso se alude a nombres, en el segundo a sus referencias. Aprovecho para aclarar el uso de las comillas simples y dobles: las comillas simples sirven para **mencionar palabras**, en tanto que las comillas dobles se usan para **nombrar conceptos**. Podemos entonces decir, *e.g.,* que el significado de 'árbol' es el concepto "árbol". Ocasionalmente uso las comillas dobles también para citar. El contexto claramente indica qué uso de las comillas se está haciendo.

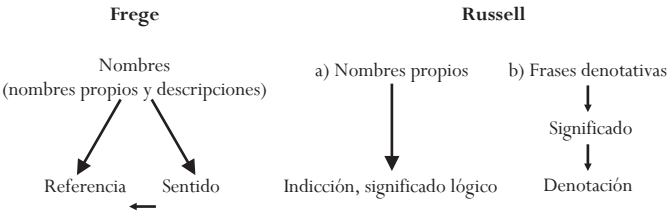
minos de dos nociones técnicas, a saber, "sentido" y "referencia". Así, decir que Napoleón es Bonaparte se explica diciendo que las palabras 'Napoleón' y 'Bonaparte' tienen diferentes sentidos pero la misma referencia. La aplicación de estas nociones se va extendiendo, pero no es parte de nuestra tarea ocuparnos de ello aquí. Por el momento, con lo que hemos dicho nos basta.

Como todos sabemos, el primer filósofo de habla inglesa que se ocupó del pensamiento de Frege, quien por primera vez lo dio a conocer en el mundo filosófico anglo-sajón y quien sin duda lo apreció en todo lo que valía fue Bertrand Russell. Éste completó su libro *Los Principios de las Matemáticas* con dos "Apéndices", uno de ellos dedicado precisamente a las teorías filosóficas y matemáticas de Frege. Dejando de lado toda clase de anacronismos, es evidente que se trata de una presentación soberbia del *corpus* lógico y filosófico de Frege de aquellos tiempos. Que Russell estaba en posición de apreciar como nadie el pensamiento del gran lógico alemán lo deja en claro el hecho de que él mismo había desarrollado por cuenta propia toda una serie de posiciones que eran si no exactamente las mismas sí **sumamente parecidas** y afines a las de Frege. En realidad, había entre ellos grandes coincidencias, inclusive si no eran totales, como pasa por ejemplo con sus respectivas versiones del logicismo: si bien fue Frege quien lo inició, Russell de manera totalmente independiente lo llevó hasta sus últimas consecuencias. Aquí casi podría establecerse un parangón interesante entre el pensamiento de Frege y el de Russell, por un lado, y el de Russell y el *Tractatus*, por el otro, si no fuera por el hecho de Russell nunca fue alumno de Frege. La idea es la siguiente: así como el autor del *Tractatus Logico-Philosophicus* se mueve en un universo russelliano, del cual da cuenta mejor que el propio Russell, así también Russell representa un paso hacia adelante en la dirección en la que Frege se movía. Así, aunque sin duda hay diferencias fundamentales entre la filosofía de Russell y la del *Tractatus*, de todos modos Russell y Wittgenstein son los dos grandes atomistas lógicos que hay; y de igual modo, aunque obviamente difieren en multitud de puntos Frege y Russell están unidos por algo más que

un mero enfoque o una cierta perspectiva. Yo diría que comparten una forma de pensar. Un modo de confirmar esto es precisamente examinando lo que Russell sostiene en *Los Principios de las Matemáticas* en relación con el nombrar y el denotar.

El punto de partida de Russell es la idea de que, aunque inexacta, la gramática escolar establece lo que son las grandes categorías filosóficas. Tenemos entonces nombres propios, "adjetivos" y verbos. De acuerdo con Russell, y siguiendo en esto a J. S. Mill, los nombres propios no tienen ningún significado, llamémosle así, 'lingüístico'. Empero, sí tienen lo que podríamos llamar 'significado lógico'. De ahí que su contribución al sentido de una oración se reduzca a indicar quién o qué es el sujeto de la proposición. Aunque lo que voy a decir no es lo que Russell sostiene, podemos presentar la idea como sigue: los nombres propios ('Napoleón', 'Cleopatra', etc.) tienen un significado lógico y ese significado es el objeto nombrado. Ahora bien, hay expresiones que **parecen** nombrar a un término, pero que en realidad no lo nombran porque no mantienen con él la misma clase de relación que mantiene un nombre con su significado lógico. Así, todas las expresiones en las que aparecen palabras lógicas como 'todos', 'algún', 'la', etc., ciertamente contribuyen al sentido de las oraciones de las que forman parte, pero no porque indiquen un objeto sino porque lo "denotan". Así, si yo digo que me topé con alguien que es simpático no quiero decir que me topé con una entelequia errante que es "alguien". Más bien, el término 'alguien' sirve para indicar que estoy hablando de una persona en concreto, sólo que la designo o me refiero a ella ambigua o vagamente. Supongamos que ese alguien es Juan. Entonces cuando hablo de alguien hablo no de un señor indefinido llamado 'alguien' sino de una persona concreta llamada 'Juan', que es su denotación. Intuitivamente entonces, Russell distingue lógicamente entre signos simples, como los nombres propios, y expresiones complejas, como las frases denotativas. Naturalmente, las funciones semánticas de esas dos clases de expresiones son diferentes: los signos simples **nombran** en tanto que las expresiones compuestas **denotan**. En aras de la claridad, podemos repre-

sentar gráficamente las posiciones de Frege y Russell en relación con los nombres y las frases denotativas. Tendríamos entonces el siguiente cuadro:



Cuadro 1
Nota: Elaborado por el autor

Vale la pena observar que, de hecho, Frege mantiene una doble teoría por cuanto para él lo que tiene tanto sentido como referencia son los **signos** (en este caso, los nombres), pero de acuerdo con él lo que nos lleva a la referencia es el sentido de la expresión. O sea, el signo tanto su referencia como refiere a través de su sentido. En otras palabras, tanto el signo como su sentido refieren (o, en terminología russelliana, denotan). Infiero que Frege tiene una doble teoría o una teoría ambigua del referir lo cual, en mi opinión, es una deficiencia de su teoría. Vale la pena señalar, asimismo, que Frege no distingue entre signos simples y signos complejos (frases denotativas). Pero ¿por qué? La explicación es relativamente simple y fácil de encontrar: si bien es a Frege a quien se le debe el haber articulado por primera vez la teoría de la cuantificación, lo cierto es que su simbolismo no le permite trazar o reconocer ciertas distinciones. Russell, que tiene una simbología distinta de la del *Begriffsschrift* fregeano, de inmediato se percata de que hay una diferencia y ciertamente la toma en cuenta. Esta diferencia simbólica, como veremos, tiene consecuencias que van más allá del mero simbolismo.

En resumen y concentrándonos exclusivamente en el caso las descripciones (frases denotativas), lo que podemos afirmar es que aunque no idénticas, las teorías de Frege y Russell sí son "simila-

res". En el caso de las descripciones, tanto uno como el otro nos hacen pasar por el sentido (o el significado) para llegar a la referencia. La diferencia grande entre ellos tiene que ver ante todo con los signos simples, *i.e.*, con los "nombres" (nombres propios y demostrativos), ya que Frege sólo tiene una misma explicación semántica para toda clase de signos, en tanto que Russell ofrece dos.

Ahora sí podemos retomar lo que Russell sostiene en "On Denoting", en el famoso argumento de la elegía de Gray. El argumento es realmente difícil de seguir, pero el objetivo que Russell persigue es bastante claro: lo que él quiere es **demostrar** que no se puede mantener sistemáticamente la distinción entre sentido y referencia (Frege) o, alternativamente, entre frase denotativa y denotación (Russell). Lo que él quiere establecer de una vez por todas es que desde un punto de vista lógico el significado de una expresión tiene que ser lo mismo que la denotación y si no se reconoce esta identificación, entonces surgen problemas insolubles. Su posición es que **lógicamente** no puede haber intermediarios entre los signos y sus denotaciones. De acuerdo con Russell, signos y objetos se tienen que conectar directamente, en tanto que en la teoría de Frege y en la de *Los Principios de las Matemáticas* se conectan via el sentido (significado). Así, pues, si el argumento de Russell en "On Denoting" es válido (como creo que lo es), Russell refuta dos teorías al mismo tiempo[5].

Como lo advertí, no voy a intentar reconstruir aquí el argumento de Russell, que es un tema en sí mismo, además de que ya expuse en otro lugar mi punto de vista y ofrecí mi reconstrucción de la argumentación russelliana[6]. Me limito entonces a hacer la siguiente observación: es innegable que en uno de los párrafos de la argumentación Russell incurre en un error menor de uso de comillas. No hay forma de salvarlo del error, pero lo que de inmediato

5 Sobre el argumento de las páginas 48-51 de "On Denoting", véase la antología: Tomasini, A. (Ed.) (1996).*Significado y Denotación. La polémica Russell-Frege*. Interlínea. Traducciones y notas de Alejandro Tomasini Bassols.

6 Véase Tomasini, A (Ed.) (1996). La Crítica de Russell a Frege: una exégesis. En *Significado y Denotación. La polémica Russelñ-Frege*. Interlínea.

hay que señalar es que el argumento no depende del uso de las comillas. Lo que hay detrás es la fundamental intuición de que la distinción <sentido/referencia o <frase denotativa/denotación> no se sostiene indefinidamente. En el argumento de Russell esto se pone de manifiesto cuando se intenta **nombrar** el **sentido** de una **frase denotativa** que tiene como **denotación** el **sentido** de **otra frase denotativa**, con lo cual se genera un regreso al infinito de tipo vicioso. Yo estoy convencido de que Russell tiene razón y de que su intuición es no sólo brillante, sino de consecuencias nada desdeñables.

Con la Teoría de las Descripciones se explica muy fácilmente el *modus operandi* semántico de las descripciones, definidas u otras. La explicación aparece con lo que Russell presenta como el "principio de la teoría", que reza como sigue: *las frases denotativas no tienen significado consideradas aisladamente (o sea, no tienen lo que llamé 'significado lógico'), pero toda proposición en cuya formulación verbal aparezcan son significativas*. En otras palabras, dado que las descripciones no son signos simples, es decir, no son nombres, entonces lógicamente carecen de significado, a diferencia de lo que pasa con los signos lógicamente simples. Dado que las descripciones están formuladas en palabras del lenguaje natural, entonces es normal que formen parte de oraciones bien formadas y que contribuyan a sus respectivos sentidos, pero no para apuntar a objetos, inclusive si aparecen como sujetos gramaticales. Sabemos entonces que si alguien afirma que el presidente de Alemania es calvo, lo que está diciendo es **no** que hay alguien que es el presidente de Alemania y que además es calvo, sino que **hay alguien o algo** (eso está todavía por especificarse) que tiene la propiedad de ser presidente de Alemania y además la de ser calvo. Esto podría parecer una distinción baladí. Trataré de hacer ver que ello dista mucho de ser así.

3. Lenguaje y lógica

Si quisiéramos explicar en términos no filosóficos lo que es la Teoría de las Descripciones podríamos, para ciertos efectos, presentarla simplemente como una máquina de traducción de oraciones del

lenguaje natural (en especial, de las que contienen descripciones, definidas u otras) al lenguaje canónico de la lógica clásica, (extensional, de primer orden). Esa faena de traducción, sin embargo, no es un mero juego simbólico y muy rápidamente se puede hacer ver que tiene consecuencias filosóficas de suma importancia. Enumeremos las más importantes.

Para empezar, la Teoría de las Descripciones se nos revela como una **teoría de la forma lógica**. En relación con ésta, el punto importante es que automáticamente hace ver que la forma gramatical de las oraciones del lenguaje natural es lógicamente incorrecta, puesto que no coincide con ella. Pero hay más. La teoría de Russell (quizá deberíamos simplemente decir: el lenguaje canónico de la lógica) distingue dos clases de signos: signos simples (constantes no lógicas) y signos complejos (descripciones, de la clase que sean), que son los signos que desaparecen en la traducción (o "análisis"). Podríamos presentar esto último como sigue: dado que el análisis lógico (de traducción) no puede extenderse *ad infinitum*, éste tiene entonces que detenerse en algún momento y por lo tanto **tiene** que haber signos simples, esto es, signos cuyos significados sean sus denotaciones. Aquí se plantea un problema que podríamos plantear como sigue: al pasar de la forma lógica a la forma gramatical nos enteramos de que **tiene** que haber signos simples, pero cuando regresamos de la forma lógica a la forma gramatical no encontramos, o no fácilmente, los signos simples que la lógica nos asegura que tiene que haber. En efecto: ¿dónde en el lenguaje natural están esos signos simples de los que nos habla la teoría de Russell? Lógicamente, las descripciones, como es obvio, son signos complejos por lo que, dejando de lado los signos sincategoremáticos, los únicos términos que podrían pasar el test de simplicidad formal serían los nombres propios, como 'Napoleón' o 'Hércules'. Desafortunadamente, el asunto no es tan simple, pues de inmediato se plantea un problema serio: si bien **sintácticamente** los nombres propios **son** simples, **semánticamente y epistemológicamente** se conducen como las descripciones. La teoría explica por qué es así: los nombres propios permiten que se predique

la existencia (o la no existencia) de sus referencias (significados), lo cual es característico de las descripciones y algo proscrito en el caso de los signos simples. Ciertamente oraciones como 'Sócrates existió' o 'Sócrates no existió' son perfectamente significativas. Por otra parte, se les puede usar sin que para ello se tenga que mantener algún contacto cognoscitivo directo con dichas referencias. Por ejemplo, Napoleón no está entre nosotros pero podemos hablar de él. Además, los nombres propios permiten que se formulen enunciados de identidad que son informativos, como cuando alguien dice que Napoleón es el vencedor de Marengo, lo cual no puede pasar con los signos cuyos significados son los objetos nombrados. Queda claro, por lo tanto, que los nombres propios del lenguaje natural no pueden constituir la clase de signos simples que la teoría de Russell parece estar demandando. Pero entonces, una vez más: ¿de qué signos hablamos cuando hablamos de signos lógicamente simples y cuáles podrían ser sus referencias?

Es aquí que se hace sentir una presuposición filosófica primordial de Russell que, pienso, es no obstante falsa. La presuposición russelliana consiste en asumir que están por un lado el lenguaje y la realidad y, por el otro el universo de la lógica y que, por alguna afortunada razón o causa, entran en contacto. Nosotros entonces, por alguna clase de inspiración, reconocemos las verdades de la lógica. Esta visión de la lógica hace que Russell incurra en una falacia: partiendo del lenguaje natural se las arregla para generar una cierta formalización del mismo y luego asume que es dicha formalización lo que regula o reglamenta el comportamiento del lenguaje. Esta conclusión a su vez le imprime al pensamiento de Russell una determinada orientación. Para empezar, se aclara lo que Russell presenta como consecuencia fundamental de su teoría, a saber, el "principio de conocimiento directo" (*knowledge by acquaintance*), un principio por medio del cual Russell vincula tres nociones importantes, a saber, las de significado, comprensión y denotación. Él lo presenta de un modo que vale la pena consignar *in extenso*. Afirma:

> Un resultado interesante de la teoría del denotar recién expuesta es este: cuando hay algo con lo que no tenemos un contacto directo inmediato sino sólo definición por medio de frases denotativas, entonces las proposiciones en las que esa cosa es introducida por medio de una frase denotativa en realidad no contienen a esa cosa como un componente, sino que contienen más bien los componentes expresados por las diversas palabras de la frase denotativa. Así, en toda proposición que podamos aprehender (i.e., no únicamente en aquellas de cuya verdad o falsedad podemos juzgar, sino en todas acerca de las cuales podemos pensar), todos los componentes son realmente entidades de los que tenemos un conocimiento directo inmediato. (1972, pp. 55-56)

La enunciación es clara, pero el sentido requiere ser parafraseado. De acuerdo con él, si "definimos" un objeto mediante descripciones, dado que éstas desaparecen en el análisis (en la traducción al lenguaje de la teoría), tenemos entonces que inferir que no hay tal objeto, es decir, que **no existe** dicho objeto. *Prima facie*, esto es contradictorio, pero en el fondo no lo es. Lo que Russell está haciendo es efectuar lo que yo llamaría una 'labor de limpieza ontológica'. Supongamos que afirmo que el rector de la UNAM es mexicano. Lo que su teoría aclara es que no hay ningún objeto que sea "el rector de la UNAM". No existe tal objeto. Lo que en cambio sí hay es "algo" (que asumimos que es una persona) que es único en tener la propiedad de ser rector de la UNAM y en ser mexicano. O sea, con la transición del lenguaje natural al lenguaje de la lógica se produce una especie de transmutación de objetos en propiedades; lo que el lenguaje natural presenta como objeto, el lenguaje de la lógica lo presenta como propiedades. Supongamos que no objetamos a este importante punto. De todos modos querremos saber cómo o por qué comprendemos una proposición que contiene descripciones cuando no existe el objeto que sea "el tal y tal". La respuesta de Russell es: sabemos que **hay un algo** (todavía no especificado) que tiene ciertas propiedades y nosotros conocemos directamente dichas propiedades. Siguiendo con el ejemplo, ni yo ni nadie conocemos al objeto "el rector de la UNAM", pero en cambio sí conocemos directamente las propiedades de un "algo" (se

supone que es una persona) que son las de ser rector de la UNAM y ser mexicano. Si hubiera problemas con esa respuesta porque se vuelven a introducir nombres propios ('UNAM', por ejemplo), el análisis lógico tendría que volverse a aplicar y entonces se vería que conocemos directamente las propiedades de una institución mexicana que es única en ser autónoma y nacional. Es porque conocemos estas propiedades independientemente de la oración en cuestión que entendemos la oración original. En principio, en todos los casos de proposiciones que comprendemos deberíamos poder llegar a los componentes últimos que son entonces nombrados por signos simples.

Es importante percatarse de que lo que Russell presenta como una consecuencia de la Teoría de las Descripciones es en realidad un principio empirista radical. Lo que en última instancia él defiende es la idea de que el conocimiento humano se funda en última instancia en la experiencia, es decir, en el "conocimiento directo" de "objetos" que podemos nombrar. Se entiende ahora por qué Russell considera que si queremos encontrar las referencias de signos simples que se ajusten a las reglas de formación de su teoría, *i.e.*, de la teoría de la cuantificación, se tiene forzosamente que efectuar un "análisis vertical", es decir, tenemos que abandonar el universo de los objetos materiales y encontrar "objetos" de otra clase que puedan ser conocidos **directamente**, es decir, **no descritos**. Dado que los nombres propios del lenguaje natural quedaron descartados como candidatos a signos lógicamente simples, Russell tiene que encontrar términos que sean tales que puedan sustituir a las variables de las expresiones de su teoría, a los signos simples que su teoría afirma que tiene que haber. ¿Qué palabras encuentra Russell que satisfagan las condiciones que él impone? En su primera respuesta él propuso cuatro, a saber, 'esto', 'yo', 'aquí' y 'ahora'. Lo denotado por esos términos es lo que conocemos directamente. Al final sólo quedaría 'esto' como paradigma de eso que Russell denominó 'nombres propios en sentido lógico'. En el uso correcto de 'esto', el sujeto emite la palabra, apunta a un obje-

to y entra en un contacto cognoscitivo con él. Es así como lenguaje, realidad y mente convergen en un punto.

Ahora bien, aquí es preciso hacer una aclaración. Tenemos que distinguir entre la Teoría de las Descripciones propiamente hablando y la **interpretación** que Russell hace de ella. Lógicamente, podemos aceptar la primera y rechazar la segunda, que es en mi opinión la opción correcta. La interpretación concierne básicamente a los signos simples (las constantes) y a las relaciones entre el lenguaje y la lógica. Por lo pronto, obsérvese que los significados de los signos simples son a la vez los objetos de la mente y de la realidad, pero no son objetos materiales. Ahora bien, que la interpretación de Russell es errada es algo que su mismo trabajo filosófico dejó en claro. Al convencerse de que la única forma de dar cuenta de los significados de sus términos primitivos era a través de un "análisis vertical", Russell se vio comprometido con un fenomenalismo radical y, como consecuencia de ello, con un descabellado programa de reconstrucción del mundo y de las otras mentes. El proyecto era simplemente descomunal, pues se trataba de reconstruir el mundo externo a partir de "particulares", es decir, los datos de la conciencia, *i.e., sense-data* (percepción, memoria e introspección) y universales, esto es, predicados y relaciones (intelección). Viéndolo retrospectivamente, puede afirmarse que era un programa filosófico fallido *a priori*. Percibiendo el peligro, Wittgenstein puso sobre aviso a Russell sólo que éste no contaba con los elementos necesarios para proponer una lectura alternativa de su propia teoría. Ya desde 1913, Wittgenstein pone en guardia a Russell respecto a su forma de usar su teoría: "Sólo quiero añadir que tu 'Teoría de las Descripciones' es ciertamente del todo correcta, inclusive si los signos individuales primitivos en ella no son los que tú pensaste" (1974, pp. 43-44).

¿En dónde y por qué se separa Wittgenstein de Russell en relación con el tema que nos ocupa? Pienso que la diferencia fundamental entre Russell y Wittgenstein en este punto radica en que Russell asume que el lenguaje de la lógica es un lenguaje meramente formal, enteramente abstracto pero descriptivo. Él, por lo

tanto, asume que la lógica conforma un universo auto-contenido, en contacto con el mundo real pero independiente por completo de él. O sea, Russell (igual que Frege, aunque haya diferencias entre sus respectivas posiciones) es un platonista convencido. Por lo tanto, para él la estructura lógica de las proposiciones y los pensamientos es independiente de la estructura gramatical de las oraciones que los expresan. Para Wittgenstein en cambio eso es mitología filosófica pura. La lógica para éste es, como él mismo la llama, el "gran espejo" del lenguaje y de la realidad (1974, p.59). El lenguaje ciertamente está regido por la las reglas de la sintaxis lógica, pero lo que hay que entender es que dichas reglas no existen al margen del lenguaje. Y dado que es a través del lenguaje como atrapamos la realidad, no podríamos tener una idea coherente de mundo si éste por su parte no se comportara "lógicamente". Es porque los objetos tienen en los nombres a sus representantes que los hechos de los que forman parte nos resultan comprensibles. Frege, por ejemplo, compartía la idea russelliana de la independencia ontológica de la lógica. Por lo tanto, también para él la conexión entre la lógica, el lenguaje y la realidad, era a final de cuentas contingente. Desde su perspectiva, la lógica es, existe o subsiste haya un mundo o no lo haya y haya o no un lenguaje para describirlo. Eso es lo que para Wittgenstein era absurdo. En el *Tractatus*, Wittgenstein se rebela en contra del platonismo de Frege y Russell con un argumento contundente que automáticamente hace sentir por qué ese punto de vista es equivocado. Dice Wittgenstein: "Y si no fuera así ¿cómo podríamos aplicar la lógica? Podría decirse: si hubiera una lógica, aunque no hubiera también un mundo ¿cómo podría haber una lógica dado que hay un mundo?" (1974, p.66)

Claramente, la idea de Wittgenstein es que sobre la base exclusivamente de la experiencia cruda no se podría percibir la diferencia entre relaciones contingentes, como la de ser tío de alguien, y relaciones necesarias, como la de implicación de una proposición y por lo tanto no podría reconocerse el *status* que la lógica tiene. Frege se contentaba con una noción puramente formal de objeto y de nombre, sólo que por ello su concepción tenía inevitablemente

consecuencias desastrosas. Así, por ejemplo, él tiene que aceptar que una expresión como 'Julio César es un número primo' es perfectamente legítima, puesto que 'Julio César' es un nombre y 'es un número primo' una función; el primero es saturado, es decir, por no contener huecos de ninguna índole no necesita ser completado, que es precisamente lo que pasa con la segunda, por lo cual encajan perfectamente bien y pueden así dar lugar a un pensamiento. En este caso, sin embargo, hay un problema, a saber, que el resultado es absurdo, pero es un resultado que Frege tiene que aceptar y que su teoría no puede bloquear. Wittgenstein, en cambio, no se ve comprometido con resultados así. Él sí puede argumentar que 'ser número primo' no es una propiedad posible de una persona, puesto que los objetos se mueven en lo que podríamos llamar **su** espacio lógico y los espacios lógicos de Julio César y de los número primos no se sobreponen, no se tocan, no coinciden. Eso precisamente es lo que el lenguaje representa o **retrata**. El *Tractatus* no tiene ningún problema con el rechazo de oraciones tan absurdas como la recién mencionada.

Podemos ahora regresar a la Teoría de las Descripciones, pero viéndola esta vez desde la perspectiva del *Tractatus*. Si la lógica es siempre la lógica **del** lenguaje, entonces la formalización no tiene por qué llevarnos a ninguna clase de "análisis vertical". Aquí lo interesante es lo que la teoría pone de relieve, a saber, la **ontología implícita** en nuestro discurso. Así, por ejemplo, si decimos que el rector de la UNAM es mexicano, aceptamos con Russell que no hay ningún objeto que sea "el rector de la UNAM". En su formulación tradicional, lo que tendríamos sería:

$$(\exists x)\,(((Rx \,\&\, (\forall y)\,(Ry \supset x = y)) \,\&\, Mx)$$

Russell, sin embargo, erróneamente considera que el universo recorrido por la variable tiene que ser un universo de *sense-data*, puesto que él cree que su teoría fuerza a un análisis vertical, dado que los nombres propios son descripciones encubiertas y que independientemente de ello tenemos que tener signos simples cuyas

denotaciones sean, por una parte, objetos de una naturaleza diferente de la de los objetos materiales y, por la otra, objetos que conocemos directamente en la experiencia. Pero eso es obviamente una inferencia inválida. La notación lógica simplemente **refleja** lo que se dice en el lenguaje. Ahora bien, si hablamos de rectores, de la UNAM, de ser mexicano: ¿de qué podemos estar hablando? Hasta donde logro ver, no hay más que una respuesta: de personas, de seres humanos. ¿Cómo nos explicamos entonces la oración? Simplemente **asumimos** que estamos hablando de personas, a las que describimos adscribiéndoles propiedades y relaciones. ¿Cómo es posible que el lenguaje simbólico nos hipnotice y nos haga hacer inferencias absurdas? Lo que pasa es que en el discurso normal no nos sentimos obligados a hacer aclaraciones que lógicamente están trazadas, aunque no las hagamos explícitas. Simplemente, damos por sentado, asumimos que los participantes en la conversación durante la cual se usa una expresión como 'el rector de la UNAM' ya saben de qué clase de cosas se está hablando o, para ser más precisos, de qué clase de entidades se está hablando. Y eso sucede sistemáticamente en todos los casos.

Si, conversando con un amigo, afirmo que el león dio un salto impresionante para atrapar a su presa, mi interlocutor y yo "sabemos" que estamos hablando de animales; si alguien afirma que los Boeing son seguros, todos sabemos que se está hablando de aviones, y así en todos los casos. Lo que la Teoría de las Descripciones hace es precisamente reflejar eso que todo el tiempo hacemos los hablantes normales, esto es, nos hace ver que en general no especificamos explícitamente de qué hablamos (cuál es "nuestra ontología") porque lo damos por supuesto, pero en ningún momento sugiere que tengamos que dar un brinco ontológico y epistemológico y pasar a buscar entidades de una naturaleza diferente de la de los objetos del sentido común. ¿Por qué surge un problema? Porque se estudia el lenguaje formal de manera completamente aislada, con total independencia del lenguaje usado. Ello inevitablemente induce a errores filosóficos. En lo que a nosotros concierne, la moraleja es clara: **lógica sin contextualización es filosófi-**

camente equívoca. Este resultado es, yo creo, importante y en total concordancia con la perspectiva tractariana.

4. Conclusiones

Cuando en 1905 Russell publica su famoso artículo todavía no ha sufrido el brutal impacto que le causará el choque con Wittgenstein 6 años después. Para entonces Russell estaba todavía muy lejos de ver en las verdades de la lógica meras tautologías. Si hay algo que literalmente se vio forzado a admitir por la incontenible argumentación wittgensteiniana fue la idea de que la lógica era radicalmente diferente de cualquier otra ciencia. Como siempre en el *Tractatus*, Wittgenstein es de una claridad pasmosa: "La explicación correcta de las proposiciones de la lógica tiene que conferirles un lugar especial entre todas las proposiciones"(1974, p.72).

O sea, no se puede pasar de manera gradual de la percepción sensorial a las ciencias empíricas, de éstas a la mecánica racional y de ésta a las matemáticas y a la lógica. No hay tal *corpus* de conocimientos humanos. Al separar tajantemente a la lógica del resto de las ciencias, Wittgenstein acabó para siempre con ese grandioso proyecto filosófico. Contrariamente a lo que Russell creía, no hay, estrictamente hablando, algo así como "conocimiento lógico", en el sentido en que podemos hablar de conocimiento de las estrellas o de los cromosomas. Para Russell la posición del *Tractatus* fue un severísimo golpe, puesto que acabó con su gran ilusión filosófica que era combinar la tradición empirista británica con una investigación enteramente *a priori*, que él pensaba que era la investigación en lógica y en la cual él era un experto. Para él, la lógica era una ciencia sólo que más abstracta que las ciencias naturales, una ciencia de máxima generalidad, una disciplina que establecía verdades en el mismo sentido en que lo hace la biología pero que lidiaba con "entidades" como las formas lógicas así como la física trabaja con átomos, campos magnéticos, energía y demás. Es cierto que no es fácil determinar si los objetivos de Russell eran siquiera conciliables, puesto que su empirismo era un empirismo con entidades abstractas, algo que a Hume o a Berkeley les habría parecido

una aberración. Pero independientemente de ello lo que es claro es que, si nos familiarizamos con lo que eran las ambiciones filosóficas del Russell pre-wittgensteiniano, podremos entender y apreciar mejor su gran Teoría de las Descripciones, al tiempo que la resguardamos de las incomprensiones de su propio forjador.

Referencias

Black, M., & Geach, P. (Trans.). (1952). *Translations of the philosophical writings of Gottlob Frege*. Basil Blackwell.

Frege, G. (1973). *Estudios sobre semántica*. (U. Moulines, Trans.). Ariel.

Russell, B. (1972). On denoting. En *Logic and knowledge*. Allen and Unwin.

Simpson, T. M. (1973). *Semántica filosófica: Problemas y discusiones*. Siglo XXI.

Tomasini Bassols, A. (1996). La crítica de Russell a Frege: Una exégesis. En A. Tomasini Bassols (Ed.). *Significado y denotación: La polémica Russell-Frege*. Interlínea.

Tomasini Bassols, A. (Ed.). (1996). *Significado y denotación: La polémica Russell-Frege*. Interlínea.

Weitz, M. (1944). Analysis and the unity of Russell's philosophy. En *The philosophy of Bertrand Russell*. The Library of Living Philosophers. Evanston and Chicago: Northwestern University.

Whitehead, A. N., & Russell, B. (1973). Incomplete symbols, Chapter III. *En Principia mathematica* to *56 (pp. 71-72). Cambridge University Press.

Wittgenstein, L. (1974). *Letters to Russell, Keynes and Moore*. Ithaca: Cornell University Press.

Wittgenstein, L. (1974). *Tractatus logico-philosophicus*. Routledge and Kegan Paul.

CAPÍTULO II

CONVENCIONES LINGÜÍSTICAS

Ricardo Mena[1]

RESUMEN

Es natural pensar que el significado literal de una oración de un lenguaje L está determinado por las convenciones lingüísticas de los hablantes de L. Este capítulo comienza con una discusión breve de lo que es el significado literal. Posteriormente se discutirá el fenómeno de las convenciones lingüísticas y cómo este ha sido usado para explicar la determinación del significado literal. Quine (1969) formuló la objeción principal que toda teoría de las convenciones

---

1 Ricardo Mena es licenciado en filosofía por la Facultad de Filosofía y Letras de la UNAM y doctor en filosofía por la Universidad de Rutgers. Es investigador titular A de tiempo completo del Instituto de Investigaciones Filosóficas de la UNAM y miembro del Sistema Nacional de Investigadores nivel C

Agradecimientos: El autor agradece al proyecto PAPIIT "Escepticismo y racionalidad" (IA 400124), de la DGAPAUNAM, por el apoyo con el que se realizó la presente investigación.

lingüísticas debe enfrentar: cuando un grupo de personas establece una convención, lo hace mediante el uso de un lenguaje determinado. Pero si esto es así, ¿cómo es que el primer lenguaje obtuvo su significado? Si fue por convención, debió de haber un lenguaje previo. Pero entonces, la teoría de las convenciones lingüísticas no ayuda a entender cómo es que el primer lenguaje obtuvo su significado. Por tanto, debe ser posible entender cómo es que un lenguaje adquiere su significado sin apelar a la teoría de las convenciones lingüísticas. El enfoque principal de este capítulo será la teoría de las convenciones de David Lewis (1969 y 1975). Dicho de manera breve, la teoría de Lewis sostiene que las convenciones son regularidades de cierto tipo: son soluciones arbitrarias, que se auto-perpetúan, a partir de problemas de coordinación recurrentes. Lewis entiende a los problemas de coordinación como se hace en teoría de juegos: son situaciones en donde cada uno de los participantes tiene que coordinar sus acciones con el resto para lograr obtener el mejor resultado posible y en donde una falta de coordinación es perjudicial para todos los participantes involucrados. Veremos cómo esta teoría puede ser usada para entender las convenciones lingüísticas en particular, cómo da cuenta de la determinación del significado literal y cómo logra responder a la objeción de Quine. También se discutirán algunas objeciones contemporáneas que se han hecho a esta teoría y se explorarán posibles respuestas.

Palabras clave: Convenciones lingüísticas, Paradoja-Quine, Lewis, Regularidades-autoperpetuantes, coordinación.

ABSTRACT

It is natural to think that the literal meaning of a sentence in a language *L* is determined by the linguistic conventions of the speakers of *L*. This chapter begins with a brief discussion of what literal meaning is. Subsequently, it addresses the phenomenon of linguistic conventions and how this has been used to explain the determination of literal meaning. Quine (1969) pointed out that the establishment of linguistic conventions requires a prior language, thereby creating a paradox: if the first language emerged through

convention, such convention would itself require a pre-existing language to be formulated. This renders it impossible to explain the origin of linguistic meaning solely through conventions. Therefore, it must be possible to understand how a language acquires its meaning without appealing to a theory of linguistic conventions. Consequently, the main focus of this chapter will be David Lewis's theory of conventions (1969 and 1975). Briefly put, Lewis's theory holds that conventions are regularities of a certain kind: they are arbitrary, self-perpetuating solutions to recurring coordination problems. We will examine how this theory can be used to understand linguistic conventions in particular, how it accounts for the determination of literal meaning, and how it manages to respond to Quine's objection. Some contemporary objections to this theory will also be discussed, and possible responses will be explored.

Keywords: Linguistic Conventions, Quine Paradox, Lewis, Self-Perpetuating Regularities, Coordination.

1. Introducción

Podemos discutir qué es, exactamente, el significado. Sin embargo, que las palabras que usamos tienen significado—aunque en ocasiones sea indeterminado cuál es ese significado—difícilmente puede ser cuestionado. De manera similar, parece difícil cuestionar que el significado de una palabra u oración no es una propiedad intrínseca de esa palabra u oración. Las palabras que usamos significan lo que significan, pero pudieron haber significado cualquier otra cosa. Es en virtud de algún tipo de actividad de una comunidad de hablantes de un lenguaje *L* que las palabras de *L* tienen el significado que tienen, y si dicha actividad hubiera sido diferente, el significado de sus palabras hubiera sido otro. Entonces, las palabras tienen el significado que tienen en virtud de cierta relación que mantienen con una comunidad de hablantes; claramente, el significado no es una propiedad intrínseca de las palabras. Las observaciones anteriores sugieren fuertemente que el lenguaje es convencional: las palabras tienen el significado que tienen porque en la comunidad de hablantes hay ciertas convenciones que así lo determinan. Da-

vid Lewis (1969) incluso pensó que es evidente que el lenguaje es convencional:

> Es una verdad de perogrullo que el lenguaje está regido por convenciones. Las palabras pueden ser usadas para significar casi cualquier cosa; y quienes las usamos las hemos hecho significar lo que significan porque de alguna manera, gradualmente e informalmente, hemos llegado a entender que así es como las usaremos. Perfectamente podríamos usar esas palabras de otra forma—o usar otras palabras, como se hace en otros países. Podríamos cambiar de convenciones si así lo queremos. (p.1)[2]

Es difícil estar en desacuerdo con Lewis en este punto: es obvio que el lenguaje es convencional. Lo que no es para nada obvio es qué es exactamente una convención y, en particular, qué es una convención lingüística. Es muy tentador pensar que las convenciones son el resultado de algún tipo de acuerdo explícito: cuando un grupo de personas *P* acuerdan explícitamente que de ahora en adelante se hará *x*—y se apegan a dicho acuerdo—hacer *x* es convencional en *P*.

Sin embargo, como Russell y Quine observaron, si las convenciones son algún tipo de acuerdo, el lenguaje no puede ser convencional. Esto es así, porque si las convenciones son acuerdos explícitos, el surgimiento del primer lenguaje sería imposible. En las palabras de Russell:[3]

> No se sabe cómo adquirieron significación esas raíces, más postular un origen convencional es evidentemente una concepción tan mítica como la del contrato social, según el cual suponían Hobbes y Rousseau que se había establecido el gobierno civil. Es difícil suponer que se haya reunido un parlamento, formado por los jefes privados hasta entonces del lenguaje, y allí se haya coincidido en llamar vaca a una

2 Todas las traducciones de Lewis (1969) serán mías.

3 Por supuesto, Quine (1936) argumentó que las verdades lógicas no pueden ser verdaderas por convención. Lewis (1969) argumenta que, de acuerdo con su teoría de las convenciones, podemos decir que las verdades analíticas son verdaderas por convención.

> vaca y lobo al lobo. La asociación de las palabras con sus significaciones debe haber surgido por algún proceso natural, aunque en la actualidad se desconozca la naturaleza de ese proceso. (1950, p. 209)

Es decir, ¿cómo es que el primer lenguaje es convencional si no hubo un lenguaje previo que la gente pudo usar para generar acuerdos? Quine, en el prólogo de (Lewis, 1969), expresa esa misma preocupación:

> Cuando era niño imaginé a nuestro lenguaje como fijado y transmitido por un grupo de administradores, sentados en convención sería en una mesa al estilo de Rembrandt. Esta imagen permaneció imperturbada por un tiempo por la pregunta de qué lenguaje los sindicados pudieron haber usado en su deliberación, o por el miedo al regreso vicioso. (parr. 3)

Por un lado, parece obvio que el lenguaje es convencional, pero, por otro lado, la noción más obvia de convención lingüística es sumamente problemática. Uno de los objetivos principales de la teoría de las convenciones de Lewis es precisamente ofrecer una teoría que no requiera que las convenciones sean algún tipo de acuerdo explícito. De acuerdo con Lewis (1969), las convenciones son, grosso modo, soluciones arbitrarias y auto-perpetuadoras a problemas de coordinación recurrentes. Lewis utiliza las herramientas de la teoría de juegos para argumentar que las convenciones, entendidas de este modo, no requieren ningún tipo de acuerdo explícito. Agentes involucrados en un problema de coordinación pueden encontrar una solución de varias maneras, sólo una de las cuales es por acuerdo explícito.

Por lo expuesto, la estructura de este artículo se organiza de la siguiente manera. En la sección 2, discutiremos la teoría general de las convenciones de Lewis y, en la sección 3, discutiremos su teoría de las convenciones lingüísticas. En las secciones 4-6, presentaré y responderé a algunas objeciones dirigidas a aspectos centrales de esta teoría.

2. Convenciones

La siguiente es una de las observaciones centrales de Lewis (1969): "El uso del lenguaje pertenece a una clase de situaciones con un carácter conspicuo: situaciones que llamaré problemas de coordinación" (p.5). Otra de las observaciones centrales de Lewis es que las convenciones son, gorsso modo, soluciones arbitrarias y auto-perpetuadoras a problemas de coordinación recurrentes. Entonces, las convenciones lingüísticas son, de acuerdo con esta manera de pensar, soluciones a cierto tipo de problema de coordinación recurrente. En esta sección discutiremos la manera en que Lewis entiende los problemas de coordinación y cómo es que las convenciones, en general, pueden ser entendidas como soluciones de cierto tipo a problemas de coordinación recurrentes. En la siguiente sección discutiremos cómo esta teoría puede ser aplicada a las convenciones lingüísticas.

3. Problemas de Coordinación

Un problema de coordinación es una situación en donde los agentes involucrados deben coordinar sus acciones para obtener resultados aceptables—si actúan de la misma manera, obtendrán resultados positivos, si no lo hacen, obtendrán resultados negativos[4].

Consideremos casos típicos de problemas de coordinación antes de presentar una teoría más detallada acerca de ellos[5]. Supongamos que Celeste y Mariana quedaron de verse en la Universidad a las 12pm, pero olvidaron especificar exactamente en dónde se encontrarán. Celeste y Mariana no se conocen bien y no tienen manera de contactarse vía electrónica. La Universidad es grande y no es obvio cómo resolver el problema. Si Celeste va a la Biblioteca Central y Mariana va al Centro Cultural, habrán fallado en coordinar sus acciones. Dado que desean encontrarse, el resultado

4 Qué constituye actuar de la misma manera requiere algo de refinamiento: puede ser que acciones que aparentemente son distintas, cuenten cómo actuar de la misma manera, bajo alguna descripción no trivial. Aclararemos esto más adelante.

5 Estos ejemplos están basados en (Lewis, 1969).

para ambas será negativo. Para obtener un resultado positivo, ambas deben actuar de la misma forma: ambas deben ir a la Biblioteca Central, al Centro Cultural, al Espacio Escultórico, o a donde fuera. Sólo así podrán encontrar una solución satisfactoria a este problema de coordinación. Importa poco en donde se encuentre, siempre y cuando ambas vayan al mismo lugar.

El siguiente es un ejemplo de Lewis (1969). Supongamos que varias personas quieren conducir en el mismo camino. Algunas quieren ir hacia el norte y otras hacia el sur. A cada una de ellas les importa poco en qué lado del camino tengan que conducir, siempre y cuándo todas manejen del mismo lado. Entonces, para solucionar este problema de coordinación, o bien todas las personas involucradas tienen que manejar de su lado derecho o todas tienen que hacerlo de su lado izquierdo. De lo contrario el resultado será negativo para todas las personas involucradas.

El siguiente es un ejemplo que Lewis tomó del *Tratado de la Naturaleza Humana*, de Hume. Supongamos que estamos en un bote con remos. Queremos llegar al otro lado del lago, aunque en realidad no nos interesa la velocidad en que lo hagamos. Si remamos a velocidades distintas, el bote se va a mover en círculos, en cuyo caso obtendremos un resultado negativo. Si remamos al mismo ritmo, podremos dirigir el bote en la dirección deseada. No nos interesa remar a un ritmo en particular, siempre y cuando lo hagamos al mismo ritmo.

Por último, consideremos otro ejemplo de Lewis. Este ejemplo es interesante porque a primera vista parece que, para coordinarse, las personas involucradas tienen que actuar de maneras distintas. Supongamos que estamos acampando y necesitamos recolectar madera para la fogata. Si todas las personas buscan madera en la misma dirección, la madera que recolectemos será insuficiente. Para recolectar madera suficiente necesitamos cubrir tanto terreno como podamos. En este caso, coordinarse consiste en actuar de una manera distinta al resto de los participantes. Sin embargo, es fácil describir el ejemplo de tal forma que los participantes actúan de la misma forma cuando logran solucionar el problema. Todas

las personas involucradas hacen lo mismo: buscan madera en direcciones opuestas al resto del grupo. Basta con considerar estos ejemplos para darnos cuenta de que, como criaturas sociales, constantemente nos enfrentamos a problemas de coordinación de diversos tipos.

Una de las características centrales de los problemas de coordinación es que las personas involucradas deben actuar, si han de hacerlo racionalmente, de acuerdo con sus expectativas acerca de los otros participantes. Si, por la razón que sea, Celeste espera que Mariana vaya al Centro Cultural, lo más racional para ella es ir al Centro Cultural. Si esperas que las demás personas van a manejar de su lado derecho, lo más racional para ti es manejar de tu lado derecho. Si esperas que tu compañera vaya a remar a cierto ritmo, lo más racional es remar a ese ritmo. Si esperas que las personas en el campamento vayan a buscar madera hacia el norte, sur, y este, lo más racional para ti es buscar madera hacia el oeste. Esta es la manera en que Lewis explica, siguiendo a Schelling (1960), el carácter general de los problemas de coordinación:

> Dos o más agentes deben escoger una de varias acciones alternativas. Con frecuencia todos los agentes tienen el mismo conjunto de acciones alternativas, pero esto no es necesario. Los resultados que los agentes quieren producir o prevenir son determinados conjuntamente por las acciones de todos los agentes. Así que el resultado de cualquier acción que un agente pueda escoger depende de las acciones de los otros agentes. Por eso—como hemos visto en cada ejemplo—cada uno debe escoger qué hacer de acuerdo con sus expectativas acerca de lo que los otros van a hacer. (1969, p. 8)

Ciertas combinaciones de acciones cobran particular importancia en un problema de coordinación. Se trata de combinaciones de acciones tales que ninguna de las personas involucradas podría haber obtenido un mejor resultado dada la manera en que el resto de las personas actuaron. Estas combinaciones de acciones son llamadas equilibrios. La situación en donde Celeste y Mariana van al Centro Cultural es un equilibrio: en esta situación, Celeste no pudo haber

obtenido un mejor resultado actuando de otra manera, dada la manera en que Mariana actuó. Por supuesto, otro equilibrio es la situación en donde Celeste y Mariana van a la Biblioteca Central. De manera similar, la situación en donde las personas en el campamento van a buscar madera hacia el norte, sur, este y oeste (respectivamente) es un equilibrio: ninguna de las personas involucradas pudo haber obtenido un mejor resultado actuando de manera diferente, dada la manera en que el resto actuó[6]. Para encontrar un equilibrio, las personas involucradas deben coordinar sus acciones, y deben hacerlo conforme a sus expectativas acerca de la manera en que las otras personas involucradas van a actuar.

Una manera estándar de representar formalmente un problema de coordinación es la siguiente:

	m1	m2
c1	1-1	0-0
c2	0-0	1-1

Figura 1
Nota: Cuadro elaborado por el autor

Este cuadro representa un problema de coordinación de dos agentes, los cuales tienen a su disposición dos acciones posibles. Supongamos que los agentes son Celeste y Mariana. Representemos a la acción de ir al Centro Cultural, con 1 y a la acción de ir al Espacio Escultórico, con 2. Entonces, *c1* representa la situación en donde Celeste va al Centro Cultural y *c2* la situación en donde Celeste va al Espacio Escultórico. *m1* y *m2* son las acciones correspondientes de Mariana. Claramente, en este problema de coordinación, los

6 Es fácil encontrar los equilibrios en los otros ejemplos.

equilibrios son los pares (*c1*, *m1*) y (*c2*, *m2*): en el primer equilibrio, dado que Mariana actuó de la manera 1, Celeste no pudo haber obtenido un mejor resultado actuando de otra manera y dada la manera en que Celeste actuó, Mariana no pudo haber obtenido un mejor resultado actuando de otra manera. Algo similar ocurre en el segundo equilibrio.

Hay muchos aspectos técnicos relacionados con los problemas de coordinación. Varios de ellos son tratados en Lewis (1969). Una presentación más reciente puede ser encontrada en Clark (1992). Para los propósitos de este artículo, lo que hemos dicho hasta ahora acerca de los problemas de coordinación será suficiente. En la siguiente sección veremos cómo es que, de acuerdo con Lewis, las convenciones pueden ser entendidas como soluciones arbitrarias y auto-perpetuadoras a problemas de coordinación recurrentes. Más adelante veremos cómo esta teoría general de las convenciones puede ser usada para entender a las convenciones lingüísticas.

4. Teoría Lewisiana de las Convenciones

De acuerdo con Lewis, las convenciones son soluciones arbitrarias y auto-perpetuadoras de problemas de coordinación recurrentes. Estas soluciones tienen que ser arbitrarias, porque las convenciones son arbitrarias: manejar del lado derecho es una convención en América Latina, pero fácilmente pudimos haber adoptado la convención de manejar del lado izquierdo, como en Inglaterra. Ninguna de las dos convenciones es preferible, en sí misma, sobre la otra: la elección de una sobre la otra es, en este sentido, arbitraria. Estas soluciones tienen que ser auto-perpetuadoras porque, de acuerdo con Lewis, las convenciones así lo son: una persona tiene razones para seguir conformándose a una convención siempre y cuando el resto lo siga haciendo. Esto provoca que los miembros de una comunidad que sigue cierta convención se sigan apegando a ella: por eso es que las convenciones son auto-perpetuadoras. Hay que notar que también es justo decir que las convenciones son regularidades de cierto tipo. Cuando una población soluciona—de manera arbitraria y auto-perpetuadora—un problema de coordinación recu-

rrente, podemos decir que esto conforma una regularidad: cada vez que surge ese problema de coordinación, la población lo soluciona de la misma manera arbitraria y auto-perpetuadora.

Lewis (1969) considera varios análisis de la noción de convención. Sin embargo, aquí consideraremos la versión más reciente y pulida :

Una regularidad *R*, en acción o en acción y creencia, es una convención en la población *P* si y sólo si, en *P*, las siguientes seis condiciones se mantienen (o por lo menos si casi se mantienen. Unas pocas excepciones a "todos" se pueden tolerar):

1. Todos se conforman a *R*.
2. Todos creen que los otros se conforman a *R*.
3. La creencia de que otros se conforman a *R* le da a todos una buena y decisiva razón para conformarse a *R*.
4. Hay una preferencia general por la conformidad general a *R* en lugar de un poco menos que una conformidad general a *R*—en particular, en lugar de la conformidad de todos menos uno.
5. *R* no es la única regularidad posible que cumple con las últimas dos condiciones. Hay por lo menos una alternativa *R′* tal que la creencia de que otros se conforman a *R′* daría a todos una buena y decisiva razón práctica o epistémica para conformarse a *R′* de igual manera […]
6. Finalmente, los varios hechos listados en las condiciones de (1) a (5) son cuestión de conocimiento común (o mutuo): son sabidos por todos, es sabido por todos que son sabidos por todos, etcétera. (1975, p. 165)

A continuación, explicaré brevemente cada una de las cláusulas. (1) y (2) son fáciles de entender, pero sí hay que enfatizar que, como en el resto de las cláusulas, se admiten excepciones: no es necesario que todas las personas en P se conformen a R, o que todas crean que todas se conforman a R, siempre y cuando casi todas lo hagan. Esto tiene sentido; si bien es cierto que en América Latina casi todo mundo maneja del lado derecho casi todo el tiempo, que algunas

personas no lo hagan ocasionalmente no impide que haya una convención de manejar del lado derecho. Tampoco sería un impedimento si no todas las personas en América Latina creen que todas las personas manejan del lado derecho.

Para explicar (3), regresemos a uno de los ejemplos. Supongamos que Celeste y Mariana se encuentran, después de todo, en el Centro Cultural. La próxima vez que se enfrenten al mismo problema de coordinación—¿en dónde encontrarse dentro de la Universidad? —habrá una diferencia importante. Ahora tendrán un precedente: ambas sabrán que en la situación anterior resolvieron el problema de coordinación de cierta manera y esto ahora les da una razón de peso para resolver el problema de la misma manera: yendo al Centro Cultural. La fuerza de los precedentes en las convenciones es muy importante; es parte de lo que hace que las convenciones sean auto-perpetuadoras. La tercera vez que se enfrenten al mismo problema de coordinación tendrán aún más razón para resolverlo de la misma manera, dados los dos precedentes anteriores. Supongamos que después de un tiempo se ha establecido una regularidad (R) entre Celeste y Mariana de encontrarse en el Centro Cultural. Supongamos que Celeste cree que Mariana se conforma a dicha regularidad, de tal forma que la próxima vez que queden de verse en la Universidad, Mariana irá al Centro Cultural. Si este es el caso, esto le da a Celeste una razón decisiva para ir al Centro Cultural la próxima vez. Sería irracional para Celeste ir a otro lugar, si cree que Mariana irá al Centro Cultural[7].

En (3) se captura la idea de que en un problema de coordinación lo más importante es encontrar el equilibrio y no tanto qué equilibrio encontrar. Es decir, en un problema de coordinación, los participantes prefieren fuertemente solucionar el problema

7 Claro, es difícil aceptar excepciones cuando se trata de una convención entre dos personas. Pero pensemos en una convención similar entre una población más grande. Supongamos que es convención entre las personas que viven en un barrio encontrarse en la única fuente que tiene ese barrio. Esta puede ser una convención, aunque no todas las personas que viven en ese barrio se encuentren en la fuente cada vez que queden de verse sin especificar un lugar.

de alguna manera y no tanto que el problema se solucione de una manera en particular. Celeste podría preferir encontrarse con Mariana en la Biblioteca Central, pero es más importante para ella encontrarse con Mariana en donde sea, que encontrarse en la Biblioteca Central. Algunas personas podrían preferir manejar del lado derecho—por la razón que sea—pero es más importante para cada persona que en su población se maneje del mismo lado, a que se maneje del lado que ellas prefieren.

Por su parte, (4) es fácil de explicar. Si se cumple la condición de que todas (o casi todas) las personas manejan del lado derecho (*R*), casi todos preferirán que cualquier otra persona maneje del lado derecho. Esta situación sería la más beneficiosa para cada miembro de la población, porque, si obtiene en la mayoría de los casos, las personas relevantes tendrán garantizado encontrarse en una situación de equilibrio. Recordemos que es de interés común resolver los problemas de coordinación de la misma manera. Entonces es de interés común que todas las personas se conformen con la regularidad que la población utiliza para solucionar el problema de coordinación en cuestión.

(5) captura el hecho de que las convenciones son arbitrarias. En América Latina manejamos del lado derecho (*R*), pero si ocurriera que la mayoría manejara del lado izquierdo (*R'*), preferiríamos que cada persona maneje del lado izquierdo (en donde no es posible manejar de ambos lados a la vez, por supuesto). Es decir, manejamos del lado derecho, pero bien pudimos haber adoptado la convención de manejar del lado izquierdo. Si creyéramos que todas las personas manejan del lado izquierdo, eso nos daría una razón decisiva para manejar del lado izquierdo.

(6) es un poco más difícil de explicar. Es conocimiento común en la población *P* que *p* si y sólo si todas las personas en *P* saben que *p* y todas las personas en *P* saben que todas saben que *p*, y todas las personas en *P* saben que todas saben que todas saben que *p*, y así, al infinito. De acuerdo con Lewis, si *R* es una convención en *P*, entonces es de conocimiento común en *P* que (1)-(5). Claro, Lewis está al tanto de que difícilmente las personas en una población tie-

nen conocimiento común explícito de (1)-(5). Sin embargo, Lewis (1975) deja claro que basta con que este conocimiento común sea tácito o incluso potencial: es decir, que sería fácil para las personas en una población adquirir conocimiento común de (1)-(5) (o alguna formulación alternativa de esas cláusulas) si se preocuparan por pensar en ello. La cláusula (6) garantiza estabilidad. Si (6) obtiene—además del resto de las cláusulas—todas, o casi todas, las personas en *P* estarán en posición de solucionar el problema de coordinación relevante de la misma manera, ya que estarán en posición de saber cómo es que otras personas en *P* confrontarán dicho problema de coordinación. Por ejemplo, cuando Celeste es confrontada con el problema de encontrarse con Mariana en la Universidad, sabrá que Mariana se conforma a la regularidad de ir al Centro Cultural y que ella sabrá que Celeste se conforma a la misma regularidad y que también sabrá que Celeste sabe que ella se conforma a la misma regularidad, etcétera. Esto garantizará—suponiendo que son racionales y que desean solucionar el problema de coordinación—que Celeste y Mariana se encontrarán en el Centro Cultural, en esta instancia del problema de coordinación y en instancias futuras[8].

4.1. Convención Lingüística

Si las convenciones lingüísticas son convenciones en el sentido que explicamos en la sección anterior, debe ser el caso que estas convenciones son soluciones regulares a problemas de coordinación. ¿Cuál es, entonces, el problema de coordinación que las convenciones lingüísticas solucionan?

Para entender la respuesta a esta pregunta será de utilidad atender primero la manera en que Lewis entiende a las señales[9]. Para ello, usemos el ejemplo clásico de Lewis (1969). Paul Revere y los sacristanes quieren organizar una defensa ante un ataque de los ingleses. Si el ataque de los ingleses llega por tierra, quieren organizar

8 Para ver casos complicados en donde (6) se necesita para alcanzar coordinación, ver Clark (1996).

9 Para una excelente elaboración de la teoría de las señales de Lewis, ver Skyrms (1996)

la defensa de cierta manera y si el ataque llega por mar, quieren organizar la defensa de cierta otra manera. Los sacristanes son quienes están en posición de saber si el ataque llegará por tierra o por mar, así que ellos emitirán una señal para indicar a Revere cuál es el caso. Los sacristanes utilizaran linternas para emitir la señal. Sin embargo, una señal no es de utilidad a menos que las partes involucradas se coordinen en una forma de interpretarla. Supongamos, para simplificar el ejemplo, que hay dos maneras de interpretar las señales: (1) una linterna señala que el ataque viene por tierra y dos linternas que el ataque viene por mar y (2) una linterna señala que el ataque viene por mar y dos linternas que el ataque viene por tierra. Los sacristanes y Revere quieren coordinarse ya sea en (1) o en (2): no importa de qué manera se interpreten las señales, siempre y cuando ambos lo hagan de la misma manera. Sin duda esta situación tiene la forma de un problema de coordinación: se obtiene el equilibrio si tanto los sacristanes como Revere eligen interpretar las señales de la misma manera (ya sea como en (1) o como en (2)). Pero, si los sacristanes eligen (1) y Revere (2), habrán fallado en solucionar el problema de coordinación: en este caso, si los sacristanes muestran una linterna, es porque el ataque llegará por tierra, pero Revere entenderá que el ataque viene por mar.

Lo que Lewis muestra con el ejemplo de los sacristanes y Revere es que podemos entender fácilmente a la comunicación por medio de señales en términos de problemas de coordinación, en donde el tipo de acción que soluciona o no el problema es la interpretación de las señales. Si este es el caso, también es fácil entender a la comunicación lingüística en términos de problemas de coordinación. El problema de coordinación consiste en cómo interpretar una emisión lingüística dada: si la persona que habla y la audiencia interpretan la emisión de la misma manera, habrán solucionado el problema, de lo contrario habrán fallado. En principio, no importa qué lenguaje sea seleccionado, siempre y cuando se haya seleccionado el mismo lenguaje: un participante del problema de coordinación no adquiere ningún beneficio al escoger un lenguaje distinto al escogido por el resto de los participantes. A continua-

ción, veremos en qué consisten las convenciones que constituyen soluciones recurrentes a este tipo de problema de coordinación. Sin embargo, antes de hacerlo, hay que entender brevemente la manera en que Lewis caracteriza formalmente a los lenguajes.

De acuerdo con Lewis (1975), un lenguaje es una función que toma como argumentos oraciones y arroja como valores proposiciones. Entonces, podemos entender al lenguaje como una asignación de significados a oraciones. Lewis entiende a las proposiciones como conjuntos de mundos posibles: los mundos posibles en donde la oración es verdadera[10]. Entonces, elegir una interpretación es cuestión de elegir un lenguaje, entendido de esta manera. Por tanto, para coordinarnos en la interpretación de una oración basta con coordinarnos en la elección de un lenguaje. Entonces, una forma de solucionar el tipo de problema de coordinación que se presenta en una instancia de comunicación lingüística es coordinándonos en la selección de un lenguaje. Para decir que un lenguaje es convencional, hay que seleccionar el mismo lenguaje de manera regular y hacerlo cumpliendo las condiciones de (1)-(6). Sin embargo, cuál es exactamente la convención relevante es una cuestión delicada.

Podríamos pensar que la regularidad que una población *P* tiene que adoptar para hablar un lenguaje *L* es simplemente la regularidad de usar *L* o la regularidad de significar con una emisión de una oración por lo que *L* asigna a *p*, para cualquier *p*. Sin embargo, Lewis (1975) deja muy claro que esta opción, por sencilla que sea, no va a funcionar. Cuando queremos explicar la convencionalidad del lenguaje, justamente lo que queremos explicar es qué es lo que hace que una población hable el lenguaje que habla o qué es lo que hace que una población significa lo que significa al usar ciertas oraciones. Pero entonces, no podemos incluir estas nociones como parte del análisis. Hace falta buscar una regularidad más sofisticada.

10 Este no es un aspecto esencial de su teoría de las convenciones lingüísticas: podemos entender a las proposiciones de otra manera sin tener que abandonar los aspectos centrales de su teoría de las convenciones.

Cuando usamos el lenguaje, lo hacemos para lograr diversos objetivos. Sin embargo, de acuerdo con Lewis (1969, 1975) uno de los objetivos centrales es el de transmitir nuestras creencias. En muchas ocasiones, nos comunicamos porque tenemos una creencia que queremos que nuestros interlocutores compartan. También, muchas veces cuando escuchamos a un hablante es porque queremos adquirir una de sus creencias. Es difícil negar que este es uno de los propósitos centrales de la comunicación lingüística. Cuando Celeste me dice que se ha descompuesto el lavabo de la cocina, parte central de lo que quiere lograr es que yo comparta su creencia de que el lavabo de la cocina se ha descompuesto. De manera similar, cuando me acerco a escuchar a Celeste, parte de lo que quiero lograr es adoptar algunas de las formas en que Celeste cree que el mundo es: confío en que Celeste tiene cierta información acerca del mundo y quiero adquirir dicha información. En este ejemplo particular, Celeste tiene información acerca del estado del lavabo de la cocina y yo tengo interés en adquirir esa información, formando la creencia de que el lavabo se ha descompuesto, si Celeste dice que el lavabo se ha descompuesto.

Teniendo lo anterior en mente, Lewis (1975) propone que la convención por medio de la cual una población usa un lenguaje *L* es la convención de veracidad y confianza en *L*. Ser veraces en *L* significa tender a emitir únicamente aquellas oraciones de *L* que creemos que son verdaderas. Por su parte, ser confiados en *L* significa tender a creer que las oraciones emitidas por otras personas son verdaderas en *L*. En mi interacción anterior con Celeste, ella y yo fuimos veraces y confiados en *Español*[11]. Ella fue veraz, porque sólo emitió una oración que ella cree que es verdadera en *Español* y yo fui confiado en *Español* porque creí que la oración que ella emitió es verdadera en *Español*.

11 Por simplicidad estoy asumiendo que el lenguaje español se puede capturar simplemente como una función de oraciones a proposiciones. Probablemente la historia completa es más complicada: por ejemplo, dar cuenta del fenómeno de la indeterminación lingüística requeriría complicar la teoría un poco.

Así es como, de acuerdo con Lewis, solucionamos el problema de coordinación presentado por la comunicación lingüística: nos coordinamos en un lenguaje entre muchos, adquiriendo una convención de veracidad y confianza en ese lenguaje. Cuando, por ejemplo, una población *P* adquiere la convención de veracidad y confianza en *Español* esto garantizaría que, por lo general, las personas en *P* serán capaces de cumplir con uno de los propósitos centrales de la comunicación lingüística: lograr que los hablantes de *P* transmitan sus creencias a sus interlocutores de *P*.

5. Objeciones a la Teoría de Lewis

La teoría de las convenciones y de las convenciones lingüísticas de Lewis ha generado una cantidad inmensa de discusión. A la fecha, sigue siendo la teoría más popular entre personas dedicadas a la filosofía del lenguaje. No obstante, esta teoría ha sido presentada con un número importante de objeciones, las cuales han motivado propuestas alternativas; aunque usualmente inspiradas en la propuesta de Lewis. Algunas de las objeciones que se han presentado a esta teoría conciernen detalles de la misma. Algunas otras conciernen aspectos centrales de la teoría. A continuación, presentaré algunas de las objeciones del segundo tipo y presentaré algunas de las maneras en que la teoría de Lewis puede responder a ellas.

5.1. La Convención de Veracidad y Confianza

No es tan difícil mirar a la convención de veracidad y confianza con sospechas; después de todo, parece que no decir la verdad y no creer que se nos ha sido dicho la verdad son prácticas comunes. No es inusual que la gente mienta. Algunas personas lo hacen de manera muy frecuente, como la gente dedicada a la política. Otras personas lo hacen de manera esporádica: mucha gente emite mentiras blancas de vez en cuando (no es inusual mentir para no lastimar a la persona con la que hablamos). En conversaciones casuales, muchas veces decimos cosas que sabemos que no son verdaderas, pero las decimos de cualquier forma por conveniencia. Por ejemplo, cuando la estamos pasando mal y nos preguntan cómo estamos, es usual

responder diciendo que estamos bien. La otra cara de la moneda es que muchas veces no creemos que se nos ha dicho algo verdadero. Las personas sensatas no creen lo que un político dice, muchas veces sabemos que se nos está diciendo una mentira blanca o que la persona que habla no está diciendo algo verdadero para evitar complicaciones, o por cualquier otra razón. Si la gente suele decir cosas falsas y suele no creer en lo que se le dice, parece que la convención de veracidad y confianza es una fantasía. Esta es de hecho la postura de Williams (2002), Davis (2003) y García-Ramírez (ms.). Si esto es correcto, la propuesta de Lewis no cuenta con una respuesta a la pregunta acerca de qué es lo que hace que una población *P* use un lenguaje *L*.

Algo que hay que resaltar de inmediato es que, de acuerdo con Lewis, las convenciones admiten excepciones. Una población puede adoptar la convención de manejar del lado derecho aun cuando algunas personas, ocasionalmente, manejen del lado izquierdo. Que esto último ocurra de vez en cuando no impide que la población en cuestión de hecho se adhiera a la convención de manejar del lado derecho. De manera similar, la convención de veracidad y confianza también admite excepciones: no es problema si de vez en cuando no actuamos en conformidad con esa convención. La objeción que ahora nos ocupa debe ser, entonces, que las personas mienten de manera muy frecuente: que lo hacen de manera tan frecuente que no tiene sentido decir que con regularidad somos veraces y confiados en el lenguaje que hablamos.

Un problema que enfrentamos a la hora de evaluar esta objeción es que Lewis no dejó claro qué tanto es tantito. Es decir, Lewis dejó claro que algunas excepciones a las convenciones son permitidas, pero no dejó claro exactamente cuantas excepciones son permitidas. Algo que es seguro es que, si en la gran mayoría de los casos las personas en *P* no actúan en conformidad con *R*, entonces *R* no es una convención en *P*. Pero no es claro, por ejemplo, qué tantas veces tenemos que manejar del lado izquierdo para que manejar del lado derecho no sea una convención. Tampoco queda

claro cuántas veces podemos mentir antes de que la veracidad y confianza deje de ser una convención.

Lo que queda claro es que si una población *P* miente en *L* en la gran mayoría de los casos, no habrá una convención de veracidad y confianza en *L* en *P*. Muy bien, pero también queda claro—y esto es crucial—que en este caso *L* no será el lenguaje de *P*. Es decir, no puede haber una población que mayormente mienta en *L* y que *L* sea el lenguaje de esa población. Las razones son las siguientes. Supongamos que estoy hablando con Celeste y que estamos observando una piedra que sabemos que es amarilla. Ahora imaginemos que yo digo—claramente apuntando a la piedra en cada instancia—"Esa manzana es azul" y más tarde digo "Esa toronja es verde" y luego "Ese chocolate es morado". Claramente aquí no estoy siendo veraz en *Español* y con toda probabilidad Celeste no sería confiada en *Español*. Si la vasta mayoría de las interacciones en *P* son así, no habrá convención de veracidad y confianza en *Español* en *P*. Pero también es cierto que el lenguaje de esta población no sería el *Español*. No parece que podamos decir que en *P* las palabras "manzana", "toronja", "chocolate", "azul", "verde" y "morado" significan lo que significan en Español. Lo mismo se puede decir acerca del resto de las palabras. Claro, es difícil pensar que una población pueda usar un lenguaje de una manera tan errática, y me parece que ese es precisamente el punto de Lewis. Para que estas palabras signifiquen lo que significan en *Español*, tiene que ocurrir que con frecuencia cuando los hablantes dicen "Esa manzana es verde" sea claro que crean que apuntan a una manzana que es verde. Es decir, para que una población P hable *Español* tiene que ser el caso que con frecuencia las personas en P son veraces en *Español*.

Sin duda, la postura de Williams, Davis y García-Ramírez no puede ser que de hecho mentimos de manera tan exagerada y frecuente como en el ejemplo anterior. El punto que quiero hacer es que para que una población *P* hable un lenguaje *L* tiene que ser el caso que las personas en *P* son veraces y confiadas en un número suficiente de casos. Quienes objetan pueden conceder esto e insistir que aun así se miente de manera tan frecuente y tan clara que es incorrecto decir

que de hecho hay una convención de veracidad y confianza. Hay varias cosas que decir acerca de esto en defensa de Lewis.

Lo primero es que no queda tan claro que mentimos tanto como la objeción sugiere. Consideren a un político cualquiera. Como tal, dice muchas mentiras. Pero también es cierto que con toda probabilidad dice muchísimas más cosas que cree que son verdaderas. Claro, no dice muchas cosas que cree que son verdaderas cuando se dirige a la población, pero si lo hace en su vida diaria. Cuando no se dirige a la población, el político dice toda serie de cosas que cree que son verdaderas acerca del clima, el color de las cosas, su fecha de nacimiento, en qué ciudad vive, el número de hijas que tiene, etcétera. Mi impresión es que en la vida diaria incluso el político dice un mar de cosas que cree que son verdaderas y que son proporcionalmente muchas más que las instancias en las que miente. La mayoría de nuestras emisiones tienen que ver con la vida cotidiana y ahí tenemos un incentivo muy claro para ser mayormente veraces, aunque en ocasiones mintamos acerca de estos temas. Es decir, un gran número de nuestras emisiones están dedicadas a coordinarnos con otras personas acerca de cuestiones prácticas, y en esos casos tenemos incentivos claros para ser veraces acerca de qué día es, cómo se llama el restaurante en el que nos vamos a ver, si ofrecen comida mexicana o japonesa, si estaré disponible la próxima semana, si soy intolerante a la lactosa, etcétera. Con frecuencia somos veraces.

Claro, un puñado de mentiras nos pueden parecer demasiadas: desde un punto de vista moral, tenemos poca tolerancia a las mentiras. Si un amigo miente tres veces en una conversación, nos puede parecer que miente mucho, aunque también haya dicho doscientas cosas que cree que son verdaderas. Pero lo que está en juego aquí no es si nos parece que un político ha dicho muchas mentiras en la semana (ciento veinte, por ejemplo), pero cuál es la proporción de mentiras a verdades. Si un político ha dicho en la semana, supongamos, cinco mil cosas que él cree que son verdaderas, parece que sí es veraz con regularidad. De manera similar, si en una semana se ha manejado cierto número de veces del lado

izquierdo, puede ser un problema dependiendo de cuántas veces se ha manejado del lado derecho. Me da pena aceptar que en mi vida probablemente he manejado del lado izquierdo aproximadamente diez veces. En algún sentido esto puede parecer mucho. Sin embargo, dado que he manejado muchísimas más veces del lado derecho, que haya manejado del lado izquierdo diez veces no es obstáculo para decir que con regularidad manejo del lado derecho y que sin duda me apego a la convención de manejar del lado derecho. Si es plausible decir que un político es veraz con regularidad, no hay obstáculo alguno para decir que la mayor parte de la población es veraz.

Si con frecuencia somos veraces en un lenguaje L, es razonable pensar que con frecuencia tenemos confianza en L. Por ejemplo, cuando nos enganchamos en conversaciones cotidianas, tenemos incentivos claros para creer que lo que se nos dice acerca del día, el nombre del restaurante, si es de comida mexicana o japonesa, etcétera, es verdadero. Este es el caso, en buena parte, porque normalmente queremos coordinar nuestras acciones con otras personas. También, uno de los objetivos principales de muchas de nuestras conversaciones es responder la pregunta bajo discusión (Roberts (1996/2012)). Es natural pensar que nuestras discusiones están en buena parte estructuradas por nuestro deseo conjunto por responder a una—o varias—preguntas. Es común plantear, hacia el inicio de una conversación, preguntas como: ¿te gustaría ir a comer en la tarde?, ¿sabes si mañana tendremos seminario?, ¿tu gatita ya se recuperó? o ¿qué piensas acerca de las próximas elecciones?[12]. Cuando este tipo de preguntas son planteadas y aceptadas en una conversación, las personas involucradas tienden a coordinarse para responderlas. A menos que tengamos una razón de peso para no

12 En ocasiones estas preguntas no se plantean de manera explícita, pero esto no quiere decir que en esas conversaciones no haya una pregunta bajo discusión. Por ejemplo, si comienzo a hablar de la recuperación de mi gatita, mis interlocutores entenderán que estoy proponiendo que la pregunta bajo discusión sea, por ejemplo, ¿cómo se está recuperando mi gatita?. Esto provocará que se acomode el discurso para incluir esa pregunta como la pregunta bajo discusión. Para más acerca de estos mecanismos pragmáticos relacionados con las preguntas bajo discusión, ver (Roberts, 1996).

creer en las respuestas que se nos ofrecen, tendemos a creerlas y, por tanto, tendemos a ser veraces.

Por supuesto, cuando escuchamos un político o una persona que nos quiere vender algo, tendemos a no creer lo que se nos dice. Hay varias formas en que podríamos explicar por qué esto no es un problema para la convención de veracidad y confianza. Creo que vale la pena explorar la siguiente manera de hacerlo. En ocasiones, cuando las circunstancias lo requieren, podemos suspender temporalmente algunas convenciones. Por ejemplo, si alguna avenida necesita ser reparada, el gobierno de la ciudad puede suspender la convención de manejar por el lado derecho cuando la gente maneja por esa avenida. Supongamos que el estado de la avenida es tal que la mejor manera de transitarla es manejando por el lado izquierdo. Incluso el gobierno podría incluir señales indicando que así es como se debe transitar por esta avenida durante las reparaciones. No me parece que un caso así cuente como uno en donde en esa avenida hay una serie de violaciones de la convención de manejar por el lado derecho. Me parece más natural pensar que la convención se ha suspendido temporalmente en ese tramo de la avenida. Creo que algo similar ocurre en algunas de nuestras conversaciones. Cuando participamos en conversaciones tales que queda claro para los participantes que la gente no va a ser veraz, la gente también tenderá a no ser confiada (pueden, por ejemplo, pretender que están siendo confiadas sin serlo).

Quizá algo similar podemos decir cuando en una conversación se hace un uso excesivo de metáforas y otras formas de lenguaje no literal. Quizá en esos casos tiene sentido decir que por el momento la convención de veracidad y confianza está siendo suspendida. Esta es una buena opción cuando las personas en la conversación no tienen como objetivo comunicar sus creencias. Si, en cambio, tienen la intención de comunicar sus creencias por medio de un uso no literal del lenguaje, tenemos a nuestra disposición la explicación que Lewis (1975) dio de esos casos.

5.2. El problema de la Innovación Léxica

Davidson (1986) argumentó que las convenciones lingüísticas no juegan un papel esencial en la explicación de la comunicación lingüística. Si Davidson está en lo correcto, la propuesta de Lewis perdería mucha de su fuerza explicativa. Recordemos que para Lewis la comunicación lingüística presenta, en todo caso, un problema de coordinación, que es resuelto cuando las personas en la conversación se coordinan en la selección de un lenguaje. Una vez que se adquiere una convención de veracidad y confianza en *L* en una población *P*, las personas en *P* solucionan los problemas de coordinación presentados por la comunicación lingüística seleccionando *L*. Parece ser que, de acuerdo con Lewis, las convenciones lingüísticas sí juegan un papel esencial en la explicación de la comunicación lingüística.

Davidson piensa que el fenómeno de la innovación léxica (como ha sido llamado por Armstrong (2016)) muestra que las convenciones lingüísticas no juegan un papel esencial en la explicación de la comunicación lingüística. Para aclarar, Davidson no niega que las convenciones lingüísticas existen, lo que niega es que jueguen un papel esencial en la explicación de la comunicación. A continuación, explicaré en qué consiste la innovación léxica. También explicaré por qué Davidson pensó que estos fenómenos presentan un problema para la teoría de las convenciones lingüísticas. Por último, argumentaré que en realidad estos fenómenos no generan los problemas que Davidson piensa que generan.

Los casos de innovación léxica son aquellos en donde una palabra que aún no tiene un significado convencional es usada y, sin embargo, esto no es obstáculo para que haya comunicación exitosa. Estas palabras que aún no tienen significado lingüístico son palabras novedosas, que no han sido usadas con anterioridad en otras conversaciones. Consideremos algunos casos de innovaciones léxicas (las palabras novedosas están en cursivas)[13]:

13 Los primeros dos ejemplos son tomados de Armstrong (2016)

A. El preso se *haudineó* de la celda.
B. Edu *tequileó* el ponche.
C. Ese barrio ya se *romeó*[14].

Cuando emitimos (A) comunicamos que Bruno se ha escapado de su celda, cuando emitimos (B) comunicamos que Edu añadió tequila al ponche, y cuando emitimos (C) comunicamos que el barrio en cuestión se ha gentrificado. Es fácil generar instancias de comunicación exitosa con estas oraciones, aunque estas contengan palabras que no tienen un significado convencional[15]. El punto de Davidson es precisamente que estos son casos de comunicación exitosa que no dependen de las convenciones lingüísticas: las personas involucradas en la conversación no tienen conocimiento común de los significados convencionales de las palabras en cursivas, porque esas palabras aún no cuentan con significados convencionales. Por tanto, no podemos explicar estas instancias de comunicación simplemente apelando al conocimiento común de convenciones lingüísticas. Entendida de la forma más directa, la objeción de Davison es claramente incorrecta. La explicación de cómo logramos comunicarnos con usos de (A)-(C) necesita apelar a nuestro conocimiento de las convenciones lingüísticas que gobiernan las palabras que no están en cursivas. No podríamos entender el ejemplo (A) a menos que entendamos el significado convencional de "el preso se" y "de la celda". Entonces, en casos de innovación léxica las convenciones lingüísticas sí son parte esencial de la explicación. Esta simple observación ya pone la objeción de Davidson en bastantes aprietos.

Quizá la objeción deba ser entendida de la siguiente manera: Usos exitosos de (A)-(C) deben ser explicados únicamente apelando al conocimiento común de convenciones lingüísticas. Si esta es la objeción de Davidson, la observación del párrafo anterior no es suficiente para derribarla. Sin embargo, no queda claro cuál es el punto de esta objeción. Si resulta ser el caso que no podemos

14 En la Ciudad de México, la Roma es una colonia totalmente gentrificada.
15 Quizá ahora esas palabras ya tengan un significado convencional en México. Si es así, pensemos en los momentos en que esas palabras fueron usadas por primera vez.

explicar las instancias de innovación léxica únicamente apelando a nuestro concomiendo de convenciones lingüísticas, no se sigue que las convenciones lingüísticas no son esenciales en la explicación de la comunicación. Únicamente se sigue que nuestro conocimiento de convenciones lingüísticas no es lo único que se necesita para explicar casos de innovación léxica, cosa que no es para nada problemática para la propuesta de Lewis.

Aún si Davidson está en lo correcto al requerir que, si las convenciones lingüísticas son esenciales para explicar la comunicación, los casos de innovación léxica deben de ser explicados únicamente en términos de nuestro conocimiento de convenciones lingüísticas, su objeción debe ser refinada. Para explicar cómo podemos comunicarnos exitosamente con un uso de, por ejemplo, "Bruno compró un banco" necesitamos apelar, no sólo a nuestro conocimiento de convenciones lingüísticas, sino también al conocimiento de los mecanismos pragmáticos que nos permiten desambiguar "banco" de la manera correcta. De manera similar, para que haya comunicación exitosa con usos de oraciones como "Celeste está aquí" es necesario tener conocimiento de ciertas convenciones lingüísticas, pero también hace falta tener conocimiento del contexto conversacional. Entonces quizá la objeción deba ser entendida de cierta manera: Lewis debe poder explicar A-C únicamente con base en mecanismos pragmáticos, nuestro conocimiento común de convenciones lingüísticas y del contexto conversacional. Nuevamente, no me queda claro que debamos tomar la objeción con mucha seriedad, ya que no queda claro que este bien motivada. Sin embargo, creo que aun así podemos cumplir con los requerimientos de Davidson. A continuación, ofreceré dos maneras en que podríamos hacerlo. Primero argumentaré que los casos (interesantes) de innovación léxica pueden ser entendidos como usos de comillas mixtas y que las comillas mixtas tienen significado convencional. Después exploraré brevemente, basado en ideas de Rita Fernández, una opción basada en mecanismos pragmáticos. Por último, consideraré una propuesta de Armstrong (2016) para solucionar el problema.

Los siguientes son ejemplos que comúnmente aparecen en la literatura sobre comillas mixtas (las comillas sencillas son comillas mixtas y las dobles comillas puras):

- Miren, 'Quine' viene a platicar con nosotros. (Burlándonos Miguel, quien piensa que "Quine" refiere a Carlos.)
- Aura piensa que su papá es un 'filtosofo'. (Reportando la creencia de una niña que está confundida acerca de cómo pronunciar "filósofo".)
- Pues sí, en ese caso los ajedrecistas son 'atletas'. (Discutiendo con alguien que tiene un uso no estándar de "atleta".)

Shan (2010), Maier (2014), Recanatti (2001) piensan que el significado convencional de las comillas mixtas ('') es similar al de un operador cuya función es asignar una interpretación contextualmente sobresaliente a las palabras bajo su alcance. Una forma de entender la semántica de las comillas mixtas es la siguiente:

- $['\sigma']i,w = [\sigma]i(x),w$

En donde ['σ']i,w es el valor semántico de 'σ' relativo a una interpretación i y un mundo w y $[\sigma]i(x),w$ es el valor semántico de σ relativo a una interpretación contextualmente determinada $i(x)$ (en donde x es la variable contextual) y un mundo w. Consideremos el siguiente ejemplo. Supongamos que relativo a la interpretación i, "Quine" refiere a Quine en el mundo w. Ahora supongamos que en un contexto c queremos interpretar "Quine" como lo hace Miguel (en nuestro ejemplo Miguel piensa que "Quine" refiere a Carlos). Entonces, en este ejemplo en particular, tenemos que ['Quine']$_{i,w}$ = [Quine]$i(c),w$ = Carlos. Entonces, las comillas mixtas en el primer ejemplo cambian la interpretación convencional de "Quine" y le asignan una interpretación contextualmente sobresaliente (la interpretación que Miguel aplica a "Quine"). Algo similar ocurre en los otros dos ejemplos.

Si esta manera de pensar acerca de las comillas mixtas es correcta, para explicar cómo nos comunicamos con usos de A-C basta con apelar a nuestro conocimiento común de convenciones

lingüísticas (incluida la convención asociada con las comillas mixtas) y nuestro conocimiento del contexto (que hay una interpretación contextualmente sobresaliente de las palabras entre comillas mixtas). Me parece que Davidson no podría objetar al convencionalismo apelando a usos de A-C. Pero los usos de A-C son muy similares a casos de innovación léxica. "Quine", "filtósofo" y "atleta" no están siendo interpretadas de acuerdo con su significado convencional y aun así hay comunicación exitosa.

Es plausible pensar que podemos explicar varios casos de innovaciones léxicas apelando a usos de comillas mixtas. Hay un par de cosas que podemos decir para apoyar este punto. La primera que quisiera mencionar es la siguiente. Si reescribimos nuestros ejemplos de innovación léxica utilizando comillas mixtas, obtenemos esencialmente el mismo tipo de ejemplo:

1. El preso se 'haudineó' de la celda.
2. Edu 'tequileó' el ponche.
3. Ese barrio ya se 'romeó'.

Estos parecen seguir siendo casos de innovación léxica que fácilmente generan instancias de comunicación exitosa. En todo caso, la inclusión de las comillas mixtas ayuda a esclarecer lo que ocurre en estos ejemplos. En el caso de "haudineó" tenemos que interpretar esa palabra como el hablante lo propone. Entonces hay una interpretación contextualmente sobresaliente que utilizamos para interpretar la innovación léxica. Pero eso es precisamente lo que ocurre en nuestros usos de comillas mixtas. Algo análogo se puede decir acerca de los otros ejemplos.

A esto se puede objetar que los casos de innovación léxica no necesitan incluir comillas mixtas. Esto es cierto, pero también es cierto que los ejemplos de comillas mixtas no necesitan incluir comillas mixtas. Podríamos reescribir nuestros ejemplos de comillas mixtas sin usarlas, usando en su lugar itálicas o cualquier otra forma de indicar que las palabras en cuestión deben ser interpretadas de manera no estándar. Esto no cambiaría de manera sustancial el carácter de los ejemplos. Mi punto es el siguiente. En ocasiones,

cuando queremos hablar o escribir con mucha precisión, marcamos las comillas mixtas explícitamente. Pero en muchas ocasiones, en lugar de marcar las comillas mixtas de manera explícita, simplemente indicamos con nuestra entonación, itálicas, o de cualquier otra manera, que se trata de un caso de entrecomillado mixto. De hecho, cuando hablamos, a diferencia de cuando escribimos, es mucho menos común que marquemos las comillas mixtas—con nuestros dedos o de alguna otra forma—y, sin embargo, las personas con las que hablamos interpretan nuestra emisión como un caso de comillas mixtas (interpretan la palabra clave en conformidad con una interpretación contextualmente sobresaliente). Creo que es plausible pensar que algo así ocurre en el caso de las llamadas innovaciones léxicas. Si este es el caso, es fácil responder a la objeción de Davidson, porque podemos explicar las instancias de innovaciones léxicas en términos de nuestro conocimiento de las convenciones lingüísticas asociadas con las comillas mixtas y nuestro conocimiento de elementos del contexto (en este caso la interpretación que es sobresaliente en el contexto). Hace falta algo de trabajo para ofrecer una defensa completa de esta propuesta. Por ahora me conformo con presentarla como una opción teórica con cierto grado de plausibilidad.

Hay otra manera de dar cuenta de los casos de innovación léxica. Rita Fernandez (2023) propuso lo siguiente. Si para entender lo que se ha dicho literalmente en ocasiones necesitamos apelar a mecanismos pragmáticos como la desambiguación (y, quizás, el enriquecimiento pragmático (Recanatti, 2010)) no hay razón por la cual no podemos también utilizar mecanismos pragmáticos para interpretar innovaciones léxicas. Podríamos, por ejemplo, pensar en una explicación griceana en donde el punto de partida es que el hablante emitió una oración que no expresa, semánticamente, una proposición completa: porque la innovación léxica (eg.“haudinear”) no cuenta con un significado convencional. Esto, asumiendo que el hablante es cooperativo, va a lanzar a los interlocutores en la búsqueda de un significado para la innovación léxica que esté en conformidad con las máximas conversacionales. Es fácil ver cómo hablantes competentes

pueden lograr esto en nuestros ejemplos de innovaciones léxicas. Davidson podría reclamar que en este caso el significado convencional de la innovación léxica no juega ningún papel en la explicación, porque no hay tal significado convencional, pero no parece que esta preocupación tenga mucho peso: los mecanismos pragmáticos son permitidos para obtener lo que literalmente se dijo y hemos visto cómo el significado convencional del resto de las palabras en la oración sí es parte esencial de la explicación.

Armstrong piensa que la manera correcta de entender los casos de innovación léxica es reconociendo que somos perfectamente capaces de generar convenciones lingüísticas sobre la marcha y que son convenciones que bien pueden solo ser utilizadas una vez. Entonces, el punto de Armstrong es que los casos de innovación léxica nos presentan con problemas de coordinación que resolvemos en el momento y cuyo resultado es una nueva convención lingüística para la palabra relevante. Esta es una forma de responder a la objeción de Davidson ya que, de acuerdo con esta propuesta, sí hay una convención lingüística asociada con la innovación léxica, es sólo que se trata de una convención que surgió en una población que comprende solo las personas involucradas en la conversación. Esta propuesta es ingeniosa, pero no es del todo clara que sea correcta. De acuerdo con Armstrong, contra Lewis, las convenciones no necesitan ser regularidades: una sola instancia basta para generar una convención[16]. Esto puede generar casos contraintuitivos. Supongamos que me encuentro con alguien en un sendero. Vamos caminando en direcciones opuestas, lo cual genera un problema de coordinación. Supongamos que lo solucionamos caminando por nuestro lado izquierdo. Armstrong diría que acabamos de generar una convención de caminar del lado izquierdo. Supongamos ahora que al día siguiente vuelvo a encontrarme con la misma persona, bajo las mismas circunstancias. Esta vez decidimos solucionar el problema de coordinación caminando por nuestro lado derecho. Parecería que ahora tenemos dos convenciones: una convención

16 Schiffer (1972) piensa algo similar.

de caminar del lado derecho y otra de caminar del lado izquierdo. Esto no parece muy plausible.

5.3. ¿Las convenciones deben solucionar problemas de coordinación?

Una objeción común a la propuesta de Lewis consiste en señalar que no todas las convenciones son soluciones a problemas de coordinación recurrentes. Por ejemplo, Schiffer (198.), Millikan (2005), Davis (2003.), Sugden (1986), argumentan que, por ejemplo, las convenciones de vestimenta no resuelven ningún problema de coordinación. Otro ejemplo es el de regalar cigarros cuando una pareja da a luz. Entonces, la objeción sostiene que el análisis de Lewis no puede ser correcto, porque hay convenciones que no son soluciones arbitrarias y atuoperpetuadoras de problemas de coordinación recurrentes. Esta objeción no niega que algunas convenciones solucionan problemas de coordinación, sólo niega que esa sea una característica esencial de las convenciones.

Me parece que este tipo de objeción es simplemente una invitación a participar en un desacuerdo meramente terminológico. Es verdad que con facilidad podemos hablar de convenciones de vestimenta o de la convención de regalar cigarros cuando una pareja da a luz. Pero también es fácil llamarlas tradiciones: parte de los usos y costumbres de una comunidad. Podríamos engancharnos en una discusión terminológica acerca de si vestirse de cierta manera y regalar cigarros bajo ciertas circunstancias son convenciones o tradiciones. En lugar de hacer eso, propongo ofrecer una lectura caritativa de Lewis. No sería una exageración decir que Lewis (1969/1975) estaba principalmente interesado en el tipo de fenómeno que surge cuando una población encuentra una solución arbitraria y auto-perpetuadora a un problema de coordinación recurrente. A esto Lewis llamó "convención". Dudo mucho que el objetivo principal de Lewis haya sido ofrecer un análisis del concepto ordinario de convención: su propósito no era capturar con su teoría todas las formas en que en el habla ordinaria usamos la palabra "convención". Si ocasionalmente llamamos "convenciones"

a regalar cigarros bajo ciertas circunstancias o si vestirse de cierta manera en ciertos eventos, no es de mucho interés teórico para el proyecto de Lewis. Entonces, objetar que esas prácticas también son convenciones no parece particularmente atinado.

Millikan (2005) podría objetar que yo soy quien no está siendo del todo caritativo con su postura. Una de las ideas de Millikan es que la noción teórica correcta de convención es una que captura el mayor número de fenómenos interrelacionados. De acuerdo con ella, las convenciones consisten de dos características: a) son patrones que son reproducidos y b) el hecho de que estos patrones son reproducidos se debe, por lo menos en parte, al peso del precedente. Esta caracterización de las convenciones es perfectamente neutral respecto de si las convenciones solucionan problemas de coordinación. De acuerdo con Millikan, su análisis puede ser usado para dar cuenta tanto de las convenciones lingüísticas como de las prácticas de regalar cigarros, entre otras. Entonces, Millikan podría argumentar que su análisis es superior al de Lewis porque logra capturar más fenómenos interrelacionados.

Podemos conceder que el análisis de Millikan logra encontrar aquello que la práctica de regalar cigarros y la de manejar del lado derecho tienen en común: ambos son patrones reproducidos en donde el peso del precedente es parte de lo que explica su proliferación. Esto no quiere decir que su análisis sea la mejor forma de dar cuenta de la práctica de manejar del lado derecho, o de cualquier otra convención que soluciona problemas de coordinación recurrentes. Los problemas de coordinación presentan situaciones en donde las personas involucradas tienen que actuar de manera deliberada y racional, si quieren encontrarse en una situación de equilibrio. Por lo menos así es como los seres humanos tenemos que confrontar los problemas de coordinación. El análisis de Lewis parece más iluminador acerca de cómo logramos resolver problemas de coordinación recurrentes. Por ejemplo, las cláusulas (3) y (4) son importantes para explicar el carácter estratégico y racional de las soluciones arbitrarias y auto-perpetuadoras de problemas de coordinación recurrente. La cláusula (3) nos habla acerca de las ra-

zones que tenemos para buscar la solución convencional: nos dice que la creencia de que otros se conforman a *R* le da a todos una buena y decisiva razón para conformarse a *R*. (4), por su parte, nos dice que una población prefiere que la mayoría se conforme a *R* y esto está fuertemente ligado con el hecho de que en un problema de coordinación nadie se beneficia si una persona decide actuar de manera distinta al resto. Millikan piensa que, con este tipo de cláusulas, el análisis de Lewis sobre racionaliza las convenciones. Este puede ser el caso si en la mira uno tiene en mente una explicación de nuestras prácticas de regalar cigarros o cómo es que especies no humanas consiguen actuar de manera regular ante ciertos problemas. Para este tipo de casos el análisis de Millikan es más adecuado. Sin embargo, el análisis de Millikan puede parecer un poco pobre, en comparación al de Lewis, cuando se trata de explicar el tipo de fenómeno que a Lewis le interesaba explicar. Mi impresión es que es mejor tratar al tipo de fenómeno que emerge cuando nos confrontamos a problemas de coordinación recurrentes de una manera, y otro tipo de fenómenos, como la práctica de regalar cigarros, o lo que hacen ciertos animales no humanos, de manera distinta.

Espero haber ofrecido buenas razones para pensar que las objeciones que aquí he discutido pueden ser refutadas. Sin embargo, a pesar de ser muy popular, la propuesta de Lewis ha recibido más objeciones de las que he discutido aquí. Para estudiar más críticas al trabajo de Lewis, recomiendo revisar Millikan (2005), Schiffer (1972), Lepore (2015), Hawthorne (1990/1993), Kölbel (1998), y Stotts (2023).

Referencias

Armstrong, J. (2016). Coordination, triangulation, and language use. *Inquiry, 59*(1), 80-112.

Armstrong, J. (2016). The problem of lexical innovation. *Linguistics and Philosophy, 39*, 87-118.

Clark, H. H. (1992). *Arenas of language use*. University of Chicago Press.

Davis, W. (2003). *Meaning, Expression, and Thought*. Cambridge University Press.

Davidson, D. (1986). A Nice Derangement of Epitaphs. En R. Grandy y R. Warner (Eds). *The Philosophical Grounds of Rationality* (pp.157–74). Oxford University Press.

García-Ramírez, E. (s.f) "The proper function of linguistic practice and linguistic meaning". [Manuscrito no publicado]

Hawthorne, J. (1990). A Note on 'Languages and Language. *Australasian Journal of Philosophy*, 68, 116–118.

Hawthorne, J. (1993). Meaning and Evidence: A Reply to Lewis. *Australasian Journal of Philosophy*, 71, 206–211.

Kölbel, M. (1998). Lewis, Language, Lust, and Lies. *Inquiry*, 41, 301–315.

Lepore, E. y Stone, M (2015). David Lewis on Convention. En B. Loewer y J. Schaffer (Eds.), *A Companion to David Lewis* (pp. 315–27). Wiley.

Lewis, D. (1969). *Convention*. Harvard University Press

Lewis, D. (1975/1983). Languages and Language. Reprinted in *Philosophical Papers*, vol. 1. Oxford University Press.

Maier, E. (2014). Mixed quotation: The grammar of apparently transparent opacity. *Semantics and Pragmatics*, 7, 1–7.

Millikan, R. (2005). *Language: A Biological Model*. Clarendon Press.

Quine, W.V.O. (1936). Truth by Convention. En A. N. Whitehead & O. H. Lee (Eds.), Philosophical essays (pp. 90–124). Russel & Russel.

Russell (1950). *Análisis del espíritu*. Paidós.

Recanati, F. (2010). *Truth-Conditional Pragmatics*. Oxford University Press.

Recanati, F. (2001). Open quotation. *Mind*, 110(439), 637–687.

Roberts, C. (1996). Information Structure in Discourse: Towards an Integrated Formal Theory of Pragmatics. *Semantics & Pragmatics*, 5 (6),1-69.

Roberts, C. (2012). Information structure: Afterword. *Semantics and Pragmatics*, 5 (7),1-19.

Schelling, T. (1960). *The Strategy of Conflict*. Harvard University Press.

Schiffer, S. (1972). *Meaning*. Oxford University Press.

Skyrms, B. (1996). *Evolution of the Social Contract*. Cambridge University Press.

Stotts, M. H. (2023). Conventions without knowledge of conformity. *Philosophical Studies, 180*(7), 2105-2127.

Sugden, Robert. (1986). *The Economics of Rights, Co-operation, and Welfare* (2nd ed). Palgrave Macmillan.

Shan, C.-C. (2010). The character of quotation. *Linguistics and Philosophy*, 33(5), 417–443

Williams, B. (2002). *Truth & Truthfulness: An Essay in Genealogy*. Princeton University Press.

CAPÍTULO III

EL CONCEPTO DE VERDAD EN LA TRADICIÓN ANALÍTICA. UNA MIRADA A LAS TEORÍAS PRINCIPALES

Mario Gómez Torrente[1]

RESUMEN

Se exponen en primer lugar algunas consideraciones introductorias sobre el concepto ordinario de verdad, los portadores de la propiedad de la verdad, y la paradoja del mentiroso. A continua-

1 Mario Gómez Torrente (Lic. Fil., Barcelona, 1990; Ph.D. in Philosophy, Princeton, 1996) es investigador en el Instituto de Investigaciones Filosóficas de la Universidad Nacional Autónoma de México (UNAM) y trabaja principalmente sobre temas de filosofía de la lógica, del lenguaje y de la mente. Entre sus publicaciones recientes destacan los artículos "Perceptual Variation, Color Language, and Reference Fixing. An Objectivist Account" (Nous, 2016), "Modal Realism and Anthropic Reasoning" (Australasian Journal of Philosophy, 2024), "Rigidity and Necessary Application" (Nous, 2025), y el libro Roads to Reference. An Essay on Reference Fixing in Natural Language (Oxford U.P., 2019)

ción, se describen someramente algunas teorías tradicionales sobre la verdad, como la coherentista, la pragmatista, y algunas variantes de la correspondentista, estableciendo la distinción entre teorías correspondentistas sustantivistas e insustantivistas. El resto del artículo está dedicado a presentar de manera crítica las dos teorías más importantes de la verdad en la tradición analítica, las teorías insustantivistas de Alfred Tarski (1933) y Saul Kripke (1975). En el caso de la teoría de Tarski, ésta se introduce presentando una definición tarskiana de un predicado de verdad para un lenguaje formal específico relativamente simple. Seguidamente se exponen las razones por las que esta definición es extensionalmente correcta para el lenguaje para el que se define, y cómo evade paradojas como la del mentiroso, en virtud del hecho de que el predicado definido no lo está para oraciones del metalenguaje en que se han dado las definiciones, de manera que el lenguaje al que se aplican no tiene su propio predicado de verdad. También se presentan críticas a la teoría tarskiana derivadas sobre todo de este último hecho. Nuestra presentación de la teoría de Kripke enfatiza este tipo de limitaciones, algunas de las cuales son superadas por esta teoría. La teoría de Kripke se presenta exponiendo su construcción de una interpretación consistente de un lenguaje que contiene su propio predicado de verdad, en la que este predicado no está definido para oraciones "no fundadas", incluidas oraciones como las que dan pie a las distintas versiones de la paradoja del mentiroso. Finalmente se describe una aparente limitación de la teoría kripkeana relacionada con la llamada "paradoja del mentiroso reforzado".

1. Preliminares sobre el concepto de verdad

¿Qué es la verdad? Hay una larga tradición de respuestas a esta pregunta filosófica. En este trabajo describiremos brevemente algunas de las respuestas ofrecidas dentro de la tradición analítica, con especial atención a dos de ellas, sin duda las más influyentes e importantes, la ofrecida por la llamada "teoría semántica de la verdad" de Alfred Tarski y la ofrecida por la teoría de Saul Kripke. Pero antes distinguiremos la pregunta (y los tipos de respuestas que admite)

de otras preguntas relacionadas con ella, y trataremos algunas otras cuestiones preliminares.

No es lo mismo preguntar qué es la verdad que preguntar cuál es la verdad (acerca de una determinada cuestión). Si preguntamos cuál es la verdad acerca de la muerte de Luis Donaldo Colosio no estamos haciendo una pregunta filosófica. Estamos preguntando, quizá, si la verdad es que lo mató un asesino solitario o que fue la víctima de una conspiración política. Si preguntamos qué es la verdad, en cambio, estamos preguntando cómo caracterizar la noción de verdad, cómo explicar esa noción de tal manera que el concepto de verdad quede delimitado, y esta es una pregunta típicamente filosófica. Además, saber qué caracteriza a la verdad no necesariamente nos permitirá averiguar cuál es la verdad acerca de una determinada cuestión: por ejemplo, como mencionaremos luego, algunos filósofos han propuesto que lo que caracteriza a la verdad es la correspondencia con los hechos; pero aceptar esta caracterización no nos ayuda mucho a saber cuál es la verdad acerca de la muerte de Luis Donaldo Colosio—para ello tendríamos que saber cuál de las hipótesis mencionadas, o de otras, se corresponde con los hechos.

No es lo mismo preguntar qué es la verdad que preguntar de qué se dice la verdad—o su contrario, la falsedad. Al preguntar de qué se dice la verdad y la falsedad estamos haciendo una pregunta filosófica, ciertamente. Pero no es la pregunta acerca de cómo caracterizar la noción de verdad, sino una pregunta preliminar a la tarea de caracterizar la noción de verdad. Al preguntar de qué se dice la verdad y la falsedad preguntamos qué tipo de cosas son susceptibles de ser verdades y falsedades.

Hay muchos tipos de cosas de las que decimos que son verdades o falsedades, verdaderas o falsas: ideas, creencias, conjeturas, afirmaciones, proposiciones, oraciones, etc. Este último tipo de cosas, las oraciones, han recibido una atención especial por parte de los filósofos analíticos que han estudiado la noción de verdad. Es importante hablar brevemente del sentido en que se dice de una oración que es verdadera o falsa. La idea fundamental que hay que

apreciar es que una oración no es nunca verdadera o falsa *por sí sola*. Considérese la oración

(1.1) Ella tiene agujetas.

La oración (1.1) no es verdadera o falsa por sí sola, independientemente de quién sea la persona a la que nos estemos refiriendo con el pronombre personal 'ella'. En algunos casos, al usar (1.1) nos estaremos refiriendo a una mujer que tiene agujetas, en otros nos referiremos a una que no tiene, o quizá no nos estaremos refiriendo a nadie, quizá porque estamos pronunciando (1.1) por el mero placer de oír cómo suenan las palabras.

Este fenómeno es más general, no tiene que ver exclusivamente con la aparición en las oraciones de pronombres y otras palabras cuyo contenido no es "permanente" y depende fundamentalmente de las intenciones comunicativas explícitas de los hablantes en las distintas ocasiones de uso. Otra manera en que (1.1) no es verdadera o falsa por sí sola tiene que ver con el hecho de que el contenido de un signo gráfico o sonoro cualquiera es arbitrario, lo cual se refleja a veces en el fenómeno de que una misma grafía o sonido tiene de hecho contenidos diferentes en idiomas o dialectos diferentes. Por ejemplo, en el español de España 'agujetas' significa aproximadamente "dolores en un músculo que no se ejercitaba y se acaba de ejercitar intensamente", mientras que en el español de México significa "cordones de las zapatillas". Incluso si ha quedado claro que con 'ella' en un uso de (1.1) nos estamos refiriendo a una determinada mujer, puede que en ese uso la oración (1.1) sea verdadera en un sentido de 'agujetas' y falso en el otro.

De manera completamente general, lo que estos ejemplos parecen mostrar es que una oración no es verdadera o falsa por sí sola, sino únicamente en cuanto que expresa un contenido, a veces llamado proposición, verdadero o falso por sí solo, que queda fijado por las intenciones comunicativas y asignaciones arbitrarias de contenidos a signos gráficos, sonoros o de otro tipo. Sin embargo, a menudo decimos en contextos ordinarios y en filosofía que tal o cual oración

es verdadera o falsa, pero sólo porque se sobreentiende cuál es el contenido de los signos de la oración. Algunas teorías filosóficas de la noción de verdad, y en particular las teorías de Tarski y Kripke, han sido formuladas dando por supuesto que la verdad o falsedad se dice de oraciones cuyo contenido se ha fijado previamente.

Consideremos la siguiente oración:

(1.2) (1.2) es falsa.

Haciendo las suposiciones obvias acerca del contenido de los signos de (1.2), (1.2) parece expresar la proposición de que la oración (1.2) es falsa. ¿Es esa proposición verdadera, o es falsa? Parece que un principio obvio acerca de la noción de verdad es que una oración es verdadera si y sólo si lo que dice es el caso, de manera que lo siguiente es una oración verdadera:

(1.3) '(1.2) es falsa' es verdadera si y sólo si (1.2) es falsa.

Pero entonces, por la implicación de izquierda a derecha, si (1.2) es verdadera, entonces es falsa y por tanto no es verdadera, lo cual es una contradicción. Por otro lado, por la implicación de derecha a izquierda, si (1.2) es falsa entonces es verdadera, y por tanto no es falsa; de nuevo una contradicción. Tanto si suponemos que (1.2) es verdadera como si suponemos que es falsa llegamos a una contradicción, y como no hay otras opciones (por el principio de tercio excluso) parece que hemos llegado a una contradicción incondicional, pura y simple. A esta contradicción se la conoce desde la antigüedad como la "antinomia del mentiroso" o "paradoja del mentiroso". No es claro todavía, aun después de que hayan pasado miles de años desde que fuera formulada por primera vez, cuál es la razón exacta de que surja la paradoja ni cuál es su solución. A veces, aunque ni mucho menos siempre, las teorías sobre qué es la verdad han sido motivadas en parte por la paradoja del mentiroso; ejemplos de ello son las teorías de Tarski y Kripke, que describiremos más abajo.

Un posible diagnóstico del problema de la paradoja del mentiroso es que la oración (1.2) no expresa realmente un contenido o

proposición, pues al fin y al cabo es sólo una oración, y por tanto no es realmente verdadera o falsa en sí misma. (Aunque a muchos les parece claro que (1.2) expresa una proposición, que es por tanto verdadera o falsa.) Así se bloquea el dilema del razonamiento de la paradoja. La pregunta inmediata es entonces si puede reproducirse una paradoja similar para proposiciones, no para oraciones. El intento empezaría seguramente considerando una oración como la siguiente:

(1.2') (1.2') expresa (en español) una proposición falsa.

¿Es la proposición presumiblemente expresada por la oración (1.2') verdadera, o falsa? Podríamos razonar así: si esa proposición es verdadera entonces esa proposición es falsa, lo cual es absurdo. Pero si esa proposición es falsa entonces es verdadera, lo cual es absurdo también. Es posible proponer que (1.2') no expresa una proposición, con lo que el nuevo razonamiento queda también bloqueado; pero queda la cuestión de cómo justificar que (1.2') no expresa una proposición. Más adelante veremos que la teoría de Kripke puede verse como proporcionando una justificación de la sugerencia de que las oraciones problemáticas no expresan proposiciones.

2. Teorías metafísicas y epistemológicas del concepto de verdad

Hay fundamentalmente dos tipos de maneras en que se ha querido caracterizar la noción de verdad en la tradición filosófica. Uno de ellos emplea nociones metafísicas, el otro emplea nociones epistemológicas.

Ya hemos mencionado la idea básica del primer tipo de caracterizaciones: la idea de que una proposición es verdadera cuando se corresponde con un hecho. Se trata de una idea asociada al nombre de Aristóteles (en virtud de su famoso aunque críptico *dictum* según el cual la verdad consiste en decir de lo que es que es, y de lo que no es que no es (*Metafísica* Γ, 1011b25)), así como a la tradición escolástica y específicamente tomista inspirada por

el Estagirita (recuérdese la famosa definición de Santo Tomás de la verdad como "la adecuación entre el intelecto y la cosa" (*Suma Teológica*, I, q. 16, a. 1)). La oposición principal a esta idea ha provenido generalmente de filósofos reacios a aceptar el valor explicativo o cognoscitivo de nociones percibidas como metafísicas. En particular, las nociones de correspondencia y de hecho han suscitado esta sospecha. A veces se ha sostenido, por ejemplo, que la noción de correspondencia es excesivamente oscura, en parte por la oscuridad misma de la noción de proposición, en parte por la de la relación entre las proposiciones verdaderas y los hechos que les corresponden. Otra objeción es que no tenemos una concepción de la noción de hecho según la cual todas las proposiciones que llamaríamos verdaderas se correspondan con un "hecho"; quizá—se ha dicho—tenemos una concepción de la noción de hecho según la cual es un hecho que la Torre Eiffel tiene 300 metros de altura, pero no según la cual haya un hecho que corresponda a la proposición verdadera de que 17 es un número primo, o a la proposición verdadera de que Cleopatra fue bella. (Las sospechas clásicas sobre las nociones de correspondencia y hecho pueden verse en los positivistas lógicos como Neurath y Carnap; véase el resumen de sus ideas al respecto en Hempel (1935).) Más adelante veremos que los proponentes de las teorías redundantista y semántica de la verdad han visto a veces sus propuestas como versiones no metafísicas de la idea correspondentista.

Una tendencia general de la modernidad fue la progresiva "epistemización" de los problemas metafísicos. Esa tendencia se manifestó también en el ámbito de la reflexión filosófica sobre la noción de verdad, con la aparición de caracterizaciones dadas en términos de nociones epistemológicas. Algunos de los filósofos que criticaron la naturaleza metafísica de la teoría correspondentista fueron también responsables de la formulación de una de las teorías epistemológicas de la verdad más conocidas, la llamada teoría coherentista. En nuestra descripción de esta teoría nos basaremos fundamentalmente en la versión que de ella dieron algunos positivistas lógicos. (Véase Hempel (1935).) Pero otras versiones

de la idea son rastreables seguramente a filósofos de las tradiciones empirista e idealista. La idea fundamental puede enunciarse así: una proposición es verdadera cuando es "coherente" con un grupo de otras proposiciones que se delimitan de una manera especial; este grupo de proposiciones especiales puede ser, por ejemplo, el de las proposiciones acerca de "datos elementales de la experiencia" que los seres humanos están dispuestos a aceptar—proposiciones como la de que veo una mancha roja en el centro de mi campo visual. ¿Qué quiere decir 'coherencia'? Dar una respuesta satisfactoria a esta pregunta es uno de los problemas básicos del filósofo con inclinaciones coherentistas. La idea intuitiva es, sin embargo, que son verdaderas las proposiciones no refutables por algún tipo de verdades acerca de "datos elementales de la experiencia". Una dificultad típica de estas teorías es que, dados ciertos supuestos bastante naturales, ninguna proposición excepto las muy básicas es refutable por las verdades acerca de "datos elementales de la experiencia". Es un hecho generalmente admitido que cualquier proposición mínimamente desligada de la experiencia inmediata es compatible con esas verdades, supuesto que uno esté dispuesto a llamar verdaderas a un sistema apropiado de proposiciones "coherentes" con aquella proposición y con las verdades acerca de "datos elementales de la experiencia". Y sin embargo parece que debería haber proposiciones que no sean básicas en este sentido pero que sean falsas en un sentido absoluto.

Otro tipo de caracterizaciones de la noción de verdad en términos de nociones epistemológicas son las caracterizaciones que se suele llamar "pragmatistas", asociadas a los filósofos norteamericanos de la escuela del mismo nombre. (Véase James (1906).) La idea general de estas caracterizaciones es que una proposición verdadera es una que los seres humanos, dadas ciertas condiciones más o menos idealizadas de su desempeño cognoscitivo, llegarían a creer o aceptar en un estadio más o menos idealizado de la historia cognoscitiva humana. Quizá no sea irrazonable admitir que todas las proposiciones con esta propiedad (suponiendo que la propiedad en cuestión haya quedado unívocamente caracterizada meramente

con lo recién dicho) serán verdaderas en un sentido intuitivo—si bien no es en absoluto claro que no vaya a haber alguna proposición falsa que los seres humanos estén condenados a aceptar a causa de la naturaleza de los mecanismos cognoscitivos de que disponen. Pero es incluso más discutible que todas las proposiciones verdaderas tengan la propiedad enunciada por el pragmatista; es razonable pensar que muchas verdades están fuera del alcance cognoscitivo de los seres humanos.

Desde luego, esto es sólo el inicio del debate acerca de las teorías coherentista y pragmatista. Los proponentes de este tipo de teorías disponen de argumentos sofisticados para defenderse de este tipo de objeciones elementales, argumentos que ocasionalmente pretenden establecer la incoherencia de la noción intuitiva de verdad en que esas objeciones se apoyan.

3. La distinción entre sustantivismo e insustantivismo

Algunos filósofos han visto en la teoría correspondentista el germen de una teoría de la verdad en términos de conceptos no metafísicos, y por tanto menos objetable que la teoría correspondentista. Una idea que los ha guiado a menudo ha sido la siguiente. En el uso más típico, uno usa las expresiones 'es verdad' y 'es verdadera' cuando uno dice de una cierta oración u oraciones que son verdad o que son verdaderas, y cabe pensar que en esos casos uno en cierto sentido sólo está haciendo de una manera algo más enfática lo que podría hacer afirmando directamente esa oración u oraciones. A esta idea, explorada quizá por primera vez por F. P. Ramsey (véanse Ramsey (1927) y Strawson (1950)), se la conoce como "la teoría redundantista" de la verdad. Por ejemplo, según la idea redundantista, cuando uno afirma

(3.1) 'Los gatos ronronean' es verdadera

uno en algún sentido no está diciendo sino lo mismo que podría decir afirmando

(3.1') Los gatos ronronean.

La idea redundantista implica la deseable conclusión de que los bicondicionales de esta forma expresan algún tipo de equivalencia fuerte (quizá incluso una equivalencia conceptual o analítica en el caso de (3.2b)):

(3.2a) 'Los gatos ronronean' es verdadera si y sólo si los gatos ronronean;
(3.2b) Que los gatos ronronean es verdad si y sólo si los gatos ronronean.

Pero debe notarse que aceptar estas equivalencias es compatible con la aceptación de otro tipo de teorías. Lo que sí parece ser una consecuencia peculiar de la teoría redundantista es que las expresiones 'es verdad' y 'es verdadera' carecen de una función "representativa", no están por una propiedad genuina de las oraciones, sino que se usan meramente de una forma redundante, quizá para añadir cierto énfasis a la afirmación de oraciones particulares. Así, la teoría redundantista tiene la consecuencia de que la noción de verdad no es tanto una noción de la que haya que dar una caracterización como una noción extremadamente tenue, que no es susceptible de caracterización más que en un sentido trivial.

La conexión con la teoría correspondentista se enfatiza a veces por quienes añaden a la idea redundantista la opinión de que 'es verdad' o 'es verdadera' son expresiones equivalentes en algún sentido fuerte a las expresiones 'se corresponde con los hechos' o incluso 'es un hecho', al menos en ciertos sentidos de estas últimas expresiones. (Véase por ejemplo Austin (1950).) No es irrazonable pensar que pueda darse una equivalencia fuerte de este tipo, quizá incluso analítica—mientras que sería difícil sostener que 'es verdad' o 'es verdadera' son analíticamente equivalentes a 'es coherente con un grupo de otras oraciones (o proposiciones)' o 'es tal que los seres humanos, dadas ciertas condiciones más o menos idealizadas de su desempeño cognoscitivo, llegarían a creerla o aceptarla en un estadio más o menos idealizado de la historia

cognoscitiva humana'. De todos modos, parece más o menos claro que si uno sostiene la idea redundantista y al mismo tiempo sostiene que 'es verdad' o 'es verdadera' son expresiones analíticamente equivalentes a las expresiones 'se corresponde con los hechos' o 'es un hecho', esto sólo puede ser así en algún sentido especialmente débil, probablemente redundantista, de estas últimas expresiones—pues al afirmar una oración particular, sin usar concepto alguno de verdad, uno no parece estar predicando ninguna propiedad de la oración.

Pero la idea redundantista parece claramente objetable. Quizá la idea resulta plausible en el caso de (3.1) y (3.1'), pero no lo es en otros casos. Si yo afirmo

(3.3) La primera oración que pronunció Sor Juana Inés de la Cruz es verdadera

no estoy por ello haciendo algo que podría hacer afirmando la oración de la que hablo—no tengo ni idea de cuál pueda ser. De la misma manera, si afirmo

(3.4) Todos los teoremas demostrados por los matemáticos son verdaderos

no estoy por ello haciendo algo que podría hacer afirmando individualmente todas las oraciones matemáticas de las que hablo—hacerlo sería simplemente imposible en la práctica.

Una variante de la teoría redundantista es la teoría "pro-oracional" (véase Williams (1995) para una introducción en español), según la cual 'es verdad' y 'es verdadera' igualmente carecen de una función "representativa", no son realmente predicados que expresan una cierta propiedad de las oraciones, pero no son estrictamente redundantes lingüísticamente, pues cumplen una función no predicativa de crear una especie de pronombres especializados para funcionar como oraciones ("pro-oraciones"), sin necesidad de formularlas explícitamente. Un pro-oracionalista sostendría que (3.3) es una pro-oración que de alguna manera sustituye a la pri-

mera oración que pronunció Sor Juana Inés de la Cruz, incluso si la hablante no tiene idea de cuál fue. Esto puede parecer ya un poco extraño, en la medida en que detecta en el lenguaje un recurso para afirmar oraciones sin pronunciarlas que no parece parafraseable en términos independientes del aparente predicado de verdad. El caso de (3.4) es aún más sospechoso. El pro-oracionalista sostiene que la afirmación "real" que abrevia (3.4) es algo que podemos entender así:

(3.4') Para todo *p*, si *p* es un teorema demostrado por los matemáticos, *p*.

Así, para ofrecer una explicación de este tipo de generalizaciones sobre oraciones que usan los predicados 'es verdad' y 'es verdadera', el pro-oracionalista emplea la llamada "cuantificación sustitucional" sobre oraciones, y debe postular que esos aparentes predicados a veces se usan para crear no pro-oraciones estrictamente hablando, sino variables pro-oracionales. De nuevo ni la cuantificación sustitucional ni las variables pro-oracionales parecen ser recursos que existan de manera natural en el lenguaje ordinario, independientemente de postulaciones teóricas como las del pro-oracionalista. Un problema de este tipo de explicación de los predicados 'es verdad' y 'es verdadera' es, por tanto, que parece postular niveles de complicación en la gramática y la sintaxis "profundas" del lenguaje que no parecen tener un apoyo natural independiente.

En vista de estas objeciones, no resulta muy atractivo sostener que las expresiones 'es verdad' y 'es verdadera' tienen una función enfática y realmente redundante, o que su función es la de crear pro-oraciones y variables pro-oracionales disfrazadas. Y los ejemplos también parecen hacer difícil usar las motivaciones redundantista o pro-oracionalista para sostener la tesis de que esas expresiones no expresan una propiedad de un cierto tipo de oraciones. Pero autores no redundantistas y no pro-oracionalistas han buscado otras maneras de defender una tesis más débil, a saber, la tesis de que 'es verdad' y 'es verdadera' sí expresan una propiedad, pero no una propiedad

"metafísica" o "sustantiva" de las oraciones (o proposiciones) que la poseen, sino una propiedad "insustantiva". Es por ello que a las teorías de la verdad descritas en la sección anterior se las conoce como "sustantivistas", mientras que a las teorías en cuestión se las conoce como "insustantivistas" (*deflationary* en inglés). (En general, se habla también de las teorías redundantista y pro-oracionalista como teorías insustantivistas, por razones obvias.)

Es razonable ver a las teorías kripkeana y tarskiana de la verdad, que nos ocuparán a continuación, como teorías insustantivistas de la verdad. Esta característica, que se apreciará cuando describamos las teorías y veamos el tipo de nociones en términos de las cuales delimitan el ámbito de las oraciones verdaderas, explica en buena medida la preeminencia de estas teorías en la tradición analítica. Otras virtudes que explican esta preeminencia son (i) el no verse afectadas por las dificultades generadas para el redundantismo y el pro-oracionalismo por oraciones como (3.3) y (3.4), (ii) el ofrecer un cierto tipo de "soluciones" para la paradoja del mentiroso y, fundamentalmente, (iii) el proporcionar caracterizaciones puramente matemáticas de las nociones de oración verdadera y de satisfacción. (En el caso de la teoría tarskiana, hay un rasgo adicional que explica su preeminencia especial, incluso por encima de la de la teoría kripkeana, y en particular su uso como piedra angular para desarrollar de una manera rigurosa la teoría matemática clásica de los modelos: la teoría tarskiana respeta el postulado clásico de que todas las oraciones de un lenguaje han de tener un valor de verdad.) Veremos por qué las teorías tarskiana y kripkeana tienen estas características (i) a (iii), y también en qué sentido delimitan el ámbito de las proposiciones verdaderas. Pero antes habremos de describir las teorías.

4. Las definiciones tarskianas de la verdad

La teoría semántica de la verdad fue formulada por A. Tarski a finales de los años 20 y principios de los años 30 del siglo XX. (Véanse Tarski (1944) para una exposición del propio Tarski en español, y Gómez Torrente (2001a) y (2001b) para una exposición y discu-

sión histórica y filosófica de la presentación original de Tarski en los años 30; véase también Barrio (2014b) para otra exposición en español.) Lo que hizo Tarski fue ofrecer un *método* para dar *definiciones* de predicados de verdad para lenguajes formales particulares a cuyas expresiones se les han asignado significados concretos. (La restricción a lenguajes formales no es infranqueable; no hay obstáculos de principio a la formulación de definiciones tarskianas para lenguajes no formales, con tal que su sintaxis se pueda especificar de una manera más o menos precisa.) Tarski no dio una definición de la noción de verdad para proposiciones en general, ni siquiera dio una definición de un predicado único de verdad para oraciones en general, ni siquiera de un predicado relacional de la forma '*O* es verdadera en *L*' que sea satisfecho por pares oración-lenguaje. Dado un lenguaje formal particular con significado *L*, el método de Tarski nos indica meramente cómo dar una definición de un predicado monádico 'es *OV* en *L*' que se aplica intuitivamente precisamente a las oraciones verdaderas de *L*.

Nótese que al decir que 'es *OV* en *L*' se aplicará intuitivamente precisamente a las oraciones verdaderas de *L* no hemos dicho que 'es *OV* en *L*' vaya a ser equivalente en un sentido fuerte a 'es verdadera' o siquiera a 'es verdadera en *L*', mucho menos que vaya a ser conceptual o analíticamente equivalente a alguna de estas expresiones. La manera como la teoría tarskiana de la verdad delimita el ámbito de la verdad es relativamente poco ambiciosa: la teoría da meramente un método para construir un predicado 'es *OV* en *L*' que es *coextensional* con (es decir, que se aplica exactamente a las mismas cosas que) el predicado intuitivo 'es verdadera en *L*'.

La definición de 'es *OV* en *L*' se formula en un *metalenguaje* apropiado para *L*, es decir, en un lenguaje por medio del cual se pueden decir suficientes cosas *acerca de L*. En un metalenguaje tarskiano para *L*, y por tanto en el *definiens* de la definición de 'es *OV* en *L*', aparecerán sólo tres tipos de términos: (a) términos de la sintaxis de *L*, en particular expresiones para nombrar a todas las expresiones de *L*, y expresiones que denoten ciertas operaciones definidas sobre expresiones de *L*, como la operación de concatenación de expresiones

de *L*; (b) términos lógicos, conjuntísticos, y matemáticos en general: símbolos para conectivas y cuantificadores y recursos para hablar de nociones como pertenencia, inclusión, funciones, secuencias finitas e infinitas, cardinalidad de conjuntos, etc.; (c) por último, deberán aparecer también los términos específicos de *L* o traducciones suyas al metalenguaje. La definición de 'es *OV* en *L*' contendrá exclusivamente términos de los grupos (a), (b) y (c). Es por esta razón que la teoría de Tarski se ha visto a menudo como una teoría insustantivista de la verdad, pues los términos de los grupos (a), (b) y (c) no parecen nombrar propiedades que hayan parecido filosóficamente "sustantivas" o "metafísicas" (salvo si hay términos "sustantivos" en el grupo (c), y por tanto en *L*).

Dar una formulación general y abstracta de su método para definir el predicado 'es *OV* en *L*' en el metalenguaje apropiado para un lenguaje cualquiera L de los que Tarski tiene en mente sería una tarea algo ardua y, lo que es peor, la descripción resultante no sería muy iluminadora. Por ello, Tarski mismo ilustró su método describiendo su aplicación a un lenguaje particular especialmente simple. Nosotros haremos lo mismo, y de hecho el lenguaje que nos servirá de ejemplo, al que llamaremos *L1*, será estructuralmente muy parecido al usado por Tarski. *L1* es un lenguaje de primer orden; en lo que sigue, daremos por supuesta la familiaridad del lector y la lectora con este tipo de lenguajes formales, así como con ciertas nociones conjuntísticas elementales.

Los signos primitivos de *L1* son el cuantificador universal ('∀'), el signo de conjunción ('∧'), el signo de negación ('¬'), paréntesis ('(', ')'), un predicado diádico ('A'), la letra 'x', y un acento subíndice (',') para generar un número infinito de variables por posposición a 'x' (para abreviar, usaremos expresiones de la forma 'x_n' para la variable en que la letra 'x' va seguida de *n* acentos subíndices). El recorrido de las variables en la interpretación deseada de *L1* es el conjunto de todas las personas. El significado deseado de 'A' es la relación que se da entre dos personas cuando la primera ama a la segunda. Las interpretaciones de los otros signos lógicos son las normales. Las fórmulas atómicas son las de la forma 'Ax_kx_1'. Las

fórmulas complejas se obtienen por las operaciones de negación, conjunción (rodeada por paréntesis) y cuantificación universal con respecto a las variables.

En el metalenguaje para *L1* hay, como dijimos, nombres para todas las expresiones de *L1* (es decir, para las series finitas formadas por signos primitivos de *L1*). Esto se consigue incluyendo en ese metalenguaje nombres para los signos primitivos y obteniendo nombres para las otras expresiones por aplicaciones sucesivas de la función binaria de concatenación. Por ejemplo, nos podemos referir a la expresión '$\forall x, A x, x,$' en el metalenguaje tarskiano para *L1* por medio de la expresión 'la expresión formada concatenando el cuantificador universal con la primera variable con la letra a mayúscula con la primera variable con la primera variable, en ese orden'. Usaremos aquí el conocido artificio de las semicomillas debido a Quine para abreviar expresiones de este tipo. (Véase por ejemplo Gómez Torrente (2000), Apéndice.) Con este artificio podemos abreviar, por ejemplo, la expresión 'el cuantificador universal' por medio de la expresión '$\ulcorner\forall\urcorner$'; la expresión 'la expresión formada concatenando el cuantificador universal con la primera variable con la letra a mayúscula con la primera variable con la primera variable, en ese orden', por medio de la expresión '$\ulcorner\forall x, A x, x,\urcorner$', y la expresión 'la expresión formada concatenando el cuantificador universal con la *n*-ésima variable con la letra a mayúscula con la *n*-ésima variable con la *n*-ésima variable, en ese orden', por medio de la expresión '$\ulcorner\forall x_n A x_n x_n\urcorner$'.

Es importante subrayar la idea básica en virtud de la cual el predicado 'es *OV* en *L1*', una vez definido, será coextensional con el predicado intuitivo de verdad para oraciones de *L1*. Para ello hay que enunciar la famosa 'convención T' de Tarski. (La 'T' es porque 'truth' empieza por 't', y el inglés es el idioma en que más se ha hablado de la convención. En el polaco de los primeros textos de Tarski la convención lleva el nombre 'P', en el alemán el nombre 'W' y en castellano estrictamente hablando debería llevar el nombre 'V'.) La convención T es una definición enunciable en el *metametalenguaje* de *L1* (similares definiciones, o "convenciones T" se

enunciarán en los metametalenguajes de otros lenguajes para los que queramos definir un predicado coextensional con el predicado intuitivo de verdad), y que intuitivamente da una condición suficiente para que el predicado 'es *OV* en *L1*' sea coextensional con el predicado intuitivo de verdad para oraciones de *L1*. La convención T para *L1*, por ejemplo, dirá que

> (T) una definición formalmente correcta de un símbolo 'es *OV* en *L1*' es una *definición adecuada de la verdad* si y sólo si tiene como consecuencias todas las oraciones que se obtienen a partir de la expresión '*x* es *OV* en *L1* si y sólo si *p*' reemplazando el símbolo '*x*' por un nombre "por concatenación" de una oración cualquiera de *L1* y el símbolo '*p*' por la expresión que es la traducción de aquella oración al metalenguaje.

(Nótese que los bicondicionales de que habla la convención T son similares a los bicondicionales de la forma de (3.2a).)

¿Cuál es el motivo por el que podemos pensar que la convención T enuncia una condición suficiente para que 'es *OV* en *L1*' sea coextensional con el predicado intuitivo de verdad para las oraciones de *L1*? El motivo lo da un argumento intuitivo como el siguiente. Abreviemos por medio de '*OV*' un predicado definido 'es *OV* en *L1*' cuya definición satisface la convención T. Supongamos primero que *p* es una oración de *L1* a la que se aplica el predicado '*OV*', independientemente de si la oración metalingüística que dice que '*OV*' se aplica a *p* es demostrable; sea *P* esta oración metalingüística. Sabemos que el bicondicional entre *P* y *p* es demostrable (es una consecuencia de una simple definición), y por tanto que es verdadero; así, por nuestro supuesto, podemos afirmar *p*, y por tanto que *p* es verdadera en el sentido intuitivo. Por otro lado, supongamos que *p* es verdadera en el sentido intuitivo; entonces podemos afirmar *p*, y así, por el mismo razonamiento que antes, afirmar *P*, y que '*OV*' se aplica a *p*.

Este razonamiento intuitivo no aparece en Tarski, y sin duda no lo habría considerado un razonamiento matemáticamente satisfactorio. Pero da una idea del tipo de consideraciones informales que sin duda justificaron para él el desiderátum de que una definición

del predicado deseado 'es *OV* en *L1*' cumpliera con la convención T. Lo que estamos buscando puede resumirse, pues, así: buscamos un predicado definible con ayuda exclusivamente de los términos de los grupos (a), (b) y (c) del metalenguaje y que cumpla con la convención T (y del que pueda demostrarse que lo hace).

Tarski menciona que la idea que surge inmediatamente es la de dar una "definición recursiva" del predicado 'es *OV* en *L1*'. Si esto es posible, la manera de hacerlo sería dar condiciones para la verdad de un cierto tipo de oraciones básicas, luego dar condiciones para la verdad de oraciones complejas en términos de las condiciones para la verdad de oraciones menos complejas, y por último definir el predicado 'es *OV* en *L1*' como aplicándose a aquellas oraciones que o son del tipo básico o son complejas pero la verdad "les ha sido transmitida" por la verdad o falsedad de las menos complejas. La idea se ve más claramente para un lenguaje aún más simple que *L1*. Digamos que *L2* es el lenguaje con signos primitivos '=', '¬', '0', '1', '2', '3', '4', '5', '6', '7', '8' y '9', cuyas fórmulas atómicas son las de la forma '*n*=*m*' (con *n* y *m* numerales decimales), y cuyas fórmulas complejas se forman anteponiendo '¬' a fórmulas menos complejas. Entonces podemos definir un predicado 'es *OV* en *L2*' apropiado para *L2* de la siguiente manera:

> una oración de *L2* es *OV* en *L2* si y sólo si (α) es de la forma '*n*=*n*' (este es el tipo de oraciones verdaderas "básicas") o (β) es de la forma '¬*O*' y *O* no es verdadera.

Así enunciada, esta definición es indeseablemente circular (pues 'es verdadera' aparece en el *definiens*), pero es fácil transformarla en una definición no circular por medio de un truco matemático debido a Frege y Dedekind (véase por ejemplo Torretti (1998), Apéndice XI):

> una oración *O* de *L2* es *OV* en *L2* si y sólo si *O* pertenece a *todo* conjunto *C* de oraciones que (α) contiene a las oraciones de la forma '*n*=*n*' y que (β) siempre que no contiene a una oración *P* contiene a la oración '¬*P*'.

Sin embargo, esta manera de proceder no es factible en el caso de lenguajes como *L1*, pues en ellos hay oraciones complejas que no se forman a partir de *oraciones* menos complejas, sino de fórmulas con variables libres (la oración (falsa) '∀x,∀x,,(Ax,x,,∧Ax,,x,)' es un ejemplo), y no se ve cómo especificar ni un conjunto básico de oraciones verdaderas ni un proceso recursivo de "transmisión" de la propiedad de la verdad en términos de la verdad o falsedad de oraciones menos complejas. Gran parte del mérito de Tarski se debe a que se diera cuenta de que, en sus propias palabras,

> se presenta la posibilidad de introducir un concepto más general que sea aplicable a cualquier fórmula, pueda ser definido recursivamente, y que, cuando se aplique a oraciones, nos lleve directamente al concepto de verdad. Estos requisitos los cumple la noción de la *satisfacción de una fórmula dada por objetos dados*.

La noción de satisfacción es una noción familiar: por ejemplo, el número 3 satisface la ecuación '*x*-2=1'; Frida Kahlo y Diego Rivera satisfacen (no importa en qué orden) la fórmula de *L1* 'Ax,x,,'; Toluca, Morelia y Puebla satisfacen, en ese orden, la "fórmula" '*x* se halla entre *y* y *z*'. Tarski define recursivamente una versión abstracta de este concepto para *L1*: la noción de satisfacción de una fórmula cualquiera *F* por una secuencia *infinita h* (con dominio los enteros positivos, o lo que viene a ser lo mismo, las variables de *L1*, y recorrido incluido en la clase de las personas)—y no una noción de satisfacción para fórmulas y objetos, pares de objetos, tríos de objetos, etc. (Como las fórmulas de *L1* pueden incluir un número arbitrario de variables, utilizar secuencias infinitas permite asignar en cada caso valores a todas las variables posibles, independientemente del número específico de variables de cada fórmula.) Intuitivamente, *F* es satisfecha por *h* cuando *F* es "verdadera" si interpretamos cada variable libre x_n que aparezca en *F* como si fuera un nombre de la persona *h*(*n*). La definición es la siguiente:

> Una secuencia infinita de personas *h* satisface una fórmula *F* si y sólo si *h* y *F* son tales que o bien (α) existen números naturales *k* y *l* tales que

> $F=\ulcorner Ax_k x_l \urcorner$ y h_k ama a h_l; o (β) hay una fórmula G tal que $F=\ulcorner \neg G \urcorner$ y h no satisface la fórmula G; o (γ) hay fórmulas G y H tales que $F=\ulcorner (G \wedge H) \urcorner$ y h satisface G y h satisface H; o finalmente (δ) hay un número natural k y una fórmula G tales que $F=\ulcorner \forall x_k G \urcorner$ y toda secuencia infinita de personas que difiera de h a lo sumo en el lugar k-ésimo satisface la fórmula G.

De nuevo esta definición recursiva es circular estrictamente hablando, pues emplea el predicado 'satisface'. Pero de nuevo puede ponerse en forma no circular por medio del truco de Frege y Dedekind. La definición no circular es la siguiente:

> Una secuencia infinita de personas h satisface una fórmula F si y sólo si el par $<h, F>$ pertenece a toda relación R entre secuencias y fórmulas tal que (α) si existen números naturales k y l tales que $G=\ulcorner Ax_k x_1 \urcorner$ y g_k ama a g_l, entonces $<g, G>$ pertenece a R; (β) si $<g, G>$ no pertenece a R, entonces $<g, \ulcorner \neg G \urcorner>$ pertenece a R; (γ) si $<g, G>$ pertenece a R y $<g, H>$ pertenece a R, entonces $<g, \ulcorner (G \wedge H) \urcorner>$ pertenece a R; (δ) para todo número natural k y fórmula G, si toda secuencia infinita de personas f que difiera de una secuencia g a lo sumo en el lugar k-ésimo es tal que $<f, G>$ pertenece a R, entonces $<g, \ulcorner \forall x_k G \urcorner>$ pertenece a R.

La definición de satisfacción tiene como consecuencia que el que una secuencia h satisfaga una fórmula F o no sólo depende de lo que h asigne a (los subíndices de) las variables que aparezcan libres en F. En otras palabras, para la satisfacción de una fórmula por una secuencia sólo importa lo que la secuencia asigne a las variables que aparecen libres en la fórmula, mientras que lo que asigne a las ligadas es indiferente. Si la secuencia h satisface la fórmula F, y la secuencia infinita de personas g es tal que para todo k, $h_k=g_k$ si v_k es una variable libre de F, entonces la secuencia g también satisface la fórmula F. (Se habla de esto como un resultado o lema de coincidencia.) Esto implica que si F es una oración —una fórmula sin variables libres— entonces o toda secuencia satisface F o ninguna lo hace. Además, la definición de satisfacción se ha construido de forma que las oraciones intuitivamente verdaderas son las del primer

tipo, las satisfechas por toda secuencia, y por tanto es posible definir de esta forma el predicado 'es *OV* en *L1*':

> *O* es *OV* en *L1* si y sólo si *O* es una oración de *L1* y toda secuencia infinita de personas satisface *O*.

5. Virtudes de la teoría tarskiana de la verdad

Tarski señaló que la comprobación de que esta definición es apropiada se haría demostrando rigurosamente que es una "definición adecuada de verdad" en el sentido de la convención T. (Véase Tarski (1944), secs. 4 y 9.) Ello requeriría la formalización de la metateoría, tarea ardua donde las haya que Tarski no estaba dispuesto a realizar. Sin embargo, el hecho es intuitivamente claro y Tarski se conformó con mostrar cómo se demostrarían en el metalenguaje algunos bicondicionales de los mencionados en la convención T. He aquí un ejemplo de cómo se haría esto con el bicondicional correspondiente a la oración de *L1* '∀x,Ax,x,', o sea

> ⌈∀x,Ax,x,⌉ es *OV* en *L1* si y sólo si para toda persona *a*, *a* ama *a a*:
> ⌈∀x,Ax,x,⌉ es *OV* en *L1* si y sólo si para toda secuencia *h*, *h* satisface '∀x,Ax,x,'
> si y sólo si para toda *h*, para toda *g* que difiere de *h* a lo sumo en lo que asigna a 1, *g* satisface 'Ax,x,'
> si y sólo si para toda *h*, para toda *g* que difiere de *h* a lo sumo en lo que asigna a 1, g(1) ama a *g*(1)
> si y sólo si para toda persona *a*, *a* ama *a* (pues dada una secuencia *h*, toda persona es asignada a 1 por alguna secuencia que difiere de *h* en lo asignado a 1).

Al satisfacer la convención T, una definición tarskiana del predicado de verdad será, como vimos, extensionalmente correcta. Quizá no es exagerado decir que, al indicar un sentido notablemente claro en que una caracterización "insustantivista" de la noción de verdad es indisputablemente extensionalmente correcta, la teoría semántica

de la verdad constituye uno de los mayores logros en la investigación filosófica sobre esta noción.

Podemos apreciar también ahora las razones por las que la teoría semántica tiene las características (i) a (iii) mencionadas al final de la sección 3:

(i) No se ve afectada por las dificultades generadas para el redundantismo y el pro-oracionalismo por oraciones como (3.3) y (3.4). A pesar de ser "insustantivista" en un sentido más o menos claro, la teoría semántica no propone que la expresión 'es verdadera' no exprese una propiedad, y, además, los predicados definidos que muestra cómo construir son predicados genuinos, que expresan propiedades de un cierto tipo, si bien propiedades "insustantivas". En el supuesto de que tengamos un predicado 'es *OV* en *L*' en un metalenguaje para un lenguaje *L* que contenga la primera oración pronunciada por Sor Juana Inés de la Cruz (y que incluya en dicho metalenguaje la descripción 'la primera oración pronunciada por Sor Juana Inés de la Cruz') y que contenga todos los teoremas matemáticos (e incluya en dicho metalenguaje el predicado 'es un teorema matemático'), cosa para la cual no hay obstáculos de principio, podremos decir sin problemas, usando recursos gramaticales y sintácticos perfectamente naturales, que la primera oración pronunciada por Sor Juana es *OV* en *L* y que todos los teoremas matemáticos son *OV* en *L*.

(ii) Ofrece un cierto tipo de "solución" para la paradoja del mentiroso. La razón es que el predicado 'es *OV* en *L*' es siempre un predicado de un lenguaje diferente al lenguaje de las oraciones a las que se aplica. En el caso de la oración (1.2) es fácil ver cómo este hecho bloquea la derivación de la paradoja para 'no es *OV* en *L*' (que expresaría "es falsa en *L*"): no es posible formar una oración del lenguaje para el que hemos definido 'es *OV* en *L*' que diga de sí misma que no es *OV* en *L*, pues 'no es *OV* en *L*' no es un predicado de ese lenguaje; por otro lado, es ciertamente posible formar una oración *O* del metalenguaje que diga de sí misma que no es *OV* en *L*, pero al ser una oración del metalenguaje es simplemente verdadera y no paradójica, pues no hay razón de que valga para ella el

bicondicional análogo a (1.3) (‘*O* es *OV* en *L* syss *O* no es *OV* en *L*’), que es esencial para la derivación de la paradoja; ese tipo de bicondicionales sólo valen para oraciones del lenguaje para el que se ha definido ‘es *OV* en *L*’, no para oraciones del metalenguaje.

Mencionemos aquí que en su trabajo sobre la verdad Tarski aplica el razonamiento de la paradoja del mentiroso en una prueba puramente matemática de que no es posible que un lenguaje de a lo sumo el mismo orden que un lenguaje *L* (y que pueda funcionar como metalenguaje de *L*, en el sentido de que pueda expresar suficientes cosas sobre la sintaxis de *L*) contenga un predicado de verdad para *L*. El ejemplo habitual es el lenguaje usual para la aritmética de primer orden y sus extensiones. Como Gödel mostró (y también Tarski, de forma independiente), estos lenguajes pueden hablar de su propia sintaxis: las expresiones de un lenguaje *L* de este tipo pueden ponerse en una correspondencia biunívoca con un subconjunto de los naturales (el de los números de Gödel de las expresiones de *L*), y para cada predicado sintáctico relevante puede definirse un predicado aritmético satisfecho precisamente por los números correspondientes a las expresiones que satisfacen el predicado sintáctico original. Usando resultados debidos en esencia a Gödel, es posible probar que, para cada fórmula (con una variable libre) $\ulcorner F(x) \urcorner$ de un lenguaje que extienda al de la aritmética, existe una oración *O* del mismo lenguaje verdadera si y sólo si el número de Gödel de *O* satisface $\ulcorner F(x) \urcorner$. Supongamos que “verdad para *L*” fuera expresable en *L* por medio de una fórmula $\ulcorner V(x) \urcorner$, esto es, que $\ulcorner V(x) \urcorner$ es satisfecha por todos y sólo los números de Gödel de oraciones verdaderas de *L*; por la proposición anterior habría una oración de *L*, *O*, verdadera si y sólo si el número de Gödel de *O* satisface $\ulcorner \neg V(x) \urcorner$; *O* “diría de sí misma” que es falsa. *O* es o bien verdadera o bien no lo es (siempre en la interpretación deseada de *L*); si es verdadera, su número de Gödel satisface $\ulcorner \neg V(x) \urcorner$ y por tanto *O* no es verdadera; si no es verdadera, su número de Gödel satisface $\ulcorner \neg V(x) \urcorner$, y por tanto es verdadera. Así, *O* es verdadera si y sólo si no lo es; la contradicción nos permite concluir que *L* no puede contener su propio predicado de verdad.

(Sobre esta prueba y las afirmaciones de este párrafo puede verse con provecho Tarski (1969).)

(iii) Por último, la teoría semántica proporciona una caracterización puramente matemática de las nociones de oración verdadera y de satisfacción, lo cual se ve claramente al inspeccionar los recursos en términos de los que se dan las definiciones tarskianas, que sólo incluyen esencialmente recursos matemáticos (los recursos sintácticos pueden matematizarse, como acabamos de mencionar). Como dijimos, quizá esta característica es la principal responsable de la preeminencia de la teoría semántica entre los lógicos e, indirectamente, entre los filósofos, especialmente entre los filósofos analíticos.

6. Algunas críticas a la teoría de Tarski y la motivación de la teoría kripkeana

A pesar de su preeminencia en los contextos donde se ha requerido un uso científico del concepto de verdad, en particular en el ámbito de la teoría de modelos en lógica matemática, la teoría semántica de la verdad no ha carecido de críticas. La raíz fundamental de todas ellas son las diferencias entre los predicados definidos por el método tarskiano y el predicado intuitivo de verdad. Una de esas diferencias es, como ya mencionamos, que son meramente extensionalmente equivalentes, pero no son analíticamente equivalentes, ni siquiera necesariamente equivalentes. Otra de esas diferencias es que no es posible decir con un predicado tarskiano—que siempre está restringido a un lenguaje particular—cosas que parecen poder decirse en virtud de la "universalidad" del predicado intuitivo (un ejemplo es la falsedad 'Todas las oraciones de todos los lenguajes son verdaderas'). Las críticas de Kripke a la teoría de Tarski tienen que ver con estas limitaciones expresivas de los predicados tarskianos.

Observemos que, si quisiéramos representar en el lenguaje natural, por medio de predicados definidos según las ideas de Tarski, el razonamiento de la antinomia del mentiroso usando la oración (1.2), al predicar 'no es verdadera' (que sería la forma tarskiana de expresar 'es falsa') de (1.2) en el razonamiento, este 'no es

verdadera' no podría ser el mismo predicado que el que aparece usado (en la forma 'es falsa') en (1.2); este último pertenecería a un lenguaje o "capa del lenguaje" "inferior" a la del predicado 'es verdadera' que predicamos de (1.2). Así, si el lenguaje natural funcionara según el modelo que parecería poder extraerse de las ideas de Tarski, deberíamos verlo como conteniendo muchos (de hecho, seguramente infinitos) predicados diferentes de verdad, todos ellos gráfica y fonéticamente idénticos, eso sí. (En realidad, Tarski no veía sus ideas como un modelo del funcionamiento del lenguaje natural, aunque fue natural para muchos conjeturar que sí apuntaban a un modelo plausible del concepto de verdad en el lenguaje natural, con muchos predicados diferentes pero gráfica y fonéticamente idénticos.) Al principio tendríamos un primer predicado de verdad que predicaríamos (o cuya negación predicaríamos) de oraciones que no contuvieran ningún predicado de verdad; después, un segundo predicado de verdad que predicaríamos (o cuya negación predicaríamos) de las oraciones que contuvieran a lo sumo el primer predicado de verdad, y así sucesivamente.

Una de las motivaciones de Kripke (1975) es ofrecer un intento de solución a la antinomia del mentiroso que supere algunas objeciones que genera el modelo del funcionamiento del concepto de verdad en el lenguaje natural que acabamos de describir—Kripke lo llama "el punto de vista ortodoxo", presumiblemente porque muchos en su época pensaban que ese modelo era plausible (¡aunque no Tarski!). Veamos algunas de las objeciones que expone el mismo Kripke. En primer lugar, aunque parece intuitivamente claro que toda oración en que aparezca el predicado de verdad debe de tener un "nivel" en el sentido explicado en el párrafo anterior, este nivel puede depender no sólo de la forma de la oración, como se supone, según Kripke, desde el punto de vista ortodoxo, sino también de ciertos hechos empíricos relativos a ella. Si Biden dice que la oración

(6.1) Todas las afirmaciones de Trump sobre la elección de Georgia son falsas

es verdadera, el "nivel" de este 'es verdadera' puede ser muy alto, dependiendo de las iteraciones de predicaciones de verdad o falsedad que haya hecho Trump en sus afirmaciones; puede no tratarse del segundo 'es verdadera' de la lista infinita del párrafo anterior. Por ejemplo, Trump puede haber dicho 'Jones no dice la verdad cuando dice que Smith dijo la verdad cuando afirmó que la elección de Georgia fue legal').

Otra de las objeciones que expone Kripke es la siguiente. Supongamos que Jones afirma (6.1), y que Trump afirma a su vez

(6.2) Todas las afirmaciones de Jones sobre la elección son falsas.

Entonces, en particular, Jones quiere afirmar que (6.2) es falsa, y Trump quiere afirmar que (6.1) es falsa; para ello, el nivel de 'es falso' en (6.1) ha de ser mayor que el de 'es falso' en (6.2), y viceversa, pero esto es imposible. Desde el punto de vista ortodoxo, al menos una de las dos oraciones estaría mal formada o no tendría sentido; sin embargo, podemos a menudo asignar valores de verdad a (6.1) y (6.2) sin ambigüedad: si al menos una afirmación de Jones sobre la elección es verdadera, (6.2) será falsa; si todas las otras afirmaciones de Trump sobre la elección son también falsas, (6.1) será verdadera. Pero no siempre podríamos decidirnos sobre el valor de verdad de (6.1) y (6.2): considérese el caso de que lo único que hubiesen afirmado Jones y Trump sobre la elección hubiesen sido, respectivamente, (6.1) y (6.2). El punto de vista ortodoxo elimina estos casos problemáticos barriendo de paso los casos en que (6.1) y (6.2) no presentan problemas.

La solución de Kripke a la antinomia del mentiroso pasa por proponer que el lenguaje natural puede verse como conteniendo únicamente un predicado de verdad. Kripke propone evitar la paradoja postulando que ciertas oraciones que contienen el predicado de verdad, entre ellas las paradójicas, no son ni verdaderas ni falsas, por no tener condiciones de verdad claras; en otras palabras, el predicado de verdad no está completamente definido, hay "huecos" en cuanto al valor de verdad (*truth-value gaps*). Ello bloquea

la aplicación del principio de tercio excluso en el argumento para obtener la paradoja. Kripke caracteriza el conjunto de las oraciones que contendrán el predicado de verdad y que serán verdaderas o falsas mediante la siguiente parábola: supongamos que queremos explicar el significado de 'es verdadera' a alguien que es en general un hablante competente del español, pero que no conoce el significado de ese predicado; si le decimos que una oración es verdadera cuando se dan las condiciones que nos permitirían afirmarla, el hablante podrá descubrir sucesivamente el valor de verdad de oraciones como "'La tierra gira alrededor del sol' es verdadera", " 'La tierra gira alrededor del sol' es verdadera" es verdadera", etc., que en última instancia podrá reducir a las condiciones de verdad de oraciones cuyo significado ya conoce; pero no podrá asignar un valor de verdad, por ejemplo, a oraciones autorreferenciales en que se predique de sí mismas la verdad o la falsedad. Verbigracia, no podrá asignar un valor de verdad a

(6.3) (6.3) es verdadero,

pues al buscar las condiciones que le permitirían afirmar (6.3) volverá siempre al punto de partida. Lo mismo le ocurriría con las oraciones paradójicas como (1.2), y con ejemplos de referencias cruzadas (como en el caso de que lo único que hubiesen afirmado Jones y Trump sobre la elección hubiesen sido, respectivamente, (6.1) y (6.2)). A las oraciones para las que el hablante del ejemplo (o nosotros mismos) podría encontrar un valor de verdad, eliminando predicados del tipo 'es verdadera', Kripke las llama "fundadas" (*grounded*).

La solución kripkeana a la paradoja del mentiroso, que hemos esbozado en lo esencial, no es la contribución más importante de Kripke a esta cuestión (de hecho, soluciones muy similares habían sido propuestas con anterioridad al artículo de Kripke). Lo más importante es que Kripke muestra cómo extender cualquier lenguaje interpretado (que pueda hablar de su propia sintaxis) a otro lenguaje interpretado que contiene su propio predicado de ver-

dad y en el que hay "huecos" en los valores de verdad, o sea en el que el predicado de verdad no está completamente definido. En concreto, los correlatos en la interpretación de las oraciones no fundadas, y por tanto los de las paradójicas, no satisfarán ni el predicado de verdad ni su negación. Naturalmente, el uso de Tarski del principio de tercio excluso en la prueba de que un lenguaje no puede contener su propio predicado de verdad presuponía que la interpretación del lenguaje considerado era clásica, esto es, que todos los predicados del lenguaje estaban completamente definidos por la interpretación—que cualquier objeto del dominio de la interpretación o bien satisface un predicado dado o su negación. A continuación, expondremos sucintamente el método de Kripke y el modelo a que da lugar. (En esta exposición de nuevo daremos por supuesta la familiaridad del lector y la lectora con varios aspectos de los lenguajes formales de primer orden, así como con ciertas nociones conjuntísticas elementales, aunque algo más avanzadas que las usadas en las definiciones tarskianas. Véase también la exposición en Teijeiro y Szmuc (2014).) Concluiremos viendo que este modelo no está sujeto a las objeciones que Kripke ponía al punto de vista ortodoxo, pero parece estar sujeto a otra objeción a su capacidad de modelar todas las intuiciones sobre el predicado de verdad en el lenguaje natural, la objeción generada por las llamadas "oraciones del mentiroso reforzado".

7. El método kripkeano para definir el predicado de verdad

Sea $\{P_1,\ldots,P_n, f_1,\ldots,f_m, c_1,\ldots,c_p\}$ el vocabulario no lógico de un lenguaje (finito) de primer orden L, en cuyo vocabulario lógico estén el cuantificador universal ('$\forall$'), el signo de conjunción ('$\wedge$'), el signo de negación ('$\neg$'), paréntesis ('(', ')'), el predicado diádico de identidad ('='), la letra 'x', y un acento subíndice (',') para generar un número infinito de variables por posposición a 'x'. (Como de costumbre, $P_1,\ldots,P_n$ son los predicados de *L*, $f_1,\ldots,f_m$ son sus símbolos de función, y $c_1,\ldots,c_p$ son sus constantes individuales.)

Diremos que una interpretación **A** para ese vocabulario y por tanto para *L* es *no clásica* si es un conjunto de la forma

$$\mathbf{A}=<A, <<P_1^{A1}, P_1^{A2}>,\ldots, <P_n^{A1}, P_n^{A2}>>, <f_1^A,\ldots,f_m^A>, <c_1^A,\ldots,c_p^A>>,$$

donde A es un conjunto no vacío, P_i^{Aj} es siempre un subconjunto de A^k (cuando k es la adicidad de P_i), $P_i^{A1} \cap P_i^{A2} = \emptyset$, f_i^A es siempre una función de A^k (cuando k es la adicidad de f_i) en A, y c_i^A es siempre un elemento de A. (Así, si para todo i=1,…,n, si P*i* es k-ádico, $P_i^{A1} \cup P_i^{A2} = A^k$, **A** es en esencia una interpretación clásica). Hablaremos de P_i^{A1} como la *extensión* de un predicado P_i, y de P_i^{A2} como su *antiextensión*. (Nótese que aquí el subíndice 'i' indica de cuál de los predicados en la lista $P_1,\ldots,P_n$ estamos hablando, mientras que los superíndices indican si estamos hablando de la extensión de P_i en **A** (P_i^{A1}) o de su antiextensión (P_i^{A2}).)

Podemos definir para *L*, una interpretación no clásica **A**, y cada secuencia h que sea una asignación de objetos de A al conjunto de las variables de *L*, una función de denotación h* del conjunto de los términos de *L* en A, análoga a la definida para interpretaciones clásicas. Kripke usa la lógica trivaluada fuerte de Kleene para tratar los predicados no completamente definidos, en el sentido siguiente (puede verse también, por ejemplo, Teijeiro y Szmuc (2014), sec. 2); siguiendo los esquemas de Kleene podemos definir recursivamente una función **A*** que asigne a cada formula de *L* un valor de entre v, f, i (leídos "verdadero", "falso", "indeterminado") mediante una secuencia h, de la siguiente manera:

A*h($\ulcorner t_1=t_2 \urcorner$) (donde t_1 y t_2 son términos de *L*) es v si $h^*(t_1)=h^*(t_2)$; f si $h^*(t_1)\neq h^*(t_2)$; nunca es i.

A*h($\ulcorner Pt_1,\ldots,t_k \urcorner$) (donde P es un predicado k-ádico de *L* y $t_1,\ldots,t_k$ son términos de *L*) es v si $<h^*(t_1),\ldots,h^*(t_k)> \in P^{A1}$; es f si $<h^*(t_1),\ldots,h^*(t_k)> \in P^{A2}$; es i si $<h^*(t_1),\ldots,h^*(t_k)> \notin P^{A1} \cup P^{A2}$.

$A^*_h(\ulcorner \neg F \urcorner)$ (donde *F* es una fórmula de *L*) es v si $A^*_h(F)$ es f; es f si $A^*_h(F)$ es v; es i si $A^*_h(F)$ es i.

$A^*_h(\ulcorner F \wedge G \urcorner)$ (donde *F* y *G* son fórmulas de *L*) es v si $A^*_h(F) =$ $A^*_h(G) =$ v; es f si $A^*_h(F) =$ f o $A^*_h(G) =$ f; es i en otro caso.
$A^*_h(\ulcorner \forall x_n F \urcorner)$ (donde *F* es una fórmula de *L*) es v si para toda secuencia j que difiera de h a lo sumo en el valor de x_n, $\mathbf{A}^*j(F) =$ v; es f si hay una secuencia j que difiere de h a lo sumo en el valor de x_n y tal que $\mathbf{A}^*j(F) =$ f; es i en otro caso.

Esta semántica permite probar fácilmente un lema de coincidencia para oraciones de *L*.

Vayamos con la construcción de Kripke. Partamos de una cierta interpretación clásica **A** de *L*. Supongamos además que *L* tiene suficientes recursos para hablar de su sintaxis (*L* puede ser el lenguaje usual de primer orden para la aritmética; también podemos imaginar que *L* es algún fragmento formalizado del lenguaje natural sin el predicado 'es verdadero'). Sea *L'* el lenguaje que resulta de añadir a *L* un predicado monádico 'V' (que desempeñará el papel de 'es verdadero'). Sea $\mathbf{A}(S_1,S_2)$ la interpretación no clásica para *L'* definida como sigue: $\mathbf{A}(S_1,S_2)$ coincide con **A** en el universo y en la interpretación de los signos de *L*; además, S_1 es $V^{A(S1,S2)1}$ y S_2 es $V^{A(S1,S2)2}$. Sea [*O*] el "código" de una oración *O*, es decir, el objeto que le corresponde en la interpretación (en el caso de la aritmética su número de Gödel; en el caso del lenguaje natural, la oración misma). Sean entonces

$S_1' = \{[O]$: *O* es una oración de *L'* y $\mathbf{A}(S_1,S_2)^*(O) =$ v$\}$
$S_2' = \{[O]$: *O* es una oración de *L'* y $\mathbf{A}(S_1,S_2)^*(O) =$ f$\} \cup$
$\{a \in A$: a no es el código de una oración de *L'*$\}$.

Si queremos interpretar 'V' como "verdad en $\mathbf{A}(S_1,S_2)$", debemos exigir que $S_1 = S_1'$ y $S_2 = S_2'$, esto es, que la extensión de 'V' sean (los códigos de) las oraciones verdaderas en $\mathbf{A}(S_1,S_2)$ y que su antiextensión sean (los códigos de) las oraciones falsas (más las cosas que no son (códigos de) oraciones). Sea φ la función que asigna a cada par de subconjuntos disjuntos de A, $<S_1,S_2>$, el par $<S_1',S_2'>$ (en la notación de más arriba). Un punto fijo de φ (esto es, un par $<S_1,S_2>$ de subconjuntos disjuntos de A para el que

$\varphi(<S_1,S_2>)=<S_1,S_2>$ ($=<S_1',S_2'>$)) proporcionará por tanto una interpretación apropiada del predicado 'V'. Así, probar la existencia de puntos fijos de φ es probar la existencia de una interpretación de *L'* bajo la cual *L'* contiene su propio predicado de verdad. El principal resultado de Kripke es la prueba de que existen puntos fijos de φ.

En una definición por recursión transfinita definimos una función con valores para todos los ordinales, incluidos los transfinitos. Definamos por recursión transfinita, para cada ordinal α, una estructura no clásica $\mathbf{A}_\alpha$ como sigue (si $\mathbf{A}_\beta = \mathbf{A}(S_1,S_2)$, ponemos $S_1=S_{1\beta}$ y $S_2=S_{2\beta}$):

$\mathbf{A}_0 = \mathbf{A}(\emptyset,\emptyset)$;
si $\alpha=\beta+1$ y $\mathbf{A}_\beta = \mathbf{A}(S_1,S_2)$, $\mathbf{A}_\alpha = A(S_1',S_2')$;
si α es un ordinal límite, $\mathbf{A}\alpha = \mathbf{A}(\cup_{\beta<\alpha}S_{1\beta}, \cup_{\beta<\alpha}S_{2\beta})$.

Hay un paralelo entre la generación recursiva de estas estructuras y el proceso imaginario mediante el que el hablante de nuestra parábola asignaba sucesivamente valores de verdad a nuevas oraciones que contenían el predicado 'es verdadero': en un primer momento, la interpretación del predicado está completamente indeterminada (en el nivel 0) y sucesivamente van apareciendo en su extensión y su antiextensión nuevas oraciones.

Digamos que $<S_1,S_2>\leq<S_1\dagger,S_2\dagger>$ ($<S_1\dagger,S_2\dagger>$ extiende a $<S_1,S_2>$) si y sólo si $S_1 \subset S1\dagger$ y $S2 \subset S_2\dagger$. Entonces φ es monótona respecto del orden ≤: si $<S_1,S_2>\leq<S_1\dagger,S_2\dagger>$, entonces $\varphi(<S_1,S_2>)\leq\varphi(<S_1\dagger,S_2\dagger>)$ (la prueba es una inducción sobre la complejidad de las oraciones de *L'*). Por la monotonía de φ, es fácil ver por inducción ordinal que si $\beta<\alpha$, la interpretación de 'V' en $\mathbf{A}_\alpha$ extiende a la interpretación de 'V' en $\mathbf{A}_\beta$. Ahora no es difícil ver que hay α tal que $\varphi(<S_{1\alpha},S_{2\alpha}>)=<S_{1\alpha},S_{2\alpha}>$, o sea que hay puntos fijos de φ. Esto se ve observando que habrá un ordinal α tal que $<S_{1\alpha},S_{2\alpha}>=<S_{1\alpha+1},S_{2\alpha+1}>$: si para todo α o bien $S_{1\alpha}$ está incluido propiamente en $S_{1\alpha+1}$ o bien $S_{2\alpha}$ está incluido propiamente en $S_{2\alpha+1}$, entonces el conjunto de oraciones de *L'* en $S_{1\omega1} \cup S_{2\omega1}$ sería no numerable, dado que la clase de todos los ordinales no es nu-

merable, pero sólo hay un número numerable de oraciones en *L'*. La existencia de un punto fijo de φ garantiza que *L'* puede interpretarse de modo que 'V' sea un predicado de verdad para *L'*.

Siguiendo con la metáfora anterior, un punto fijo sería el análogo del momento en que el hablante imaginario ya no puede dar más valores de verdad a oraciones del español; fuera de la extensión y la antiextensión para 'V' que suministra el punto fijo quedarán las oraciones no fundadas, y entre ellas las paradójicas. Se puede ahora dar una definición formal precisa de lo que es una oración fundada (de *L'*): es una oración que es verdadera o falsa en la interpretación que suministra el menor punto fijo de φ. Si *O* es una oración fundada, el menor ordinal α para el que $\mathbf{A}_\alpha$ da un valor de verdad a *O* proporciona una medida del nivel de *O*, en el sentido intuitivo de 'nivel' del que habíamos hablado. Es fácil ver, por inducción ordinal, que las oraciones de *L'* en las que aparecen autopredicaciones de verdad o falsedad (por ejemplo, las oraciones de la aritmética de las que hablábamos al final de la sección 5, o la oración (6.3) y las oraciones autorreferenciales paradójicas como (1.2) en el caso del lenguaje natural) no serán ni verdaderas ni falsas en ninguna interpretación $\mathbf{A}_\alpha$; así, no serán fundadas.

Veamos cómo se resuelven en el modelo recién descrito los problemas que Kripke detectaba en el "punto de vista ortodoxo" sobre la paradoja del mentiroso. En primer lugar, en el caso de oraciones como (6.1), la interpretación inicial (**A**, en la construcción anterior) proporciona la extensión del predicado 'ser una afirmación de Trump sobre la elección de Georgia'; si en ella hay alguna oración no fundada, (6.1) será no fundada; si todas las oraciones en ella son fundadas, (6.1) también lo será, y pertenecerá a la extensión o la antiextensión de 'V' en alguna $\mathbf{A}_\alpha$ con α suficientemente grande; en cualquier caso, el nivel de (6.1) (en el sentido preciso mencionado en el párrafo anterior) dependerá de **A** de un modo esencial. Lo mismo ocurrirá en el caso de la referencia cruzada de (6.1) y (6.2): que sean fundadas dependerá de las extensiones de 'ser una afirmación de Trump sobre la elección de Georgia' y de 'ser una afirmación de Jones sobre la elección', proporcionadas

por **A**, de la que también dependerá el nivel de (6.1) y (6.2) si son fundadas. En el ejemplo anterior, si una de las afirmaciones de Jones sobre la elección es fundada y verdadera, de nivel α, (6.2) será falsa de nivel α+1; si además todas las afirmaciones de Trump distintas de (6.2) son falsas, entonces (6.1) será verdadera y de un nivel superior a α+1.

También podemos ver que la teoría de Kripke cumple con los desiderátums (i) a (iii) de la sección 3. En cuanto a (i), no se ve afectada por las dificultades generadas para el redundantismo y el pro-oracionalismo por oraciones como (3.3) y (3.4), pues, al igual que la teoría de Tarski, la teoría de Kripke propone que el predicado de verdad expresa una propiedad de un cierto tipo, si bien una razonablemente "insustantiva". De nuevo un predicado kripkeano de verdad para un lenguaje apropiado *L* podrá usarse para expresar de manera natural que a la primera oración pronunciada por Sor Juana se aplica ese predicado y que a todos los teoremas matemáticos se les aplica ese predicado. Por otro lado, la teoría evidentemente ofrece un cierto tipo de solución para la paradoja del mentiroso, como vimos hace un par de párrafos, según requería (ii). Por último, según requería (iii), la teoría de Kripke proporciona una caracterización meramente conjuntística, y por tanto puramente matemática, de la noción de oración verdadera para un lenguaje, lo cual de nuevo comprobamos inspeccionando los recursos utilizados en la construcción de más arriba.

¿Quiere esto decir que la teoría de Kripke es claramente la teoría correcta de la verdad, o el modelo correcto del funcionamiento del predicado de verdad en el lenguaje natural, o al menos la teoría o el modelo correcto dados los desiderátums insustantivistas? Hay al menos una objeción que podría usarse para poner esto en duda, creada por las llamadas "oraciones del mentiroso reforzado". Una oración del mentiroso reforzado es, en la versión más simple y típica, una oración que afirma de sí misma que no es verdadera, a diferencia de una oración del mentiroso normal, que afirma de sí misma que es falsa (el caso de (1.2), recordemos); por ejemplo,

(7.1) (7.1) no es verdadera

es una oración del mentiroso reforzado.

Si aceptamos el principio de bivalencia de la lógica clásica, (7.1) es equivalente a

(7.2) (7.1) es falsa.

Si suponemos, por contra, que hay oraciones sin valor de verdad, ni verdaderas ni falsas, (7.1) y (7.2) no son equivalentes. (7.1) es intuitivamente paradójica: si es verdadera, entonces no lo es; si no es verdadera, entonces es verdadera, pues afirma precisamente que no es verdadera.

En los modelos con huecos en cuanto al valor de verdad, las oraciones paradójicas como (7.1) carecen de valor de verdad. Por ejemplo, consideremos el modelo proporcionado por un punto fijo como el construido inductivamente por Kripke para el lenguaje de la aritmética de primer orden con un predicado de verdad, 'V'. Si n* es el nombre estándar (numeral) de n y n es el número de Gödel de $\ulcorner \neg Vn^* \urcorner$, entonces n no está ni en la extensión ni en la antiextensión de 'V' en el modelo del punto fijo, y así $\ulcorner \neg Vn^* \urcorner$ no es ni verdadera ni falsa en el modelo. (Esto puede verse mostrando por inducción ordinal que n no pertenece a la extensión ni a la antiextensión de 'V' en ninguna de las estructuras $\mathbf{A}_\alpha$.) El problema resulta entonces de la siguiente observación: si $\ulcorner \neg Vn^* \urcorner$ no es ni verdadera ni falsa, entonces no es verdadera, pero esto parece ser precisamente lo que afirma $\ulcorner \neg Vn^* \urcorner$, y por tanto $\ulcorner \neg Vn^* \urcorner$ será verdadera después de todo; en el lenguaje natural, el argumento paralelo procedería así: si '(7.1) no es verdadera' no es ni verdadera ni falsa, entonces '(7.1) no es verdadera' no es verdadera, pero entonces (7.1) no es verdadera, y así '(7.1) no es verdadera' será verdadera después de todo.

Naturalmente, esta contradicción no puede reproducirse formalmente, pues, como señala Kripke, el lenguaje objeto (aquel para el que se construye el modelo) no puede expresar que una oración suya es o falsa o carece de valor de verdad; es decir, no

puede expresar que una oración no es verdadera, en el sentido en que decimos en el metalenguaje (por ejemplo, en el argumento anterior que llevaba a la contradicción) que no es verdadera. Ello se debe a la interpretación de la negación en la lógica trivaluada de Kleene, que se desprende de las instrucciones de más arriba: que ⌜¬Vt⌝ sea verdadera en una estructura no clásica **A** quiere decir que ⌜Vt⌝ es falsa, es decir, que el número (de Gödel) nombrado por t pertenece a la antiextensión de 'V' en **A**, o sea que la oración cuyo número de Gödel es el número denotado por t es falsa; no quiere decir que esta oración es o falsa o carece de valor de verdad.

Así, las oraciones del mentiroso reforzado parecen poner de manifiesto una limitación en los medios de expresión de los lenguajes interpretados mediante una semántica trivaluada: el predicado 'V' del lenguaje objeto y el predicado 'es verdadero' del metalenguaje no expresan lo mismo, pues es verdadero que una oración del mentiroso reforzado como ⌜¬Vn*⌝ no es verdadera, según la semántica intuitiva del metalenguaje, pero suponer que la traducción obvia de esta afirmación metalingüística es verdadera en el modelo trivaluado, es decir, suponer que el número de Gödel de ⌜¬Vn*⌝ está en la extensión de 'V', implica una contradicción. Una de las motivaciones de las construcciones de modelos trivaluados es sin embargo mostrar que un predicado del lenguaje objeto puede ser interpretado como aplicándose a todo lo que en el metalenguaje podemos establecer como verdadero; por tanto, las oraciones del mentiroso reforzado cuestionan el que esto sea posible (y refutan de hecho que ello se haya conseguido en el caso del modelo de Kripke). En definitiva, las oraciones del mentiroso reforzado ponen en duda que la semántica trivaluada sea una maqueta fiel y lo más rica posible del comportamiento del predicado de verdad del lenguaje natural.

Hay muchas otras teorías matemáticas (y por tanto "insustantivistas") de la verdad en la tradición analítica reciente, que han buscado aprovechar las lecciones que dejaron las teorías de Tarski y Kripke en la construcción de modelos que superen los problemas que estas teorías enfrentan. (Sobre algunas de estas teorías, en español pueden

verse Buacar y Picollo (2014), Martínez (2014), Rosenblatt y Pailos (2014) y Tajer (2014).) Pero también hay varias teorías de otros tipos, dentro y fuera de la tradición analítica (algunas de las cuales están representadas en la antología Nicolás y Frápolli (1997)). La situación es fluida y compleja, e intentar describirla panorámicamente pero con un detalle satisfactorio es una tarea complicada donde las haya, que habrá de quedar para otra ocasión.

Referencias

Austin, J. L. (1950). Verdad, en Nicolás y Frápolli (1997), 225-242.

Barrio, E. A. (comp.) (2014a). *La Lógica de la Verdad*, Eudeba. Buenos Aires.

Barrio, E. A. (2014b). Definiciones tarskianas de la verdad, en Barrio (2014a), 25-73.

Barrio, E. A. (2014c) (comp.). *Paradojas, Paradojas y más Paradojas*, College Publications, Londres.

Buacar, N. M. y L. M. Picollo (2014). La teoría revisionista de la verdad. En Barrio (2014a), 117-186.

Gómez Torrente, M. (2000). *Forma y Modalidad. Una Introducción al Concepto de Consecuencia Lógica*. Eudeba, Buenos Aires.

Gómez Torrente, M. (2001a). Notas sobre el *Wahrheitsbegriff*, I, *Análisis Filosófico* 21, 5-41.

Gómez Torrente, M. (2001b). Notas sobre el *Wahrheitsbegriff*, II, *Análisis Filosófico* 21, 149-185.

Hempel, C. G. (1935). La teoría de la verdad de los positivistas lógicos. En Nicolás y Frápolli (1997), 481-493.

James, W. (1906). Concepción de la verdad según el pragmatismo. En Nicolás y Frápolli (1997), 25-43.

Kripke, S. (1975). Esbozo de una teoría de la verdad. En Nicolás y Frápolli (1997), 109-143.

Martínez, J. (2014). La paradoja del mentiroso. En Barrio (2014c), 11-26.

Nicolás, J. A. y M. J. Frápolli (comps.) (1997). *Teorías de la Verdad en el Siglo XX*. Tecnos, Madrid.

Ramsey, F. P. (1927). La naturaleza de la verdad. En Nicolás y Frápolli (1997), 265-279.

Rosenblatt, L. y F. Pailos (2014). Paracompletitud sofisticada. En Barrio (2014a), 187-248.

Strawson, P. F. (1950). Verdad. En Nicolás y Frápolli (1997), 281-307.

Tajer, D. (2014). Dialeteísmo. Una teoría contradictoria de la verdad. En Barrio (2014a), 249-292.

Tarski, A. (1944). La concepción semántica de la verdad y los fundamentos de la semántica. En Nicolás y Frápolli (1997), 65-108.

Tarski, A. (1969). Verdad y Demostración. *Disputatio. Philosophical Research Bulletin* 4 (2015), 367-396.

Teijeiro, P. y D. E. Szmuc (2014). Teoría de puntos fijos de Kripke. En Barrio (2014a), 75-116.

Torretti, R. (1998). *El Paraíso de Cantor. La Tradición Conjuntista en la Filosofía Matemática*. Editorial Universitaria, Santiago de Chile.

Williams, C. J. F. (1992). La teoría pro-oracional de la verdad. En Nicolás y Frápolli (1997), 309-320.

CAPÍTULO IV

TEORÍA DE MODELOS

Raymundo R. Meza[1]

RESUMEN

En la actualidad, la teoría de modelos se ha consolidado como una disciplina de la lógica matemática cuyos resultados han enriquecido

---

1 Raymundo R. Meza, Posgrado en Filosofía de la Ciencia, Universidad Nacional Autónoma de México (PFC-UNAM); Universidad Rosario Castellanos (URC). Licenciado en Filosofía y Maestro en Filosofía de la Ciencia por la UNAM. Actualmente, se encuentra realizando estudios de doctorado en Filosofía de la Ciencia con especialidad en Filosofía de las Matemáticas y Lógica de la Ciencia dentro de la misma institución. Desde 2020, es miembro del Programa de Estudiantes Asociados del Instituto de Investigaciones Filosóficas de la UNAM. Asimismo, de 2021 a 2024, se desempeñó como profesor ayudante en varios cursos de lógica de nivel licenciatura y, desde 2024, como profesor titular en el área de lógica en la URC.

Facultad de Filosofía y Letras, Universidad Nacional Autónoma de México. Este trabajo fue realizado como parte de los proyectos de investigación PAPIIT IN406225 "Los principios de la logica" y PROINV_24_31 "Fortalecimiento a la Red de Difusión de Lógica y Argumentación (ReDLA)", así como con el apoyo económico del Consejo Nacional de Humanidades, Ciencias y Tecnologías

el análisis filosófico en varias áreas. El propósito de este capítulo es mostrar cómo esta disciplina puede emplearse para abordar con precisión cuestiones de diversa naturaleza filosófica. Para ello, introduzco algunas nociones básicas y resultados importantes de la teoría de modelos. Luego, examino el contexto histórico en el que se originó, así como las motivaciones que subyacen al surgimiento y desarrollo de sus conceptos fundamentales. Posteriormente, exploro las relaciones entre estos últimos y señalo su papel en la investigación lógica de la sintaxis y la semántica de las teorías formales. Con base en estas consideraciones, destaco el valor de la disciplina para generar interpretaciones de teorías científicas y matemáticas. Esto con la finalidad de abordar una cuestión que atraviesa múltiples áreas de la filosofía, a saber, el problema de la inescrutabilidad de la referencia. La relevancia de este problema en diferentes ámbitos y su análisis mediante el uso de instrumentos modelo-teóricos permiten ilustrar, hacia el final del capítulo, el modo en que la teoría de modelos puede contribuir al planteamiento, tratamiento y esclarecimiento de asuntos de interés para el filósofo así como de discusiones interdisciplinares.

Palabras clave: modelos, lenguajes formales, teorías, lógica, verdad, interpretación, inescrutabilidad de la referencia.

ABSTRACT

Nowadays, model theory has established itself as a branch of mathematical logic whose results have enriched philosophical analysis in various areas. This chapter aims to show how model theory can address questions of a diverse philosophical nature with preciseness. Toward this end, I introduce some basic notions and significant results of model theory. I then examine the historical context in which it originated, along with the motivations underlying the emergence and development of its fundamental concepts. Subsequently, I explore the relationships between these concepts and highlight their role in the logical investigation of the syntax and semantics of formal theories. Based on these considerations, I emphasize the value of the discipline in generating interpretations of

scientific and mathematical theories to address a question that cuts across multiple areas of philosophy, namely, the problem of the inscrutability of reference. The magnitude of this problem in different domains and its analysis through model-theoretic tools help us to show, toward the end of the chapter, how model theory can take part in the formulation, treatment, and clarification of issues of interest to philosophers, in addition to interdisciplinary discussions.

Keywords: models, formal languages, theories, logic, truth, interpretation, inscrutability of reference.

1. Introducción

Una de las formas principales de "representar el mundo" es a través del *lenguaje*. Si bien tal modo de caracterizar esta función del lenguaje —como medio para describir la realidad— puede parecer descuidado y poco informativo, difícilmente le resultará muy desconocido a alguien: muchas veces, cuando buscamos recuperar o identificar ciertos aspectos de la realidad, *interpretamos* nuestras expresiones lingüísticas, reconociendo su contenido o significado, con la intención de determinar si algún objeto, *verdaderamente* o no, posee alguna propiedad o mantiene alguna relación con otros objetos. Evidentemente, este primer acercamiento tampoco goza de mucha claridad ni precisión; sin embargo, basta al menos para motivar algunas preguntas iniciales: ¿cómo es que la interpretación le asigna y especifica significado a las palabras?, ¿cómo es que la interpretación establece una relación entre las palabras y los fragmentos de la realidad?

Con frecuencia, la *semántica* se asocia con el estudio del significado de las expresiones de un cierto lenguaje y, en este sentido, resulta natural formular las preguntas anteriores dentro del marco de esta disciplina. No obstante, la teoría del significado —como cabe imaginar— es bastante amplia y abarca cuestiones relacionadas con múltiples aspectos alrededor de su objeto de estudio[2]: ¿qué se entiende por significado?, ¿qué factores sociales, políticos,

2 Para una introducción sobre diferentes teorías del significado, véase Speaks (2021).

psicológicos y biológicos intervienen en la especificación del significado de determinadas palabras?, ¿por qué las palabras adquieren el significado que tienen?, ¿cómo es que las expresiones lingüísticas son procesadas e interpretadas por las personas, y cómo son asociadas con su contenido?, ¿cuál es la naturaleza de esa interpretación y cuáles son sus propiedades?, ¿estas propiedades varían en virtud de los diferentes tipos de lenguaje?, etcétera.

Por lo tanto, todo parece indicar que, si deseamos abordar las preguntas del primer párrafo con mayor seriedad, al menos debemos establecer de manera más precisa las nociones que se encuentran involucradas y la clase de respuesta que se intenta ofrecer.

Desde este punto de vista, la *teoría de modelos* es considerablemente más reducida y tan solo se concentra en una parte de las investigaciones en semántica. En particular, la teoría de modelos se ocupa del estudio matemático de las relaciones que hay entre los enunciados de las teorías formales y las estructuras matemáticas que los interpretan. Desde luego, quienes no conocen esta teoría probablemente no encuentren lo anterior especialmente muy esclarecedor, pero al menos les podrá sugerir una idea inicial sobre el tipo de análisis que se puede llevar a cabo dentro de ella y alrededor de qué está restringido. En cambio, quienes sí están familiarizados con ella seguramente se asombrarán por lo escueto de esta caracterización, pues es verdad que actualmente los teóricos de los modelos se han dedicado en gran medida a la construcción, comparación y clasificación de estructuras matemáticas. Por otra parte, tampoco es muy difícil encontrar propuestas modelo-teóricas que caen más allá del territorio de la matemática pura y de las teorías formales. Lamentablemente, debido al carácter introductorio de esta presentación, no podremos revisar aquí estas aproximaciones[3] y, por lo mismo, nos veremos limitados a presentar principalmente la teoría de modelos clásica para la lógica de primer orden, es decir, aquélla más cercana a la semántica de los lenguajes formales

3 No obstante, el lector interesado en estos temas puede consultar Davidson (1970) y Montague (1974). Para un recorrido histórico alrededor de la relación entre lógica y lingüística, véase Lenci y Sandu (2009).

que cuantifican únicamente sobre individuos y que se sirve de la maquinaria de la teoría de conjuntos para ofrecer sus definiciones. Aunque existen otras maneras de acercarse al análisis semántico de las teorías formales, la teoría de modelos clásica es posiblemente la más conocida y esto se debe en gran medida a lo que explicaremos más adelante.

En primera instancia, este enfoque de la teoría de modelos puede parecer algo especializado y su alcance extremadamente limitado, pero la sorprendente precisión de sus definiciones y sus poderosas herramientas conceptuales han logrado colocar a esta disciplina dentro de diferentes discusiones en áreas como lógica, filosofía del lenguaje, filosofía de las matemáticas, filosofía de la ciencia, metafísica, epistemología y filosofía de la ciencia de la computación.

El propósito de este capítulo es mostrar que la adopción de la teoría de modelos puede llegar a ser bastante útil tanto para identificar problemas como para motivar discusiones en varias áreas de la filosofía analítica. Sin embargo, para llevar a cabo esta empresa, primero es necesario introducir una serie de conceptos y resultados paradigmáticos, así como básicos.

De esta manera, el texto se divide en dos partes. Primero, se presentan nociones tales como estructura, interpretación, homomorfismo, signatura, teoría, satisfacción y modelo, entre otras; se exploran los vínculos entre ellas y se especifica su función en la investigación lógica de la sintaxis y la semántica de las teorías formales. Además, se ofrece una breve revisión del panorama histórico y de las motivaciones que llevaron al origen y desarrollo de estos conceptos.

La segunda parte, por su lado, se consagra al análisis de un problema filosófico que atraviesa diversas áreas de la filosofía (como filosofía del lenguaje, metafísica, epistemología, filosofía de la ciencia y filosofía de las matemáticas): el *problema de la inescrutabilidad de la referencia*. A través de este ejemplo, se espera que el lector pueda apreciar cómo las herramientas de la teoría de modelos

pueden contribuir a la precisión, el planteamiento y la claridad de diferentes discusiones filosóficas.

2. Algunas Nociones Básicas de la Teoría de Modelos[4]

Como adelantamos, la teoría de modelos clásica se ha consolidado como una disciplina de la lógica matemática que explora las relaciones que hay entre las estructuras matemáticas y las teorías formales, al grado de convertirse en la manera estándar para describir la semántica de los lenguajes formales. No obstante, al tratarse de una teoría puramente matemática, es importante anticipar que varios de sus términos se emplean en un sentido técnico que, en muchos casos, no coincide estrictamente con su uso coloquial. Por esta razón, resulta pertinente que, desde un principio, introduzcamos con claridad las nociones en que se apoyan sus definiciones básicas.

2.1. Lenguajes Formales y Estructuras[5]

Un *lenguaje formal* consiste en *i*) un conjunto de símbolos o *alfabeto* y *ii*) *reglas de formación* rigurosas (habitualmente *recursivas*) para construir sucesiones gramaticalmente correctas de tales símbolos. De este modo, si entendemos por *expresión* cualquier sucesión de símbolos, entonces el objetivo de estas reglas es determinar clara y distintamente una clase muy particular de expresiones, a saber, la clase de los *términos* y de las *fórmulas bien-formadas* (o simplemente *fórmulas*) del lenguaje. Aunque no es estrictamente necesario, contar con reglas de naturaleza recursiva resulta ser bastante útil, pues esto nos permite tener a nuestra disposición un *procedimiento efectivo* para *decidir* si cualquier expresión se trata o no de un término o de una fórmula.[6]

4 La siguiente exposición presupone cierto conocimiento mínimo de teoría de conjuntos, lógica cuantificacional de primer orden y metalógica. Por otro lado, cabe mencionar que todas traducciones que aparecen en este texto son propias.

5 El lector interesado en revisar más detalladamente el contenido de esta sección, y todas aquéllas en que se presenten definiciones formales, puede dirigirse a Manzano (1999), Hodges (1993; 1997), Chang y Keisler (1990), así como cualquier texto introductorio de lógica matemática como Enderton (2001) o Mendelson (1997).

A primera vista, el papel de la recursión puede parecer un mero capricho matemático, pero esto es así si tomamos muy al pie de la letra la noción de "símbolo". Si bien es cierto que en un inicio la concepción del lenguaje formal era más cercana a su contraparte coloquial del lenguaje natural, de manera que sus símbolos solían ser asociados con símbolos escritos susceptibles de reconocimiento por la sola inspección visual (Manzano, 1999, p. 35), hoy día la naturaleza ontológica de estos símbolos resulta ser irrelevante e incluso puede llegar a ser bastante abstracta (*e. g.*, las mismas entidades de la matemática pueden ser consideradas como nombres, ¡incluso de sí mismas!).

Dentro del alfabeto de un lenguaje formal, podemos distinguir el conjunto (posiblemente vacío) de símbolos no-lógicos junto con su *aridad* (o *adicidad*) correspondiente[7], este es, la *signatura* $\mathcal{L}$[8]. Si I, J y K son conjuntos de índices para los símbolos de constante, función y relación, respectivamente, entonces podemos definir la aridad como una función ar: $\cup\{I, J, K\} \rightarrow \mathrm{N}$. La aridad de los sím-

6 Una de las ventajas de trabajar con lenguajes formalizados es que, por lo menos en principio, nos permiten dar un tratamiento más estructural y formal que en el caso de los lenguajes naturales. Esto se debe a que abstraemos del lenguaje su forma sintáctica y eliminamos, por lo menos momentáneamente, su semántica y su pragmática. Justamente uno de los objetivos de la teoría de modelos es generar una semántica formal y precisa (idealmente recursiva) de los lenguajes formales.

7 Como se verá en unos momentos, el lenguaje que se considera en este trabajo cuenta con el símbolo de igualdad (=) que usualmente se interpreta como la *identidad*. Aunque frecuentemente el símbolo de igualdad se toma por lógico en la literatura, muchos autores lo consideran como no-lógico. De cualquier manera, en caso de no tener símbolos no-lógicos adicionales, incluso podríamos construir fórmulas únicamente a partir de la igualdad y trabajar con el llamado *lenguaje de la pura igualdad* (*o identidad*). Si consideramos como lógico al símbolo de la identidad, el lenguaje de la pura identidad sería un *lenguaje puramente logico*, es decir, uno en que todos sus símbolos son lógicos. Es importante destacar que los símbolos no-lógicos son aquellos cuya interpretación puede variar de un modelo a otro, mientras que los símbolos lógicos tendrán una misma interpretación en todos los modelos. Estos últimos son los que nos permitirán establecer la forma lógica de nuestras teorías.

8 Algunos autores le adjudican el mismo nombre tanto al lenguaje como a la signatura. Esto se debe a que el primero no es más que el conjunto de todas las fórmulas que pueden formarse a partir de los elementos de la signatura y los símbolos lógicos, en virtud de las reglas de formación.

bolos no-lógicos determina qué cantidad de lugares requiere ser ocupada por términos en cada caso para, o bien obtener un término (en el caso de los símbolos funcionales), o bien obtener una fórmula atómica (en el caso de los símbolos relacionales).

En adelante, cuando hablemos de $\mathcal{L}$ asumiremos que nos referimos a algún *lenguaje de la lógica cuantificacional de primer orden con identidad de signatura* $\mathcal{L}$. Para nuestros fines, es suficiente considerar cualquier lenguaje con una cantidad numerable de variables, una signatura finita o infinitamente contable, un conjunto funcionalmente completo de conectivas lógicas[9], al menos un símbolo de cuantificador y el símbolo de identidad; así como las reglas usuales de formación de términos y fórmulas.

El propósito del lenguaje formal es representar objetos y relaciones entre ellos: *grosso modo*, la función que juegan los términos es muy parecida a la que juegan los sustantivos, los pronombres y las frases nominales del lenguaje natural, es decir, nombran objetos; mientras que las fórmulas —siempre y cuando no cuenten con variables libres[10]— se asemejan a las oraciones declarativas o a los enunciados, los cuales establecen algo acerca de dichos objetos (sus propiedades o relaciones).

Los objetos sobre los que nos interesa hablar por medio de $\mathcal{L}$ forman un conjunto no-vacío[11] conocido como *dominio* (o *universo*) *de discurso*. Adicionalmente, dentro de este dominio, llamémoslo A, podemos distinguir ciertos elementos ($c^{\mathfrak{A}}$), así como definir operaciones ($f^{\mathfrak{A}}$) y relaciones ($R^{\mathfrak{A}}$) n-arias sobre él, lo cual nos permite generar una *estructura algebraico-relacional* $\mathfrak{A} = (A, (\mathrm{c_i}^{\mathfrak{A}})_{\mathrm{i}\in\mathrm{I}}, (f_j^{\mathfrak{A}})_{\mathrm{j}\in\mathrm{J}}, (R_{\mathrm{k}}{}^{\mathfrak{A}})_{\mathrm{k}\in\mathrm{K}})$. En este caso, el dominio de discurso A también es llamado *dominio de la estructura* $\mathfrak{A}$ (es decir, $\mathrm{dom}(\mathfrak{A}) = A$). Las $(\mathrm{c_i})_{\mathrm{i}\in\mathrm{I}}$ son los símbolos de constante de individuo de nuestro lenguaje, el

9 Un conjunto de conectivas es *funcionalmente completo* si y sólo si toda función veritativo-funcional *n*-aria puede ser expresada mediante los miembros del conjunto.

10 Es decir, variables que no se encuentran ligadas a un cuantificador (esto se precisará más adelante).

11 De considerar dominios vacíos, entre otras cosas, tendríamos que realizar una serie de modificaciones técnicas sobre las definiciones que veremos en unos momentos. No obstante, lo usual asumir que los dominios no son vacíos.

subíndice *i* que pertenece al conjunto de índices *I* nos permite distinguir unas constantes de otras. Hasta aquí ninguno de estos símbolos tiene un referente asociado. Esto cambia cuando se les interpreta en $\mathfrak{A}$. Así, $(c_i^{\mathfrak{A}})_{i\in I}$ se trata de los referentes de los símbolos en la estructura mencionada. De manera análoga, esto ocurre para los símbolos funcionales y relacionales (véase más abajo). Si los conjuntos de índices *I* y *J* fueran vacíos, entonces diríamos que es una *estructura relacional*; mientras que, si *K* fuera vacío, se trataría de una *estructura algebraica*. Como las funciones son un tipo especial de relación y como los elementos distinguidos pueden considerarse como funciones *0*-arias, en cierto sentido, todas las estructuras son simplemente relacionales. Aunque, por comodidad, seguiremos haciendo esta distinción entre constantes, funciones y relaciones. Cuando dos estructuras tienen la misma cantidad de constantes, funciones y relaciones, y la aridad de cada una de ellas coincide exactamente en ambas, entonces afirmamos que dichas estructuras son *similares*.

Con base en lo anterior, ahora podemos definir una *estructura de signatura* $\mathcal{L}$, o $\mathcal{L}$-estructura, como una estructura tal que cada uno de sus elementos distinguidos, funciones y relaciones *n*-arias son asignados a cada símbolo de constante, función y relación de aridad correspondiente en $\mathcal{L}$. Por esta razón, varios autores definen una $\mathcal{L}$-estructura $\mathfrak{A}$ sencillamente como un par $(\mathrm{dom}(\mathfrak{A}), \iota)$; donde ι es una función que asigna extensiones apropiadas a cada elemento de la signatura $\mathcal{L}$, en ocasiones llamada *función interpretación*. Más precisamente: si c_i es un símbolo de constante, entonces $\iota(c_i)$ es $c_i^{\mathfrak{A}} \in \mathrm{dom}(\mathfrak{A})$; si f_j es un símbolo de función de aridad $ar(j)$[12], entonces $\iota(f_j)$ es una operación $f_j^{\mathfrak{A}}$: $\mathrm{dom}(\mathfrak{A})^{ar(j)} \longrightarrow \mathrm{dom}(\mathfrak{A})$[13]; y, si R_k es un símbolo de relación $ar(k)$-ario, entonces $\iota(R_k)$ es $R_k^{\mathfrak{A}} \subseteq \mathrm{dom}(\mathfrak{A})^{ar(k)}$.

12 Esto es, con una cantidad $ar(j)$ de lugares o de entradas; es decir, la expresión $ar(j)$ afirma que la función *ar* asigna un número natural a *j*, el cual corresponde con la aridad de la función f_j.

De lo anterior, se sigue inmediatamente que cualesquiera dos $\mathcal{L}$-estructuras (con la misma signatura) son similares.

El término "estructura" aquí se aproxima en gran medida al empleado en el lenguaje ordinario para referir a un arreglo, disposición u organización de determinados objetos: estos pueden agruparse en un conjunto —el dominio de la estructura— sobre los que posteriormente se puede establecer una serie de relaciones y formar, por decir algo, diferentes tipos de orden, grupos, anillos, campos, módulos.

Con todo, el uso de tal término dentro del contexto de la teoría de modelos es relativamente reciente. En un principio, lo usual era hablar de "sistema matemático" para referirse a clases de objetos junto con ciertas relaciones definidas sobre ellos, es decir, lo que actualmente entendemos como estructura; pero en su momento la una y la otra no se consideraban equivalentes, sino que la segunda era algo subyacente a la primera (Mancosu *et al.*, 2009, pp. 420-1). Sin embargo, no podemos apresurarnos y sostener que el uso de dicha noción estaba propiamente estandarizado. Como Hodges (2018, pp. 439-40) advierte, podemos encontrar que Richard Dedekind utilizaba la palabra "sistema" para aludir tanto a colecciones como a estructuras (en sentido actual) y que David Hilbert hablaba de "sistemas de cosas" como colecciones de objetos reunidas de manera ordenada; incluso podemos observar casos en que mencionan "sistemas de interpretaciones" con un sentido similar. De hecho, "en la teoría de modelos el nombre 'sistema' perduró hasta que fue reemplazado por 'estructura' hacia finales de los cincuenta, parece ser que bajo la influencia de [Abraham] Robinson y [Nicolas] Bourbaki" (p. 440)[14].

13 En otras palabras, $f_j^{\mathfrak{A}}$ refiere a función que asigna $ar(j)$-adas de objetos a un único objeto de dom($\mathfrak{A}$), pues recordemos que el valor que toma *ar* en j es un número natural. Dicho de otra manera, $\iota(f_j) = f_j^{\mathfrak{A}}$, o sea, ι asocia fj con una función definida sobre la estructura $\mathfrak{A}$ de la aridad adecuada.

14 Mi traducción.

En fin, las $\mathcal{L}$-estructuras —por decirlo de algún modo— sirven como referencia una vez que los diferentes símbolos de $\mathcal{L}$ son interpretados; pero, en varias ocasiones, nos interesa también considerar interpretaciones alternativas para los mismos símbolos o, posiblemente, para otros. En cualquier caso, resulta pertinente dar cuenta de las distintas relaciones entre $\mathcal{L}$-estructuras, pero, para hacerlo, debemos detenernos antes sobre las relaciones que se dan entre las estructuras independientemente del lenguaje formal.

2.2. Algunas Relaciones entre Estructuras

Como señala Manzano, "hay esencialmente dos maneras de relacionar estructuras: *a*) desde fuera del lenguaje de primer orden; y *b*) por medio del lenguaje de primer orden" (1999, p. 102)[15]. En esta sección, revisaremos rápidamente lo concerniente con el inciso *a*). Desde luego, caracterizar estructuras y establecer relaciones entre ellas "desde fuera" de nuestros lenguajes de primer orden no significa que lo hagamos completamente con independencia del lenguaje. (Si así fuera, tendríamos que dejar esta sección en blanco.) El lenguaje que emplearemos para formular las siguientes definiciones forma parte del *metalenguaje*, del cual hablaremos más adelante.

Aunque el punto que pretende destacar Manzano en la cita de arriba es claro, conviene señalar que no es extraño asumir —como, de hecho, hacemos ahora— una *metateoría* enriquecida, entre otras cosas, con teoría de conjuntos, la cual resulta suficiente para representar casi cualquier estructura y relación matemática posible. Por esto, en rigor, muchas veces basta con ocupar tan solo una porción de la teoría de conjuntos en primer orden para expresar las relaciones entre estructuras. De esta manera, sean $\mathfrak{A} = (A, (c_{\mathrm{i}}^{\mathfrak{A}})_{\mathrm{i}\in\mathrm{I}}, (f_{\mathrm{j}}^{\mathfrak{A}})_{\mathrm{j}\in\mathrm{J}}, (R_k^{\mathfrak{A}})_{\mathrm{k}\in\mathrm{K}})$ y $\mathfrak{B} = (B, (c_i^{\mathfrak{B}})_{\mathrm{i}\in\mathrm{I}}, (f_{\mathrm{j}}^{\mathfrak{B}})_{\mathrm{j}\in\mathrm{J}}, (R_k^{\mathfrak{B}})_{\mathrm{k}\in\mathrm{K}})$ dos estructuras similares. Decimos que una función $h\colon A \longrightarrow B$ es un *homomorfismo débil* de $\mathfrak{A}$ en $\mathfrak{B}$ sii:

15 Mi traducción.

i) para cualquier $i \in I$, $h(c_i^{\mathfrak{A}}) = c_i^{\mathfrak{B}}$;
ii) para cualquier $j \in \mathrm{J}$ y cualesquiera $a_1,\ldots,a_{ar(j)} \in A$, $h(f_j^{\mathfrak{A}}(a_1,\ldots,a_{ar(j)})) = f_j^{\mathfrak{B}}(h(a_1),\ldots,h(a_{ar(j)}))$; y
iii) para cualquier $k \in K$ y cualesquiera $a_1,\ldots,a_{ar(k)} \in A$, $(a_1,\ldots,a_{ar(k)}) \in R_k^{\mathfrak{A}} \Longrightarrow (h(a_1),\ldots,h(a_{ar(k)})) \in R_k^{\mathfrak{B}}$.

Si el regreso también se cumple en iii), entonces se trata de un *homomorfismo fuerte*[16]. En pocas palabras, un homomorfismo es una función del dominio de una estructura a otra tal que preserva todas las relaciones y las funciones.

Por otro lado, un homomorfismo *fuerte* de $\mathfrak{A}$ en $\mathfrak{B}$ es un *encaje* (o una *inmersión isomorfa*) sii es inyectivo. En caso de que el encaje también sea sobreyectivo, entonces decimos que se trata de un *isomorfismo*[17], esto es que $\mathfrak{A}$ y $\mathfrak{B}$ son *isomorfas* (en símbolos $\mathfrak{A} \cong \mathfrak{B}$). En pocas palabras, esto significa que tanto $\mathfrak{A}$ como $\mathfrak{B}$ son estructuralmente idénticas.

Por último, tenemos que $\mathfrak{A}$ es una *subestructura* de $\mathfrak{B}$ (o que $\mathfrak{B}$ es una *extensión* de $\mathfrak{A}$), en símbolos $\mathfrak{A} \subseteq \mathfrak{B}$, sii la inyección canónica i: dom($\mathfrak{A}$) $\rightarrow$ dom($\mathfrak{B}$) es un encaje. De este modo, un encaje de $\mathfrak{A}$ en $\mathfrak{B}$ es simplemente un isomorfismo de $\mathfrak{A}$ sobre alguna subestructura de $\mathfrak{B}$.

2.3. Modelos y Teorías

Hasta ahora, hemos introducido nociones relacionadas con la teoría de modelos como "lenguaje" y "estructura"; pero, en primer lugar, ¿qué es un modelo?

La palabra "modelo" no suele emplearse de manera unívoca en el habla cotidiana y esta discrepancia de sentidos puede ser la causa de que el objeto de estudio de la teoría de modelos pueda resultar algo desconcertante en un inicio. Día a día, hablamos de modelo de arquitectura, modelo de vehículos, modelo a escala, modelo atómico, modelo del sistema solar, modelo de enseñanza, modelo de vestimenta, modelo de una pintura, modelo de modas, etcétera. Ya

16 Un homomorfismo de A en A se conoce como *endomorfismo*.
17 Un isomorfismo de A en A se conoce como *automorfismo*.

sea como un ejemplar, un diseño, un patrón, un molde, una figura, un esquema (incluso teórico); o como una copia, un análogo, un ejemplo, un símil; o incluso como un arquetipo, un prototipo, un parámetro, un punto de referencia, o algo del mundo; en algunos casos, ocupamos el término con el sentido de *i*) una representación; en otros casos, como *ii*) un representante; y en otros, como *iii*) lo representado. Aunque algunos autores tienden a asociar el concepto de modelo que aquí nos interesa con el último[18], quizá lo más conveniente sería no apresurar esta conclusión y reconocer que, en cierta forma, la noción dentro de este contexto efectivamente posee un poco de cada una[19].

Un *modelo* —en sentido modelo-teórico— se trata de una noción abstracta y relativa al lenguaje; se habla de "modelo de un enunciado" o "de un conjunto de enunciados"[20], entendido como una *interpretación del lenguaje* que los *hace verdaderos*. Por supuesto, establecer con precisión cuál es la naturaleza de esa interpretación, así como el modo en que le confiere verdad a los enunciados, es una tarea delicada que requiere un análisis cuidadoso si es que pretendemos estudiarla con todo el rigor matemático. Para ello, en la siguiente sección revisaremos rápidamente la concepción semántica de la verdad de Alfred Tarski, la cual impulsó el surgimiento de la noción contemporánea de *verdad en un modelo*[21].

18 Por ejemplo, Mosterín en su prefacio al libro de Manzano (1999), sostiene que el uso del término "modelo" en la teoría de modelos es más cercano al que ocupan los pintores o los fotógrafos cuando se refieren al objeto que están pintando o fotografiando.

19 Para un recorrido histórico sobre el origen y desarrollo de las distintas nociones de "modelo", véase Müller (2009).

20 Tal vez el caso más interesante es cuando el enunciado o los enunciados conforman una teoría axiomática.

21 Este es un ejemplo muy claro de cómo se le puede dar un tratamiento *extensional* a un concepto que tradicionalmente se ha considerado como *intensional*, en este caso, al concepto de verdad. Esto se debe, entre otras cosas, a que Tarski define un predicado de verdad en un lenguaje particular y, como consecuencia de su teorema de la indefinibilidad de la verdad, renuncia a la posibilidad de tener un predicado de verdad *absoluto* que se aplique a todos los lenguajes y que sea satisfecho por todos y cada uno de los enunciados verdaderos. Como se verá en unos momentos, lo que Tarski logra es definir un predicado de verdad extensionalmente adecuado para fragmentos bien

3. La Concepción Semántica de la Verdad

Durante la segunda mitad del siglo XIX, el desarrollo de diferentes teorías axiomáticas en áreas como el análisis, la geometría y el álgebra promovió gradualmente el estudio de sistemas formales desprovistos de significado[22]. Fue en esta época que —explícita o implícitamente— comenzaron a manifestarse nociones de carácter semántico dentro del discurso matemático y como elementos esen-

---

determinados del lenguaje, pero que no está asociado con una teoría completamente general que explique, en cada caso, por qué tal o cual enunciado es verdadero. Para los detalles, véase Tarski (1933; 1944).

Sin embargo, no todos los autores consideran a la verdad como concepto extensional (*e. g.*, Frege, 1892; 1918/1919), sino como intensional, en tanto que no es una noción cuyo significado cambie de contexto a contexto —o de lenguaje a lenguaje (como en el caso de Tarski)— y cuyo significado y extensión deberían permanecer invariantes bajo interpretación.

22 Originalmente, los sistemas axiomáticos eran *materiales*, es decir, sus términos básicos tenían asociados significados pretendidos o referencias específicas que permitían generar descripciones de sistemas concretos. En este sentido, por ejemplo, se concebían como teorías verdaderas sobre el espacio físico (en el caso de la geometría euclidiana) o sobre el movimiento de los cuerpos (en el caso de la mecánica newtoniana). No obstante, en el siglo XIX se desarrolló una concepción distinta de los sistemas axiomáticos, que prescindía del significado o la referencia de los términos primitivos y dejaba abierta la posibilidad de asignarles diferentes contenidos. Desde esta perspectiva, los sistemas axiomáticos se denominan *formales*, y los axiomas no son verdaderos ni falsos hasta que se les asigna un significado determinado a los términos primitivos. De esta manera, el paso de la axiomática material a la formal permitió concentrarse más en la estructura del sistema y en las relaciones entre los términos que propiamente en su contenido.

Esta tendencia por *abstraer* el contenido de las representaciones lingüísticas y por *establecer* desde un principio definiciones de las propiedades internas o estructurales de los objetos de estudio de la matemática responde a varias razones. Entre ellas, destacan la introducción de metodologías no-constructivas en matemáticas y la necesidad de justificar o fundamentar herramientas matemáticas con independencia de consideraciones espaciotemporales. Sin embargo, quizá una de las motivaciones más relevantes, promovida principalmente por Hilbert, radicaba en la importancia de estudiar las propiedades internas de los sistemas axiomáticos. Esto dio lugar al estudio metamatemático de los sistemas formales, cuyo propósito, entre otros, era establecer la consistencia de los sistemas. De este modo, al tratar los términos de una teoría como carentes de significado, era posible asignarles contenidos que podían asociarse con los términos de otras teorías y, con ello, trazar conexiones entre ellas. En ocasiones, esto permitía demostrar la consistencia relativa de un sistema axiomático en función de la consistencia de otro más fuerte. Para una breve historia del método axiomático, véase Cassini (2009).

ciales en la demostración de cuestiones de naturaleza *meta*matemática (*e. g.*, consistencia, independencia, completud). No obstante, en aquel tiempo, el tratamiento de estas nociones era ofrecido de manera informal y con un sentido intuitivo. De esta manera, su relevancia progresiva en las investigaciones de la matemática, así como la carencia de rigor en sus definiciones impulsaron el surgimiento de los estudios semánticos de los lenguajes formales[23].

En medio de este panorama, Tarski publicó "El concepto de verdad en lenguajes formalizados" (1933) como respuesta a la necesidad de una definición satisfactoria de una de estas nociones semánticas, a saber, la noción de *verdad*. En este texto, Tarski —inspirado por la concepción clásica de *verdad por correspondencia*[24]— consiguió ofrecer de manera exitosa una definición "materialmente adecuada" y "formalmente correcta[25]" de *verdad en un lenguaje* (en adelante, *verdad-en-ℒ*), la cual, entre otras cosas, evitaba apelar a conceptos semánticos que no hubieran sido reducidos previamente a otros conceptos (1933, p. 152-3). Este acercamiento a la noción de verdad es lo que denominaría como "la concepción semántica de la verdad" (*e. g.*, Tarski, 1944, p. 345). Cabe señalar que, al tratarse de la definición de un predicado sobre enunciados —los cuales están expresados en un cierto lenguaje particular—, el pro-

23 Véase Mancosu *et al*. (2009).

24 Tarski dice: "Nos gustaría que nuestra definición le hiciera justicia a las intuiciones que se adhieren a la *concepción clásica aristotélica de verdad*" (1944, p. 342, traducción propia). Tarski está pensando aquí en las palabras de Aristóteles en el libro de la Metafísica: "Decir de lo que es que no es, o de lo que no es que es, es falso; mientras que decir de lo que es que es, o de lo que no es que no es, es verdadero" (1994, p.198).

25 Con "materialmente adecuada", Tarski quiere decir que la definición especifica las condiciones para determinar efectivamente cuándo algo cae bajo ella (véase Tarski, 1944, p. 341). Por su parte, "formalmente correcta" quiere decir que la definición especifica los conceptos involucrados en la definición y las reglas a las que debe apegarse (véase Tarski, 1944, p 342).

En el caso particular de la definición formalmente correcta de verdad, la adecuación material puede formularse a partir de la condición dada por la llamada *Convención T*, la cual, *grosso modo*, requiere que la definición tenga como consecuencia todas las instancias del "esquema T" y que la clase de enunciados verdaderos sea subconjunto de la clase de enunciados (véase Tarski, 1933, p. 187-8). Para un estudio detallado sobre estas cuestiones, véase David (2008) y Feferman (2008).

pósito de Tarski era dar sencillamente con una versión de "verdad en *un* lenguaje específico". Por otro lado, por la naturaleza de los lenguajes naturales[26], esta clase de definición creía que no era posible para lenguajes de este tipo[27]. Asimismo, es importante notar que la definición tarskiana de verdad no pretende determinar qué es *la* verdad en un sentido amplio ni absoluto —la intención *no* es establecer *qué es lo que realmente hay* o *cómo es genuinamente el mundo*—, sino que busca ofrecer con claridad y precisión condiciones bajo las cuales podemos determinar la verdad de los enunciados *en un cierto lenguaje*.

Como podemos apreciar, el alcance de la definición, en algún sentido, resulta bastante modesta: la definición se encuentra restringida a los lenguajes formales, cuyas fórmulas pueden especificarse con exactitud —por las características mencionadas en la sección anterior— y que están libres de expresiones que permiten la autorreferencia de sus propiedades semánticas; esto último es lo que nos obliga a definir la noción de verdad por separado en otro lenguaje. "Por esta razón, —dice Tarski— cuando investigamos el lenguaje de una ciencia deductiva formalizada, siempre debemos distinguir claramente entre el lenguaje *acerca* del que hablamos y el lenguaje *en* el que hablamos" (1933, p. 167)[28]. Estos lenguajes son los que comúnmente conocemos como *lenguaje-objeto y metalenguaje*, respectivamente.

Otro aspecto que resulta pertinente destacar es que la definición de verdad-en-$\mathcal{L}$ está diseñada para lenguajes "plenamente interpretados", es decir, lenguajes cuyos símbolos no-lógicos ya tie-

26 A diferencia de los lenguajes formales, Tarski concluye que no es posible ofrecer una definición de verdad con las características deseadas para el caso de los lenguajes naturales. En pocas palabras, esto se debe, en primer lugar, a que este tipo de lenguajes contiene, además de sus enunciados, los nombres de sus enunciados, así como expresiones de carácter semántico, lo cual permite la autorreferencia y genera paradojas del tipo mentiroso; y en segundo, a que estos lenguajes, al no estar acabados ni tener límites claros, no son susceptibles de que sus expresiones sean determinadas estructuralmente en su totalidad. Véase Tarski (1933, pp. 154-65).

27 Obviamente, como se mencionó al principio de este trabajo, hay autores que han desafiado esta idea (véase la nota 3).

28 La traducción es mía.

nen asociados un significado de antemano y cuyos enunciados son todos o bien verdaderos, o bien falsos por este motivo. Además de esto, estos lenguajes cuantifican irrestrictamente sobre todo lo que hay. Con todas estas consideraciones en mente es que Tarski, finalmente, se dispone a construir la definición de verdad-en-$\mathcal{L}$ para un lenguaje en específico, a saber, el lenguaje de la teoría de clases; y, después, para lenguajes arbitrarios de tipo finito[29].

No obstante, a pesar de estos requisitos, el matemático polaco aún no podía dar inmediatamente la codiciada definición. En palabras de Mancosu *et al*. (2009):

> Idealmente, a uno le gustaría proceder con la definición de verdad por recursión sobre la complejidad de enunciados. Desafortunadamente, en vistas de que los enunciados en general no son obtenidos a partir de otros enunciados, sino, más bien, a partir de fórmulas (las cuales, en general, pueden contener variables libres), una definición recursiva de "enunciado verdadero" no podía ofrecerse directamente. (p. 440, traducción propia)

Si bien es cierto que esto ocurre con los enunciados, el caso de las fórmulas es distinto: éstas sí pueden construirse recursivamente a partir de otras fórmulas de menor complejidad. Por este motivo, el matemático se ve en la necesidad de recurrir en un primer momento a la noción de *satisfacción*.

Lamentablemente, por cuestiones de espacio, no nos será posible formular aquí estas definiciones. Esto no representa un problema para nosotros, pues estas nociones —como comentamos— son demasiado "específicas" para nuestros fines. Hasta ahora, más que presentar propiamente su definición, lo que buscábamos era destacar más bien el análisis que le permitió al lógico polaco concebir una primera versión de la definición de verdad. Además, en unos momentos más, compensaremos este vacío con la definición de otro concepto de "carácter relativo" y que Tarski (1933) men-

29 En realidad, Tarski (1933) también explora otras definiciones, pero no son relevantes para nuestros fines.

ciona más adelante, el cual "juega un papel mucho mayor que el concepto absoluto de verdad y que lo incluye como caso especial. Este es el concepto de *enunciado correcto o verdadero en un dominio individual a*" (p. 199, traducción propia).

4. La Noción de Verdad en la Teoría de Modelos

Como mencionamos al inicio de la sección pasada, dentro de la práctica matemática era común considerar que las teorías no tuvieran un significado asociado de antemano, de tal modo que la terminología no-lógica que aparecía en ellas podía ser variable. Desafortunadamente, la definición tarskiana de verdad-en-$\mathfrak{L}$ estaba diseñada exclusivamente para lenguajes plenamente interpretados[30]. Esta restricción fue la que eventualmente motivaría la búsqueda de una nueva definición de verdad, una que mantuviera el significado de los lenguajes por separado y que permitiera distintas interpretaciones de los símbolos no-lógicos presentes en ellos. De esta manera, "la noción de modelo es definida en términos de funciones que le asignan objetos teórico-conjuntistas a estas constantes [no-lógicas]. Las nociones de significado y referencia se desvanecen dentro de la teoría de conjuntos" (Hodges, 2009, p. 483)[31].

Aunque la definición contemporánea de verdad suele rastrearse justamente al pasaje con el que cerramos la sección anterior, la versión que utilizamos hoy en día está basada en la ofrecida por Tarski y Robert Vaught en 1956[32]. A continuación, formularemos la noción que realmente nos interesaba desde un inicio, a saber, la de *verdad en un modelo* (o *verdad-en-*$\mathfrak{M}$). Para ello, partiremos desde el punto en que interrumpimos la exposición de arriba y, con base en la definición de satisfacción para fórmulas, definiremos esta nueva noción de verdad como lo hacemos en la actualidad. Como hemos dicho, esperamos ofrecer una definición recursiva, primero, para fórmulas y, luego, para enunciados.

30 Véase Hodges (2018; 2022, secc. 1).
31 La traducción es mía.
32 Véase Hodges (2022, secc. 3).

Sea $\mathfrak{A} = (A, (c_i^{\mathfrak{A}})_{i\in I}, (f_j^{\mathfrak{A}})_{j\in J}, (R_k^{\mathfrak{A}})_{k\in K})$ una $\mathcal{L}$-estructura y s: *Var* $\longrightarrow$ dom($\mathfrak{A}$) una *asignación de variables* (o simplemente *asignación*); donde *Var* es el conjunto de variables. (*Grosso modo*, el papel de la asignación es asociar a las variables que ocurren libres en fórmulas con denotaciones.) Extendemos recursivamente la función s a una nueva función $\check{s}$: *Ter* $\longrightarrow$ dom($\mathfrak{A}$), la *función denotación*, para abarcar el conjunto *Ter* de términos de $\mathcal{L}$. Sean t y $t_1,\ldots,t_n$ términos de *Ter*[33], y sea ι una función interpretación como la definida previamente[34], entonces:

i) para cualquier $x \in Var$, $\check{s}(t) = s(x)$ si $t = x$;
ii) para toda $i \in I$, $\check{s}(t) = \iota(c_i) = c_i^{\mathfrak{A}}$ si $t = c_i$; y
iii) para toda $j \in J$, $\check{s}(t) = (\iota \circ f_j)(\check{s}(t_1),\ldots,\check{s}(t_n)) = f_j^{\mathfrak{A}}(\check{s}(t_1),\ldots,\check{s}(t_n))$ si $t = f_j(t_1,\ldots,t_n)$ y $n = ar(j)$.

Con base en la función denotación, finalmente podemos ofrecer la definición recursiva de satisfacción. Así, decimos que una $\mathcal{L}$-estructura $\mathfrak{A}$ *satisface* una fórmula φ con s, escrito $\vDash_{\mathfrak{A}} \varphi[s]$, sii[35]:

iv) $\check{s}(t_1) = \check{s}(t_2)$ si φ es de la forma $\ulcorner t_1 = t_2 \urcorner$;
v) para cualquier $k \in K$, $(\check{s}(t_1),\ldots,\check{s}(t_n)) \in \iota(R_k) = R_k^{\mathfrak{A}} \subseteq \text{dom}(\mathfrak{A})^n$ si φ es de la forma $\ulcorner R_k t_1,\ldots,t_n \urcorner$ y $n = ar(k)$[36];
vi) $\nvDash_{\mathfrak{A}} \psi[s]$ si φ es de la forma $\ulcorner \neg\psi \urcorner$;
vii) o bien $\nvDash_{\mathfrak{A}} \psi[s]$, o bien $\vDash_{\mathfrak{A}} \phi[s]$, si φ es de la forma $\ulcorner \psi \to \phi \urcorner$; y

33 Notemos que t y $t_1,\ldots,t_n$ son metavariables.
34 Véase la sección "Lenguajes formales y estructuras".
35 Como mencionamos anteriormente, para nuestros fines basta considerar un lenguaje con uno de los cuantificadores usuales y un conjunto completo de conectivas lógicas. Para esta exposición, sencillamente consideraremos el cuantificador universal y el conjunto completo $\{\to, \neg\}$.
36 Esta definición también aplica cuando $n=0$. En este caso, el papel que juega el símbolo de relación *0*-ario es muy parecido al de las variables proposicionales. Para los detalles, véase van Dalen (2013, p. 55).
37 La función s(x|d) es una función "x-variante", es decir, una función que es exactamente como s excepto que en x toma el valor d.

viii) para cualquier $d \in \text{dom}(\mathfrak{A})$, $\vDash_{\mathfrak{A}} \varphi[s(x|d)]$ si φ es de la forma $\ulcorner \forall x\, \psi \urcorner$ [37].

Una vez que tenemos la definición de satisfacción, podemos obtener fácilmente la que corresponde con la verdad: sea $\mathfrak{A}$ una $\mathcal{L}$-estructura y φ una fórmula de $\mathcal{L}$, decimos que φ es *verdadera en* $\mathfrak{A}$ (en símbolos $\vDash_{\mathfrak{A}} \varphi$) sii $\vDash_{\mathfrak{A}} \varphi[s]$ para toda asignación s. En particular, si φ es una fórmula cerrada, o sea, un enunciado, entonces o bien cualquier asignación s la satisface (*i. e.,* es verdadera en $\mathfrak{A}$), o bien ninguna lo hace (*i. e.*, es falsa en $\mathfrak{A}$)[38]. Precisamente, si ocurre lo primero, también podemos hablar de que $\mathfrak{A}$ es un *modelo* (o una *realización*) de φ[39].

Ciertamente, nos ha llevado tiempo llegar a esta definición, pero no tardaremos en convencernos de que se ha tratado de una buena inversión. En efecto, a partir del concepto de modelo podemos definir de manera inmediata una serie de nociones derivadas y que juegan un papel protagónico en varias discusiones filosóficas. Esto es justamente lo que haremos en las próximas secciones[40].

5. Algunas Nociones Lógicas en Versión Modelo-Teórica

Probablemente, unas de las primeras nociones que podemos obtener de manera natural a partir de las definiciones de satisfacción y

38 No se pueden cumplir ambas porque $\text{dom}(\mathfrak{A}) \neq \emptyset$.

39 En adelante, procuraremos reservar las letras σ y τ para enunciados y el resto de las letras griegas minúsculas para fórmulas en general. De manera análoga, las letras Σ y T procuraremos emplearlas para conjuntos de enunciados y el resto de las letras mayúsculas para conjuntos de fórmulas en general.

40 Es importante aclarar que la anterior no es la única manera de definir la semántica de los lenguajes de orden uno, nosotros estamos ofreciendo una de las versiones tarskianas que emplea la teoría de conjuntos en la metateoría. Tarski originalmente intentaba hacerlo a la Frege cambiando terminología no-lógica por variables y cuantificando sobre ellas en algún orden superior adecuado (véase Hodges, 2018). Robinson también ofreció una manera de hacerlo con la añadidura de una cantidad suficiente de términos para nombrar a todos los objetos del dominio de cuantificación. Incluso existe una manera de mezclar las versiones tarskiana y la robinsoniana, véase Button y Walsh (2018, cap. 1).

de modelo son aquellas que sencillamente las amplían hasta abarcar no sólo fórmulas individuales, sino conjuntos de ellas: sea Γ un conjunto de fórmulas y $\mathfrak{A}$ como antes, decimos que $\mathfrak{A}$ *satisface* Γ con s ($\vDash_{\mathfrak{A}} \Gamma[s]$) sii para toda $\gamma \in \Gamma$, $\vDash_{\mathfrak{A}} \gamma[s]$; de manera similar, $\mathfrak{A}$ es un *modelo* de Γ ($\vDash_{\mathfrak{A}} \Gamma$) sii, para toda $\gamma \in \Gamma$, $\vDash_{\mathfrak{A}} \gamma$. En caso de que Γ sea unitario, entonces obtenemos las definiciones originales. Cuando resulta que efectivamente hay alguna $\mathcal{L}$-estructura y asignación que satisface Γ, entonces decimos que es *satisfacible*[41].

Sabemos que si $\vDash_{\mathfrak{A}} \varphi[s]$ con cualquier s, entonces φ es verdadera en $\mathfrak{A}$, ¿y si ahora nos preguntáramos cuando φ es verdadera en *cualquier* $\mathcal{L}$-estructura? En ese caso, alcanzaríamos uno de los conceptos más paradigmáticos de la lógica en su versión modelo-teórica, a saber, la *verdad lógica* o *validez universal* (en símbolos, $\vDash \varphi$): una fórmula es *lógicamente verdadera* sii es verdadera en cualquier $\mathcal{L}$-estructura. Sobra señalar la importancia de este concepto —muchos autores incluso lo consideran el objeto de estudio de la lógica—[42], desgraciadamente, detenernos sobre él nos alejaría de los propósitos de este capítulo.

Una vez que tenemos esta noción, cabe introducir otro par de nociones que aparecen con frecuencia en lógica en sus versiones modelo-teóricas, a saber, la *consecuencia* y la *equivalencia lógicas*. Más precisamente: φ *implica lógicamente* ψ (en símbolos, $\varphi \vDash \psi$) sii, para toda $\mathfrak{A}$ y s, si $\vDash_{\mathfrak{A}} \varphi[s]$, entonces $\vDash_{\mathfrak{A}} \psi[s]$; por su parte, decimos que φ y ψ son *lógicamente equivalentes* ($\varphi \dashv\vDash \psi$) sii tanto $\varphi \vDash \psi$ como $\psi \vDash \varphi$[43]. Estas definiciones no pretenden describir cómo es que evaluamos intuitivamente las nociones lógicas, sino que preten-

41 Por simplicidad, en lo sucesivo, nos suscribiremos a la práctica común de considerar que las fórmulas son de longitud finita. Razón por la cual, asumiremos que cualquier conjunto de fórmulas puede expresarse —con el uso de las conectivas— simplemente como una sola fórmula (posiblemente larga, pero finita).

42 Por ejemplo, Frege (1918/1919) y Quine (1973, cap. 4).

43 De nuevo, podríamos pensar en ampliar estas nociones a conjuntos de fórmulas, en uno o ambos lados de los símbolos que acabamos de definir. Sin embargo, no consideraremos estos casos por lo mencionado en la nota 41 y porque tradicionalmente no pensamos en los casos de multi-conclusión. Aunque hay acercamientos que sí lo hacen, por ejemplo, Shoesmith y Smiley (1978).

de ofrecernos una reconstrucción teórica basada en nociones más precisas (matemáticas). En palabras de Gómez-Torrente:

> Pero una teoría filosófica de una noción intuitiva como la de consecuencia lógica no tiene por qué tener como único objetivo la explicación de su significado intuitivo. El principal objetivo de Tarski al proponer su definición no es éste, sino el objetivo de, caracterizar *extensionalmente* la noción intuitiva de consecuencia lógica; en términos de un aparato de conceptos mejor comprendidos que el intuitivo de consecuencia lógica. Consideremos la noción intuitiva de "arco iris". Esta noción tiene un significado intuitivo para la gran mayoría de la gente; prácticamente todo el mundo puede asociar ideas intuitivas con ella, pero la cantidad de gente familiarizada con la base óptica y física en general del arco iris es relativamente escasa. Supongamos que el físico *define* "arco iris" por medio de una serie de conceptos físicos técnicos, por ejemplo "arco de colores prismáticos causado por la refracción y reflexión de los rayos del sol en las gotas de lluvia". Con esta definición, el físico no intenta explicar el significado de la noción intuitiva de arco iris, sino, en parte, ofrecer una caracterización en términos menos intuitivos, mejor comprendidos, que se aplique *exactamente* a los arcos iris (y que de este modo explique o ilumine el hecho de que haya arcos iris). (Podría disputarse que un objetivo del físico o del científico en general sea asegurar la coextensionalidad de conceptos intuitivos y conceptos técnicos. En muchos casos, la investigación puede revelar que el concepto intuitivo es problemático, quizá incluso contradictorio, y, por tanto, su sustitución por un concepto técnico coextensional es ociosa. Pero en muchos otros casos la ciencia parece hacer justamente caracterizaciones extensionales de conceptos intuitivos en términos de conceptos técnicos; un ejemplo famoso en la bibliografía filosófica es la identidad extensional entre agua y H_20.) Como otras definiciones científicas, la definición de consecuencia lógica de Tarski tiene como uno de sus objetivos la caracterización extensional de una noción intuitiva en términos de conceptos técnicos, mejor comprendidos que el concepto intuitivo de consecuencia lógica. (2000, p. 43-4)

Si seguimos pensando diferentes escenarios en que se presentan casos de satisfacción, de repente podemos desprender otros conceptos

bastante útiles. Supongamos que ahora tenemos una $\mathcal{L}$-estructura $\mathfrak{A}$ y una fórmula abierta φ, cuyas únicas variables libres son $x_{l_1},\ldots,x_{m_k}$, representado por $\varphi(x_{l_1},\ldots,x_{m_k})$, y que $\vDash_{\mathfrak{A}} \varphi(x_{l_1},\ldots,x_{m_k})[s]$ para alguna s. Podemos apreciar que si reunimos todas esas asignaciones, cualquier secuencia $(a_1,\ldots,a_k)$ de elementos de dom($\mathfrak{A}$) con $a_1,\ldots,a_k$ en los lugares $l_1,\ldots,m_k$, respectivamente, satisfaría $\varphi(x_{l_1},\ldots,x_{m_k})$ en $\mathfrak{A}$[44]; en otras palabras, todo segmento inicial $(a_1,\ldots,a_k)$ de cualesquiera de tales asignaciones satisfaría $\varphi(x_{l_1},\ldots,x_{m_k})$ en $\mathfrak{A}$, para abreviar: $\vDash_{\mathfrak{A}} \varphi[a_1,\ldots,a_k]$. De este modo, dicha fórmula sería satisfecha por todos los elementos de algún subconjunto de dom($\mathfrak{A}$)^k, esto es, por los elementos de alguna relación k-aria definida sobre dom($\mathfrak{A}$)[45]. Desde aquí, es fácil plantear un escenario con $k = 1$, es decir, donde φ tiene una única variable libre x tal que $\vDash_{\mathfrak{A}} \varphi[a]$. En esta ocasión, podemos observar que φ es satisfecha por todos los elementos de algún subconjunto de dom($\mathfrak{A}$). Por último, podemos contemplar el caso en que $\vDash_{\mathfrak{A}} \varphi[a]$ y a es única, o sea, $\mathfrak{A}$ satisface $\varphi(x)$ con un solo objeto de dom($\mathfrak{A}$).

A partir de lo de arriba, hemos conseguido una nueva noción, esta es, la noción de *definibilidad en primer orden* (*sin parámetros*)[46] *de una relación en una estructura*[47]. Concretamente: una relación k-aria X es *definible en una estructura* $\mathfrak{A}$ sii existe una fórmula $\varphi(x_{l_1},\ldots,x_{m_k})$ tal que para cualesquiera $a_1,\ldots,a_k \in$ dom($\mathfrak{A}$), $X = \{(a_1,\ldots,a_k) \in \text{dom}(\mathfrak{A})^k : \vDash_{\mathfrak{A}} \varphi[a_1,\ldots,a_k]\}$.

Ahora, recordemos que una fórmula es verdadera en $\mathfrak{A}$ siempre y cuando es satisfecha en $\mathfrak{A}$ con toda asignación de variables, esto no es otra cosa que decir que su *clausura* (o *cerradura*) *universal*

44 De hecho, es probable que varias secuencias distintas coincidan exactamente con las a_i en $(a_1,\ldots,a_k)$.

45 Esta relación no necesariamente debe corresponderse con algún símbolo relacional de $\mathcal{L}$.

46 Como advertimos, nosotros asumimos que $x_{l_1},\ldots,x_{m_k}$ se trataban exactamente de las variables libres en $\varphi(x_{l_1},\ldots,x_{m_k})$; sin embargo, esta notación suele no ser tan restrictiva: $x_{l_1},\ldots,x_{m_k}$ pueden aparecer o no en φ e incluso podría haber variables libres en φ que no están entre $x_{l_1},\ldots,x_{m_k}$.

47 Una vez más, recordemos que las funciones y las constantes se pueden considerar como un tipo particular de relación.

es verdadera en $\mathfrak{A}$. Si consideramos esto y retomamos lo de arriba, podemos pensar en definir o describir relaciones, funciones y objetos distinguidos mediante las fórmulas de algún lenguaje $\mathcal{L}$ para después cerrarlas universalmente, con el propósito de caracterizar la clase de las $\mathcal{L}$-estructuras que las hace verdaderas. Si bien muchas veces nos puede interesar la caracterización de clases con dos o más $\mathcal{L}$-estructuras, muchas otras, nos pueden interesar clases unitarias de ellas[48]. En este caso, solemos fijar de antemano un significado preferido a la terminología no-lógica (muy parecido a lo que Tarski hacía con los lenguajes plenamente interpretados al definir la verdad en $\mathcal{L}$) e interpretamos los enunciados en esas estructuras predilectas. Es usual referirse a estas últimas como los *modelos pretendidos* (o *estándar*) de una clase de enunciados o —como veremos a continuación— de una teoría. Desgraciadamente, como se verá más adelante, no es posible caracterizar de manera única a los modelos pretendidos de tamaño infinito, lo cual quiere decir que existen modelos estructuralmente distintos de ellos (no-isomorfos) que realizan exactamente los mismos enunciados. Estos últimos son los que comúnmente se conocen como *modelos no-estándar*.

En primera instancia, esto puede parecer catastrófico para los lenguajes de orden uno (y, hasta cierto punto, lo es), pero —como veremos en un momento— también puede representar una ventaja por encima de los lenguajes de orden superior.

6. Teorías Formales

Hasta este momento, habíamos considerado fórmulas y conjuntos de fórmulas de manera arbitraria, pero "la parte más interesante de la teoría de los modelos comienza cuando nos adentramos a la investigación sobre conjuntos de enunciados que conforman una teoría" (Manzano, 1999, p. 115, traducción propia), pues "la teoría de modelos clásica se refiere al estudio en general de los modelos de las teorías

48 De conformidad con la clasificación de Shapiro (1991), las primeras son las llamadas *teorías algebraicas*; las segundas se tratan de las *no-algebraicas*.

formales y sus relaciones con los lenguajes de los que son interpretaciones" (Amor, 2008, p. 86). Dicho esto, y con las observaciones que hemos realizado hasta aquí, resulta pertinente definir a continuación la noción de *teoría formal*, así como otras relacionadas.

Sea T un conjunto de *enunciados* expresados en algún lenguaje $\mathcal{L}$, T es una *teoría* sii para cualquier enunciado σ, $T \vDash \sigma \Longrightarrow \sigma \in T$. En otras palabras, T es un conjunto de enunciados de $\mathcal{L}$ cerrado bajo la relación de consecuencia lógica; por lo tanto, T es el conjunto de sus *consecuencias* ($Cn(T)$), esto es, $T = \{\sigma : T \vDash \sigma\} = Cn(T)$. Cabe señalar, que algunos autores prefieren definir esta noción con base en otra relación, a saber, la de *deducibilidad* ($\vdash$); sin embargo, esta diferencia no es muy significativa para este trabajo, ya que ambas relaciones en primer orden son extensionalmente equivalentes por los teoremas de correctud y completud de cualquier cálculo deductivo "adecuado"[49] con la semántica usual[50].

Como adelantábamos al final de la sección pasada, muchas veces nos interesa estudiar la clase de modelos determinados por un cierto conjunto de enunciados T, llamemos $\mathcal{K}$ al conjunto de modelos de T (o sea, $\mathcal{K} = \text{Mod}(T)$). En estos casos, no esperamos dar por lo general el conjunto infinito mismo de todos los enunciados que deseamos que se cumplan[51] en $\mathcal{K}$ (pocas veces lo sabemos en realidad)[52], sino que esperamos poder ofrecer un conjunto recursivo Σ de enunciados —cerrado bajo consecuencia lógica— que nos permita obtener el conjunto de dichos enunciados verdaderos en $\mathcal{K}$. Sea $Th(\mathcal{K})$ tal conjunto; es decir, el conjunto de todos los enunciados verdaderos en cada elemento de $\mathcal{K}$[53]. En este caso,

49 "Adecuado" aquí es empleado informalmente. Nos referimos a cualquier cálculo deductivo presentado en algún buen libro de texto de lógica (*e. g.*, Mendelson, 1997; Enderton, 2001).

50 Por lo mismo, las nociones de validez y satisfacibilidad también coinciden exactamente con las de teoremicidad y consistencia, respectivamente.

51 Afirmamos con seguridad que es infinito porque el conjunto de verdades lógicas es de por sí infinito.

52 Podemos conocerlo solamente en el caso de las teorías completas de estructuras finitas.

53 Más formalmente: $Th(\mathcal{K}) = \{\sigma : \vDash_{\mathfrak{A}} \sigma$, para toda $\mathfrak{A} \in \mathcal{K}\}$. El conjunto $Th(\mathcal{K})$ se conoce como la *teoría de una clase de estructuras* $\mathcal{K}$ ("*Th*" por *theory*, en inglés). Véase, por ejemplo, Manzano (1999. p. 119) y Enderton (2001, p. 227).

tendríamos que T es una *teoría recursivamente axiomatizable* sii hay algún conjunto recursivo de enunciados Σ (*axiomas* de T) tal que $T = Cn(\Sigma)$. Lamentablemente, encontrar esa Σ para cualquier T no siempre es posible.

Para ilustrar, supongamos que nos interesa una cierta clase de modelos $\mathcal{K}$ y que proponemos un conjunto de axiomas que describe las características de los elementos de esa clase. ¿Tenemos la garantía de haber considerado una cantidad suficiente de axiomas? Dicho de otro modo, ¿sabemos que cualquier enunciado verdadero en cada miembro de $\mathcal{K}$ siempre se sigue lógicamente de los axiomas? La respuesta es no. La razón es que *no* toda teoría T es *completa para la negación*, esto es, no siempre es el caso que, para cualquier φ en $\mathcal{L}$, o bien $T \vDash \varphi$, o bien $T \vDash \neg\varphi$[54]. Sin embargo, cualesquiera dos modelos de una teoría hacen verdaderos exactamente a los mismos enunciados *(i. e.,* son elementalmente equivalentes —véase más adelante—) sii la teoría es completa[55].

Hay unas cuantas propiedades más que nos interesa definir en relación con las teorías y sus modelos, pero hacerlo requiere considerar a su vez las relaciones entre estructuras que presentamos anteriormente.

7. Algunas Relaciones entre $\mathcal{L}$-estructuras y Otras Nociones Modelo-Teóricas

Al principio del capítulo, mencionamos la función de la teoría de modelos como una disciplina semántica que vincula a los elementos de un lenguaje formal —símbolos y secuencias de símbolos— con los objetos correspondientes en las estructuras que representan; el puente entre ellos —como hemos visto— es la noción de verdad

54 O bien, como consecuencia de la definición de teoría, $\varphi \in T$, o bien $\neg\varphi \in T$.

55 Si suponemos que no es completa, entonces habría un enunciado independiente, que no pertenece ni él ni su negación a la teoría, lo cual quiere decir que ninguno está implicado lógicamente por la teoría (dado que la teoría está cerrada bajo consecuencia lógica). Esto tiene como consecuencia que existen al menos dos modelos de T, uno donde φ es verdadera y otro donde es falsa. Por lo tanto, no hacen verdaderos a los mismos enunciados.

que introdujimos en términos de modelo (Manzano, 1999, p. 35). En esta sección, reuniremos lo considerado hasta ahora para establecer relaciones entre las estructuras y aquello que podemos decir sobre ellas por medio de nuestros lenguajes de primer orden (o sea, el inciso *b*) que mencionamos en la sección sobre relaciones entre estructuras). Por desgracia, pronto nos daremos cuenta de que hay ciertos límites insuperables respecto de lo que podemos expresar a través de estos lenguajes y que, en última instancia, puede generar una serie de problemas importantes en discusiones al interior de diferentes áreas de la filosofía; por ejemplo, relacionadas con epistemología, metafísica y filosofía del lenguaje, entre otras. (Esto quedará más claro cuando estudiemos la indeterminación de la referencia a partir de la teoría de modelos en la última sección.)

Antes de continuar, resulta conveniente presentar un resultado de importancia para la exposición, este es, el famoso *teorema del homomorfismo*. Este teorema explícitamente nos permite estudiar la relación entre estructuras apelando a la noción de verdad: sea h un homomorfismo de $\mathfrak{A}$ en $\mathfrak{B}$ y s una asignación, entonces $\vDash_{\mathfrak{A}} \varphi[s]$ sii $\vDash_{\mathfrak{B}} \varphi[h \circ s]$[56].

A partir del teorema del homomorfismo, es fácil observar que cualesquiera dos estructuras isomorfas realizan exactamente los mismos enunciados de primer orden. En otras palabras, dos estructuras isomorfas con la misma signatura son indiscernibles dentro de las teorías de primer orden. Esto no resulta muy sorprendente, pues "las estructuras isomorfas se parecen de cualquier manera 'estructural'; no sólo satisfacen los enunciados de primer orden, sino que también satisfacen los de segundo orden (y superiores)" (Enderton, 2001, p. 146). Sin embargo, la situación conversa no siempre se cumple: no es el caso que cualesquiera dos estructuras que realizan los mismos enunciados son isomorfas. Esto nos sugiere que la una y la otra no se tratan de nociones equi-

56 Por simplicidad, hemos omitido algunos detalles en esta formulación. Para una versión más precisa, véase Enderton (2001, pp. 143-4).

valentes y que, por consiguiente, deben manejarse por separado: decimos que dos estructuras $\mathfrak{A}$ y $\mathfrak{B}$ son *elementalmente equivalentes* ($\mathfrak{A} \equiv \mathfrak{B}$) sii para cualquier enunciado σ de $\mathcal{L}$, $\vDash_{\mathfrak{A}} \sigma \Leftrightarrow \vDash_{\mathfrak{B}} \sigma$[57]. Aun así, en ocasiones, resulta conveniente hablar de teorías tales que cualesquiera de sus modelos son únicos, salvo isomorfismo; este es el caso de las llamadas *teorías categóricas*. Lamentablemente, las únicas teorías categóricas expresadas en primer orden son las teorías completas de estructuras finitas[58].

Hasta este momento, hemos dado por sentados resultados característicos de metalógica como la completud, correctud, compacidad, etcétera, de los sistemas formales de primer orden y que, en algún sentido, también pueden considerarse como parte de la teoría de modelos, pues sus demostraciones usuales apelan a las herramientas de esta última. No obstante, por su relevancia en la próxima discusión, cabe detenernos en uno de ellos.

Probablemente, la contribución de uno de los primeros grandes resultados de la teoría de modelos (si no es que el primero) fue establecer algo sustancial y aparentemente paradójico acer-

57 El lector familiarizado con los estudios de metalógica no encontrará lo anterior particularmente impactante al recordar los resultados de Löwenheim-Skolem.

58 Esto no se refleja en órdenes superiores donde teorías con modelos infinitos sí pueden ser categóricas.

59 Intuitivamente, una teoría expresada en un lenguaje de primer orden nos permite cuantificar sobre objetos (*i. e.*, hablar de subconjuntos del dominio de discurso en términos cuantitativos), pero no sobre propiedades. Una forma más precisa de presentarlo es la siguiente.

La mayoría de los lenguajes, ya sean formales o naturales, son *tipados*. Esto quiere decir que sus expresiones se clasifican en diferentes categorías, llamadas *tipos*, de tal manera que hay reglas que establecen qué combinaciones entre expresiones de cierto tipo son gramaticalmente correctas (Bacon, 2024).

Probablemente, los lenguaje formales más conocidos de este estilo son los de las teorías de tipos donde cada variable tiene asociado un *símbolo de tipo*, el cual comúnmente se coloca como superíndice. Aunque estos símbolos de tipo usualmente son números naturales, nada impide extenderlos a ordinales arbitrarios. *Grosso modo*, la interpretación estándar es que las variables del *tipo 0* (*i. e.*, las variables de individuo) pueden tomar como valor individuos del dominio de discurso; las del *tipo 1*, conjuntos de individuos; las del *tipo 2*, conjuntos de conjuntos de individuos, y así sucesivamente. (En realidad, la cuestión es más complicada si consideramos secuencias de tipos como tipos o

ca de la relación entre una teoría formal —expresada en primer orden[59]— y su interpretación (Arsenijević, 2012, p. 64). Por supuesto, nos referimos al teorema de Löwenheim-Skolem. Ahora bien, ¿qué es esto sustancial y aparentemente paradójico?

Formalmente, el teorema no parece muy problemático[60]: si un conjunto de enunciados expresados en un lenguaje de cardinalidad numerable[61], llamémoslo Γ, tiene un modelo infinito[62], entonces tiene modelos de cualquier cardinalidad infinita. Sin embargo, al momento de mirar cuidadosamente el resultado, es que puede surgir la perplejidad y el desconcierto. Lo primero que salta a la vista es que el teorema implica que Γ es satisfacible por una infinidad de modelos de distinto tamaño infinito, lo cual, para empezar, quiere decir que la empresa de cualquier teoría (de primer orden) que pretende caracterizar de manera única un modelo pretendido de tamaño infinito está destinada al fracaso. Esto evidentemente significa que Γ no es categórica y, por ende, que no podemos distinguir formalmente lo que es distinguible informalmente desde afuera (Asenijević, 2012, p. 66).

---

incluso tipos mixtos; sin embargo, para los fines de esta exposición basta con los ya mencionados.)

De este modo, decimos que un lenguaje es de *orden n* si y sólo si todas sus variables son de tipo estrictamente menor que *n* (de nuevo, *n* podría ser cualquier ordinal). Por ejemplo, los *lenguajes de primer orden* (o de *orden uno*) solamente cuentan con variables de tipo *0*; mientras que los *lenguajes de segundo orden* (o de *orden dos*) cuentan tanto con variables de tipo *0* como con variables de tipo *1*.

Por otra parte, la diferencia expresiva entre los lenguajes de ordenes mayores con respecto de los menores, en pocas palabras, se debe a que los primeros incluyen cuantificadores que pueden cuantificar sobre más tipos de variables y a que pueden representar más abstracciones mediante expresiones con variables libres de más tipos. Para una discusión más detallada sobre estos temas, véase Tarski (1933, secc. 4 y 7), Linnebo y Rayo (2012) y Bacon (2023).

60 Por lo regular, el teorema suele presentarse en dos versiones: una ascendente y otra descendente. No obstante, para nuestros fines basta con considerar el corolario siguiente, que colapsa ambas.

61 *Grosso modo*, un lenguaje de cardinalidad numerable es un lenguaje con a lo más una cantidad numerable de fórmulas. Nótese que esto es compatible con que la signatura de tal lenguaje sea finita o a lo más numerable (véase la nota 8).

62 Dicho brevemente, un *modelo infinito* es un modelo cuyo dominio de discurso tiene una cantidad infinita de objetos.

Como mencionamos anteriormente, esto puede representar un grave problema cuando pretendemos caracterizar mediante una teoría una estructura pretendida o una clase bien delimitada de estructuras naturales o prototípicas de determinado tamaño infinito, pues la pérdida de categoricidad implica la presencia inevitable de una cantidad infinita de modelos no-estándar que no son isomorfos con los estándar. Pese a esto, mencionamos a su vez que esto puede resultar favorable frente a lógicas de órdenes superiores. En palabras de Väänänen (2021):

> Una razón de por qué la lógica de primer orden posee una teoría de modelos de semejante riqueza es que la lógica de primer orden es relativamente débil. Las teorías contables de primer orden que tienen modelos infinitos tienen modelos de todas las cardinalidades infinitas. Esta es consecuencia de una debilidad de la lógica de primer orden: sus enunciados no pueden delimitar el tamaño del modelo a no ser que sea finito. En teoría de modelos esta debilidad es convertida en una fortaleza: la estructura de los modelos de las teorías de primer orden puede ser analizadas de maneras interesantes. (secc. 7)[63]

En otra parte también dice:

> Puede sostenerse que la teoría de modelos de primer orden se basa en la incapacidad de la lógica de primer orden para expresar cosas tales como la finitud, la buena-fundación, la contabilidad, y la compleción de un orden lineal. Esta incapacidad la ha dirigido hacia una rica teoría de las estructuras, la cual constituye mucho de la teoría de modelos —la debilidad se ha convertido en una ventaja—. La lógica de segundo orden no "sufre" de esta incapacidad, y tampoco posee una rica teorías de las estructuras, al menos hasta donde se puede ver hoy en día. (2023, p. 289)[64]

En relación con esto, vale la pena considerar no sólo cuando, entre todas las $\mathcal{L}$-estructuras que satisfacen una teoría, algunas resultan

63 La traducción es propia.
64 Mi traducción.

ser subestructuras de otras, de modo que resultan indiscernibles por la teoría en cuestión, sino también un caso todavía más fuerte: cuando las $\mathcal{L}$-estructuras satisfacen las mismas fórmulas que sus subestructuras con las mismas asignaciones. Bajo estas circunstancias, aun cuando podría tratarse de $\mathcal{L}$-estructuras distintas, no obstante, resultan ser indiscernibles por el mismo lenguaje $\mathcal{L}$.

De esta manera, tenemos que una $\mathcal{L}$-estructura $\mathfrak{A}$ es una *subestructura elemental* de $\mathfrak{B}$ (o que $\mathfrak{B}$ es una *extensión elemental* de $\mathfrak{A}$), en símbolos $\mathfrak{A} \preccurlyeq \mathfrak{B}$, sii $\mathfrak{A} \subseteq \mathfrak{B}$ y, para cualquier asignación s: *Var* $\longrightarrow$ dom($\mathfrak{A}$) y cualquier fórmula φ de $\mathcal{L}$, $\vDash_{\mathfrak{A}} \varphi[s] \Leftrightarrow \vDash_{\mathfrak{B}} \varphi[s]$[65]. Es importante destacar que, si bien $\mathfrak{A} \preccurlyeq \mathfrak{B}$ implica que $\mathfrak{A} \equiv \mathfrak{B}$, en cambio, $\mathfrak{A} \subseteq \mathfrak{B}$ y $\mathfrak{A} \equiv \mathfrak{B}$ no implican que $\mathfrak{A} \preccurlyeq \mathfrak{B}$. Esto es claro cuando observamos que la definición de equivalencia elemental exige que ambas $\mathcal{L}$-estructuras satisfagan sencillamente los *mismos enunciados*; mientras que las subestructuras elementales requieren, aparte de ser subestructuras, satisfacer exactamente las *mismas fórmulas* con los *mismos objetos*[66].

Por último, cabe introducir una noción relacionada con la que podemos "convertir" una $\mathcal{L}$-estructura en otra, mas no mediante una modificación directa sobre aquélla. En su lugar, simplemente "olvidamos" la interpretación de algunos símbolos de $\mathcal{L}$, con lo que generamos —como diría Quine[67]— una reducción *ideológica*, y no *ontológica* (véase Button y Walsh, 2018, p. 15). Sean $\mathcal{L}_{_}$ y $\mathcal{L}$ dos signaturas tales que $\mathcal{L}_{_} \subseteq \mathcal{L}$, y $\mathfrak{A} = (\text{dom}(\mathfrak{A}), \iota)$ una $\mathcal{L}$-estructura; una $\mathcal{L}_{_}$-estructura $\mathfrak{B}$ es un $\mathcal{L}_{_}$*-reducto*[68] de $\mathfrak{A}$ (o que $\mathfrak{A}$ es una $\mathcal{L}$*-expan-*

65 Dicho de otro modo, $\mathfrak{A} \preccurlyeq \mathfrak{B}$, sii $\mathfrak{A} \subseteq \mathfrak{B}$ y, para cualquier $\varphi(x_{i_1},\ldots,x_{i_n})$ y cualesquiera $a_1,\ldots,a_n \in \mathfrak{A}$, $\vDash_{\mathfrak{A}} \varphi[a_1,\ldots,a_n] \Leftrightarrow \vDash_{\mathfrak{B}} \varphi[a_1,\ldots,a_n]$.

66 Conviene a su vez considerar el caso en que una estructura es isomorfa con alguna subestructura elemental de otra estructura. En este caso, decimos que hay un *encaje elemental* de la primera estructura en la superestructura.

67 Véase Quine (1951a).

68 También se dice que $\mathfrak{B}$ es el reducto de $\mathfrak{A}$ a $\mathcal{L}_{_}$.

69 La notación $A\restriction B$ suele ocuparse en teoría de conjuntos para representar la *restricción de una función A* (o posiblemente también de una *relación*) *a un conjunto B*. Dicho de modo más preciso, $A\restriction B = \{(x, y) \in A : x \in B\}$.

sión de $\mathfrak{B}$) sii es la única $\mathcal{L}_-$-estructura tal que $\mathfrak{B} = (\mathrm{dom}(\mathfrak{A}), \iota')$; donde $\iota' = \iota \restriction \mathcal{L}_-$[69].

8. Teoría de Modelos y Filosofía Analítica

Hasta aquí, hemos introducido una serie de nociones que aparecen con frecuencia al momento de estudiar los modelos de las teorías formales. Aunque, en primera instancia, el alcance de la teoría de modelos puede parecer muy limitado y exclusivo de las teorías y modelos de la matemática, el rigor característico de sus resultados y el poder de sus herramientas conceptuales le han abierto paso dentro de la discusión en varias áreas de la filosofía analítica.

Justamente, en esta última parte, nos proponemos ilustrar la manera en que podemos disponer de las nociones modelo-teóricas para clarificar y abordar con precisión distintos problemas de carácter filosófico. Para ello, nos concentraremos en un ejemplo que atraviesa por diferentes áreas del análisis filosófico, a saber, el problema de la *inescrutabilidad de la referencia*. Como veremos, abordar este problema desde el marco conceptual de la teoría de modelos nos permitirá dar cuenta de que otra motivación para adoptar sus recursos metodológicos es que facilita la interacción entre las diferentes disciplinas filosóficas, lo que posibilita a su vez un tratamiento más uniforme e integral de los problemas en cuestión.

8.1 El Problema de la Inescrutabilidad de la Referencia

Iniciamos este capítulo con la afirmación de que una de las funciones principales del lenguaje era representar o describir el mundo, y confesamos que, si bien no resultaba totalmente extraño, este primer acercamiento no era muy preciso. Aunque en su momento no profundizamos mucho sobre este asunto, esto lo sostuvimos, por

[69] En el caso en cuestión, la función interpretación ι' no es más que la restricción de la función ι al conjunto de los símbolos de $\mathcal{L}_-$. Es decir que ι' es exactamente como ι salvo que $\mathcal{L}_- \subset \mathcal{L}$, en cuyo caso no tendrá las interpretaciones correspondientes con los símbolos de $\mathcal{L}$ que no están en $\mathcal{L}_-$.

un lado, sobre la base de que no era muy claro cómo es que se interpretaban las expresiones del lenguaje que pretendían representar distintos aspectos de la realidad y, por otro, en tanto que no era muy inmediato comprender lo que queríamos decir cuando hablamos del "mundo" o de la "realidad". En cuanto a lo primero, la breve presentación de la teoría de modelos en secciones anteriores, entre otras cosas, puede darnos una idea del modo en que llevamos a cabo la interpretación del lenguaje. Sin embargo, en relación con lo segundo, las cuestiones sobre lo que hay y cómo lo conocemos caen más allá del terreno de la lógica o, incluso, de la filosofía misma. En este sentido, no es raro considerar que la ciencia, con su amplia variedad de disciplinas, se encarga de investigar diferentes fenómenos y de desarrollar teorías para explicarlos, y cuyo análisis puede contribuir sobre nuestra mejor comprensión del mundo[70].

Ya sea para identificar su forma lógica, formular su contenido sin ambigüedad, o para facilitar el reconocimiento de nuestros compromisos ontológicos, por decir algo, muchas veces nos resulta conveniente formalizar nuestras teorías científicas. En este caso, podemos implementar naturalmente nuestra maquinaria modelo-teórica para describir la semántica de las correspondientes teorías formales. Lo que intentaremos mostrar a continuación es que la adopción de la teoría de modelos puede llegar a ser bastante útil tanto para poner de manifiesto problemas como para motivar una serie de discusiones en varias áreas de la filosofía analítica. Sin embargo, para simplificar el planteamiento de este problema en el caso de las teorías científicas, consideremos primero el siguiente escenario paradigmático, proveniente de la filosofía de la matemática.

Frecuentemente, cuando trabajamos dentro de alguna disciplina matemática, nos encontramos con que no es muy claro qué son exactamente las entidades matemáticas que constituyen sus objetos de estudio. Un caso evidente de esto puede apreciarse con la teoría elemental de números, donde la claridad de la cuestión sobre

70 Obviamente, esto es muy discutible. Asumir este papel de la ciencia ya presupone ciertas posturas filosóficas, por ejemplo, una suerte de realismo o un compromiso de la ciencia con la búsqueda de *la verdad*.

qué son los *números naturales* se ve disminuida adicionalmente por el primer teorema de incompletud de Gödel, el cual, *grosso modo*, implica que la aritmética es incompleta para la negación; si, como dice Quine (1969), "la aritmética es todo lo que hay para el número" (p. 45), entonces inevitablemente la noción de número se verá impedida, entre otras cosas, porque habrá enunciados independientes que probablemente contienen información sobre ésta y que no podrán ser aprehendidos por medios formales combinatorios. Además, el esclarecimiento de esta noción se ve obstaculizada especialmente porque la teoría elemental de números[71] es susceptible del teorema de Löwenheim-Skolem, cosa que a su vez dificulta la especificación de su universo de discurso o su reducción a otro, pues, en consecuencia, la teoría es realizable por una cantidad infinita de modelos, la mayoría de ellos, no-estándar.

Aunque lo anterior exhibe una serie de dificultades adicionales para la demarcación de los objetos de la matemática, no necesitamos ir tan lejos para ilustrar el punto que aquí nos interesa destacar. Una de las observaciones más sobresalientes hechas por Benacerraf (1965) fue que, al trabajar dentro de alguna teoría matemática, como la aritmética, para los propósitos de la matemática, resulta indiferente si, por decir algo, interpretamos los términos de la teoría de números como ordinales finitos de Zermelo o como ordinales finitos de von Neumann[72]. En este sentido, parece que no resulta muy relevante el tipo de constructos conjuntistas ni los objetos que elijamos como referencia del concepto de número, siempre que satisfagan los enunciados de la teoría (en buena medida porque la aritmética en primer orden no ofrece una caracterización lo suficientemente precisa del concepto en cuestión). ¿Cómo, entonces, puede determinarse la auténtica referencia de este concepto? En general, ¿cómo podría establecerse la referencia

71 Aquí el término "elemental" es sinónimo de "primer orden".

72 Un número natural de Zermelo es el conjunto unitario del número que le precede; mientras que un número natural de von Neumann es el conjunto de todos los naturales anteriores a él. Ambas construcciones satisfacen todas las oraciones de la aritmética de Peano.

pretendida de cualquier noción cuyos únicos indicios estén dados por su teoría formal expresada en orden uno y que a su vez admite una cantidad infinita de modelos de diferente tamaño?

La aritmética no es el único caso que se enfrenta con este problema, el análisis real y la teoría de conjuntos (en sus versiones clásicas de primer orden) son otro par de ejemplos donde claramente ocurre lo mismo. En efecto, los enunciados de la aritmética son los únicos que pueden decir algo acerca de los números, pero no parece correcto decir que cualquier cosa que los satisfaga son números por igual e identificables entre sí[73]. Esto es precisamente uno de los puntos señalados por Benacerraf. Consideremos, por ejemplo, las ya mencionadas reconstrucciones conjuntistas de los naturales ofrecidas por Zermelo y von Neumann. Al momento de intentar igualarlas en la metateoría, podemos apreciar que los sustitutos conjuntistas de los números difieren entre sí en la satisfacción de gran parte de los enunciados metateóricos (como que el cero pertenece al tres en la segunda de ellas, pero no así en la primera). Ahora bien, ¿cuál de ellas, si es que alguna, es la que trabaja con números genuinos?

Más aún, cualquier biyección sobre el dominio de un cierto modelo de la teoría puede utilizarse para definir un isomorfismo y —como vimos en secciones anteriores— esto quiere decir que, mientras dispongamos de suficientes objetos, podemos construir una estructura isomorfa que necesariamente satisfará los mismos enunciados que el modelo original (Button y Walsh, 2018, p. 36). Así, la pregunta surge de nuevo prácticamente por sí sola: ¿qué razones podemos aludir para preferir un modelo en lugar del otro?[74]

En este momento, podríamos tener la impresión de que este tipo de problemas ocurren exclusivamente en áreas relacionadas con la matemática, donde la pregunta por la naturaleza intrínseca

73 Frege (1884) ya había notado este problema. Incluso llegó a sostener que la referencia de los términos numéricos debía ser única.

74 Para una discusión más detallada sobre este punto, véase Button y Walsh (2018, cap. 2).

de sus objetos, para empezar, es de por sí extraña, pues no tenemos muchas intuiciones sobre qué clase de objetos son desde un punto de vista ontológico. Sin embargo, en lo que sigue, nos percataremos de que no es muy complicado extrapolar esta situación a otras áreas de la filosofía como la filosofía de la ciencia[75].

Imaginemos que nos encontramos dentro de alguna disciplina científica estudiando algún fenómeno de interés. Podemos pensar, por ejemplo, que nos encontramos investigando interacciones entre partículas subatómicas, algún tipo de reacción química, la influencia de los procesos fisicoquímicos de la célula en la evolución de los organismos, etcétera[76].

Asumamos que hay una teoría dentro de la mencionada disciplina que ha logrado explicar exitosamente el fenómeno en cuestión. Por mor del argumento, supongamos a su vez que esta teoría ha sido formalizada mediante un lenguaje de primer orden. Esto quiere decir que, una vez que interpretamos los símbolos del lenguaje de manera particular, los enunciados de nuestra teoría resultan ser verdaderos en ese modelo (en especial, aquéllos que se relacionan directamente con el fenómeno bajo investigación). Desgraciadamente, por observaciones como las arriba, sabemos que es posible construir un modelo isomorfo y alternativo al primero en la medida que dispongamos de una cantidad suficiente de objetos. El problema estriba en que, al ser isomorfos, ambos modelos son elementalmente equivalentes, pero, al tener dominios distintos, difieren en la referencia de alguno de los términos. De esta manera, la pregunta que nos inquietaba en el caso de la matemática regresa dramáticamente, salvo que en esta ocasión puede reformularse dentro del ámbito de la ciencia: si pretendiéramos esta-

75 La mayoría de los siguientes argumentos fueron recuperados de Button y Walsh (2018, cap. 2), aquí simplemente hemos adaptado sus observaciones para el caso de las teorías científicas.

76 Parece que, en cada uno de estos fenómenos, nos encontramos estudiando propiedades de cierto tipo de objetos y relaciones entre ellos, así como con otros. Esto, en principio, parece adecuarse al tipo de análisis que puede generarse a partir de la teoría de modelos, en tanto que consideramos un dominio de cuantificación —la clase de objetos que deseamos estudiar—, y una serie de relaciones definidas sobre éste.

blecer la referencia de los términos de nuestro lenguaje científico mediante nuestra noción modelo-teórica de interpretación, ¿sobre qué base podemos asegurar que cierto modelo explica mejor la referencia de nuestras teorías científicas que otro?

Frente a esta situación, podríamos responder de varias formas. Por mencionar un par, podríamos alegar que no disponemos de suficientes objetos o que la interpretación alternativa no es "natural" y no respeta el significado del resto de los términos empleados en la explicación del fenómeno. Lo que mostraremos a continuación es que cada uno de estos intentos está condenado al fracaso por una u otra razón.

Aun cuando sostuviéramos, como en la primera objeción, que no hay una cantidad suficiente de otros objetos, pronto caeríamos en cuenta de que nuestros esfuerzos serían inútiles, pues de cualquier modo siempre podemos inducir un isomorfismo por medio de una *permutación* de los objetos, es decir, a partir de una biyección del dominio de nuestro modelo original sobre sí mismo. De este modo,

> podemos ver que cualquier nombre podría ser utilizado para referir *cualquier cosa*, que cualquier predicado de un lugar podría ser empleado para seleccionar *cualquier* colección de objetos (únicamente en la medida que haya una cantidad suficiente de ellos), y de manera similar para todas las otras expresiones de nuestro lenguaje. Estaremos contemplando dentro del abismo de la *indeterminación radical de la referencia*, donde cada palabra refiere igualmente a todo, lo que es decir tan solo que nada refiere en absoluto. (Button y Walsh, 2018, p. 40, traducción propia)

De cualquier forma, podríamos insistir que estamos en desacuerdo, ya que en ese caso los términos no respetarían su significado pretendido. No obstante, esta clase de respuesta nos lleva inmediatamente a la segunda de nuestras objeciones.

La segunda objeción a grandes rasgos nos dice que podemos mantener nuestra postura y fijar la referencia de nuestras teorías científicas de conformidad con cierta *predilección* por determinados

modelos en lugar de otros. Sin embargo, todavía podemos rechazar este intento con base en un elegante argumento ofrecido por Putnam (1980). *Grosso modo*, la idea es que podemos contrarrestar la réplica si observamos que quien la sostiene debería ofrecernos razones de por qué determinados modelos son más *preferibles* que el resto. En otras palabras, el partidario de esta posición tendría que proporcionar algo como una "teoría de la preferencia" que clarifique por qué ciertas referencias son superiores a las demás cuando pretendemos explicar los fenómenos estudiados por nuestras teorías científicas. Desafortunadamente, esta teoría adicional —como cualquier otra— no es una excepción y resulta ser "sólo más teoría", por lo que podemos adaptar y reproducir el argumento de la permutación una vez más para este caso.

Como podemos apreciar, si bien la maquinaria de la teoría de modelos nos ha permitido poner de manifiesto problemas que yacen al fondo de cuestiones de naturaleza ontológica, por otra parte, también puede incentivar nuestras reflexiones dentro del terreno de la epistemología.

Así como la aritmética es todo lo que tenemos a nuestra disposición para establecer propiedades y relaciones entre números, de manera análoga, muchas veces lo único con lo que contamos para ampliar nuestro conocimiento acerca del mundo es aquello que nuestras teorías científicas nos dicen sobre él. De esta forma, nos podemos aproximar a otro de los problemas que podemos desprender desde los resultados que presentamos en secciones anteriores: si lo que podemos saber sobre los objetos del mundo está dado en parte por la referencia de los términos que forman parte de nuestras mejores teorías sobre el mundo (Quine, 1951b), entonces estamos en un grave problema porque, como hemos dicho, la satisfacción de un conjunto consistente de enunciados no determina estructuras de manera única[77]. Dicho de otro modo, si nuestras teorías científicas son lo mejor que tenemos para conocer

77 En realidad, determinar los criterios de identidad entre estructuras y teorías es otro problema que podemos explorar por medio de la teoría de modelos (véase, Button y Walsh, cap. 5).

genuinamente las cosas, sin importar qué tanto progrese la ciencia ni cuanto ampliemos nuestras teorías, nunca podremos contar con que efectivamente hayamos alcanzado la estructura última de la realidad, pues la equivalencia elemental entre estructuras no implica que sean isomorfas.

Al respecto de esto, podríamos tratar de resistir y mantener que, aun cuando por lo general contienen terminología cuestionable, nuestras teorías científicas contienen términos claramente especificables sobre los cuales quizá podemos determinar los demás. En este sentido, la propuesta pretende apelar de algún modo a la clásica distinción entre *términos observacionales* y *términos teóricos*[78]. En pocas palabras, la idea es la siguiente: dentro de la signatura del lenguaje formal que empleamos para expresar nuestra teoría científica, llamémosla $\mathcal{S}$, podemos realizar una partición que separe los términos observacionales de los teóricos, esto con la intención de considerar un $\mathcal{S}^{-}$-reducto que proporcione sólo la referencia de la terminología observacional. Pero esto no es todo, la respuesta sugiere además que los términos teóricos sean definibles en el lenguaje generado por $\mathcal{S}^{-}$. Por medio de esta estrategia, pues, pretendemos deshacernos de los términos problemáticos en favor de aquéllos "mejor comprendidos". Desgraciadamente, esto tampoco es suficiente, pues la razón detrás de la preferencia por los términos observacionales se apoya sobre el hecho pragmático de que podemos determinar su referencia por medios ostensivos. Sin embargo, como Quine (1969) ya nos ha mostrado, mecanismos como la ostensión no están libres de la indeterminación de la referencia.

Aun si no estuviéramos de acuerdo con Quine, podemos encontrar otras razones por las cuales la propuesta de arriba resulta problemática. Por decir algo, incluso si concedemos que en varias ocasiones podemos sostener enunciados que involucran directamente términos observacionales con base en nuestra mera percepción, podríamos reconocer —como hace Hanson (1965)— que inevitablemente estas afirmaciones están sesgadas por la carga teó-

78 Para una discusión sobre esta distinción, véase Olivé y Pérez-Ransanz (1989).

rica de la observación y, una vez más, podríamos aplicar el argumento de la permutación sobre esta última.

A pesar de todo esto, incluso si pudiéramos establecer con éxito y unívocamente la referencia de nuestras teorías, todavía cabría la posibilidad de que se tratara de una subestructura elemental y que dejáramos fuera varios objetos que efectivamente existen. En fin, parece que de cualquier forma todos nuestros esfuerzos por conocer la estructura última de la realidad están destinados al fracaso. Con todo, esta versatilidad de las teorías también permite explorar perspectivas alternativas (modelos) de cómo podría ser la realidad —con distintas ontologías e interpretaciones de las relaciones—, que satisfacen los enunciados de la teoría y que pueden ser estudiadas y contrastadas por el científico. La teoría por más precisa que sea no determinará una única manera de ser del mundo, pues tendrá múltiples modelos que la satisfagan. Del mismo modo, toda la evidencia empírica que podemos recuperar del mundo puede ser subsumida por más de una teoría científica, esto puede verse fácilmente considerando las reconstrucciones modelo-teóricas. Dado que la evidencia empírica de la cual podemos disponer siempre es finita, pero la cantidad de fenómenos posibles a considerar es potencialmente infinita (o quizá finita, pero mucho mayor que la evidencia de la que disponemos), siempre será posible generar teorías distintas que den cuenta de las observaciones, las cuales, sin embargo, tendrán consecuencias distintas aún no observadas. La evidencia empírica nunca será suficiente para determinar una única propuesta teórica que dé cuenta de ella. Esta es justamente la tesis conocida como la tesis de la subdeterminación empírica de las teorías científicas —también conocida como la tesis Duhem-Quine—.

Por último, tal vez podríamos adoptar una postura más radical y asegurar, como Quine (1969), que preguntarse por la referencia de manera absoluta no tiene sentido. Esto se debe a que, dada nuestras limitaciones para fijar la referencia, parece que inevitablemente debemos recurrir a una teoría de fondo más amplia que nos permita establecer (si bien parcialmente) la referencia de los

términos de la teoría original. De manera semejante a cuando pretendemos reducir objetos de un nivel teórico a otro (e. g., cuando intentamos reducir o explicar fenómenos químicos mediante fenómenos físicos). Aunque parece que el problema para establecer la referencia de los términos de la teoría original se ha resuelto, esto depende por completo de que la teoría de fondo logre a su vez determinar la referencia de sus propios términos. Por desgracia, para que consiga establecer la referencia de sus términos, esta teoría de fondo debe recurrir a una nueva teoría de fondo más amplia, y así sucesivamente. No es muy difícil advertir cómo esto puede generar un regreso al infinito.

9. Conclusiones

A lo largo de este capítulo, hemos introducido nociones fundamentales de la teoría de modelos que permiten estudiar la relación entre las teorías formales y las estructuras que las interpretan. En particular, explicamos qué es un lenguaje formal y una estructura, así como la manera en que esta última asigna significado a los símbolos del primero; examinamos las distintas formas en que las estructuras pueden relacionarse, ya sea de manera independiente o en función del lenguaje en el que se expresan, y exploramos el contexto histórico en el que surgió la noción de verdad en teoría de modelos. A partir de ello, definimos la noción de modelo como una estructura que hace verdadero un conjunto de enunciados y destacamos su papel en la precisión de conceptos lógicos clave, como la consecuencia, la verdad y la equivalencia lógica.

A partir de estas nociones y de algunos resultados relacionados con ellas, abordamos el problema de la inescrutabilidad de la referencia, el cual atraviesa diversas áreas de la filosofía, desde la metafísica y la epistemología hasta la filosofía de las matemáticas y la filosofía de la ciencia. Grosso modo, este problema consiste en que no podemos determinar la referencia de los términos singulares de un lenguaje con la información que tenemos a nuestra disposición.

En la teoría de modelos, este problema puede reproducirse si pretendemos caracterizar una estructura única salvo isomorfismo

mediante una teoría expresada en primer orden, pues las limitaciones expresivas de los lenguajes de primer orden a menudo nos impiden proporcionar información suficiente para determinar de manera exclusiva y exhaustiva la referencia de los términos de la teoría y sus relaciones entre sí. En otras palabras, dado que la equivalencia elemental entre estructuras no implica que sean isomorfas, las teorías solo ofrecen condiciones necesarias, pero no suficientes, para fijar referencias. Esto quiere decir que las relaciones estructurales descritas por estas teorías pueden ser satisfechas igualmente por distintos candidatos. Sobra decir que el problema se vuelve aún más complejo si las teorías poseen un modelo infinito, en cuyo caso tendrían múltiples modelos de tamaño infinito, cada uno de los cuales ofrecería una interpretación adecuada para la teoría.

De esta manera, sostuvimos que este problema tiene relevancia en varias áreas de la filosofía. Por ejemplo, en filosofía de las matemáticas, al demostrar que una teoría formal puede admitir múltiples modelos no isomorfos, la teoría de modelos nos muestra que el problema de la inescrutabilidad de la referencia se manifiesta en la dificultad para determinar la identidad de los números y de otras entidades matemáticas. Esto se puede observar en la discusión de Benacerraf sobre las reconstrucciones conjuntistas de los números naturales, que sugiere la ausencia de un modelo privilegiado. En el caso de la filosofía de la ciencia, por su parte, el problema se puede plantear en relación con la subdeterminación de las teorías científicas por la evidencia empírica: si la misma evidencia puede ser explicada por múltiples teorías no equivalentes, y estas a su vez pueden admitir diferentes interpretaciones como indica la teoría de modelos, entonces la determinación de la referencia de los términos científicos puede entorpecerse. Esto, además, puede plantear un desafío epistemológico, ya que impide establecer una correspondencia única entre los términos de las teorías científicas y los objetos del mundo, lo cual puede despertar dudas acerca del realismo científico y de la posibilidad de conocer la estructura última del mundo. Asimismo, el tratamiento modelo-teórico del pro-

blema de la inescrutabilidad de la referencia nos permite destacar los límites de nuestro acceso al mundo a través del lenguaje y las teorías: si la relación entre el lenguaje y el mundo no es una correspondencia directa y nuestras descripciones del mundo no pueden capturar una única estructura ontológica subyacente, entonces la existencia de múltiples modelos nos ofrece diversas posibilidades de conceptualizar la realidad. Esto puede llevarnos a cuestionar la posibilidad de adquirir un conocimiento objetivo o de obtener un acceso epistémico privilegiado a una única realidad.

Por otro lado, desde un punto de vista ontológico, la dificultad de establecer una única estructura ontológica subyacente a nuestras descripciones del mundo podría sugerir que, si nuestras mejores teorías científicas determinan lo que hay, la referencia de nuestros términos es relativa a un sistema conceptual. Esto, a su vez, podría favorecer una forma de pluralismo ontológico, dado que la teoría de modelos revela la existencia de múltiples modelos para nuestras teorías.

En conclusión, el propósito de presentar este problema desde el marco conceptual de la teoría de modelos fue mostrar la precisión y claridad con las que puede formularse en términos de los conceptos de dicha teoría. Asi mismo, al exponer sus manifestaciones en diferentes áreas de la filosofía y presentarlas en los mismos términos conceptuales, pudimos mostrar cómo se pueden establecer conexiones entre las distintas reflexiones filosóficas características de cada área en relación con este problema.

De ninguna manera el tratamiento ofrecido en este capítulo ha pretendido ser exhaustivo ni agotar todas las aplicaciones de la teoría de modelos en la filosofía. La única intención ha sido ofrecer una pequeña muestra de su importancia y utilidad para la labor filosófica. Espero que, a partir de la maquinaria conceptual aquí presentada, el lector interesado pueda acercarse y explorar el fascinante espacio de posibilidades que la teoría de modelos tiene para ofrecer.

Referencias

Amor, J. A. (2008). La teoría de modelos de la teoría de conjuntos: un concepto delicado. En A. García de la Sienra, *Reflexiones sobre la paradoja de Orayen* (pp. 79-91). IIFs-UNAM.

Aristóteles. (1994). *Metafísica*. Gredos.

Arsenijević, M. (2012). The Philosophical Impact of the Löwenheim-Skolem Theorem. En Trobok, M., Miscevic, N. y Zarnic, B., *Between Logic and Reality: Modeling Inference, Action and Understanding* (Vol., 25, pp. 59-81). Springer.

Bacon, Andrew (2023). *A Philosophical Introduction to Higher-order Logics*. Routledge.

Benacerraf, P. (1965). What numbers could not be. *The philosophical review*, 74(1), 47-73.

Button T. y Walsh S. (2018). *Philosophy and Model Theory*. Oxford University Press.

Cassini, A. (2009). *El juego de los principios: Una introducción al método axiomático*. A-Z Editora.

Chang, C. C. y Keisler, H. J. (1990). *Model Theory*. Elsevier.

David, M. (2008). Tarski's Covention T and the Concept of Truth. En D. Patterson (Ed.). *New Essays on Tarski and Philosophy* (pp. 133-156). Oxford University Press.

Davidson, D. y Harman, G. (1972). *Semantics of Natural Language*. Dordrecht: Reidel.

Enderton, H. B. (2001). *Una introducción matemática a la lógica*. IIFs-UNAM.

Feferman, S. (2008). Tarski's Conceptual Analysis of Semantical Notions. En D. Patterson (Ed.), *New Essays on Tarski and Philosophy* (pp. 72-93). Oxford University Press.

Frege, G. (1884). Los fundamentos de la aritmética: Una investigación lógico-matemática sobre el concepto de número. En M. M. Valdés (Ed.), *Escritos sobre lógica, semántica y filosofía de las matemáticas* (pp. 361-487). IIFs-UNAM.

Frege, G. (1892). Sobre sentido y referencia. En M. M. Valdés (Ed.), *Escritos sobre lógica, semántica y filosofía de las matemáticas* (pp. 321-48). IIFs-UNAM.

Frege, G. (1918/1919). El pensamiento; una investigación lógica. En M. M. Valdés (Ed.), *Escritos sobre lógica, semántica y filosofía de las matemáticas* (pp. 321-48). IIFs-UNAM.

Gómez-Torrente, M. (2000). *Forma y Modalidad: Una introducción al concepto de consecuencia lógica*. Eudeba.

Hanson, N. R. (1965). *Patterns of discovery: An inquiry into the conceptual foundations of science*. CUP Archive.

Hodges, W. (1993). *Model Theory*. Cambridge University Press.

Hodges, W. (1997). *A Shorter Model Theory*. Cambridge University Press.

Hodges, W. (2009). Set Theory, Model Theory and Computability Theory. En L. Haaparanta (Ed.). *The Development of Modern Logic* (pp. 471-98). Oxford University Press.

Hodges, W. (2018). *A Short History of Model Theory*. En T. Button y S. Walsh (Eds.), *Philosophy and Model Theory* (pp. 439-76). Oxford University Press.

Hodges, W. (2022). Tarski's Truth Definitions. En E. N. Zalta y U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University.

Hodges, W. (2023). Model Theory. En E. N. Zalta y U. Nodelman (Eds.). *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University.

Hodges, W., y Scanlon, T. (2024). First-order Model Theory. En E. N. Zalta y U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University.

Lenci, A. y Sandu, G. (2009). Logic and Linguistics in the Twentieth Century. En L. Haaparanta (Ed.), *The Development of Modern Logic* (pp. 775-847). Oxford University Press.

Linnebo, Ø., & Rayo, A. (2012). Hierarchies Ontological and Ideological. *Mind, 121*(482), pp. 269–308.

Mancosu, P., Zach, R. y Badesa, C. (2009). The Development of Mathematical Logic from Russell to Tarski, 1900-1935. En L. Haaparanta (Ed.). *The Development of Modern Logic* (pp. 318-470). Oxford University Press.

Manzano, M. (1999). *Model Theory*. Oxford University Press.

Mendelson, E. (1997). *Introduction to Mathematical Logic*. Springer.

Montague, R. (1974). *Formal Philosophy. Selected Papers of Richard Montague*. Yale University Press.

Müller, R. (2009). The notion of a model: A historical overview. En D. M. Gabbay *et al.*. *Handbook of the Philosophy of Science,Volume 9: Philosophy of Technology and Engineering Sciences* (pp. 637–64). Elsevier.

Olivé, L. y Pérez-Ransanz A. R. (1989). *Filosofía de la ciencia: teoría y observación*. Siglo XXI-UNAM.

Putnam, H. (1980). Models and Reality. *The journal of symbolic logic, 45*(3), pp. 464-82.

Quine, W.V.O. (1951a). Ontology and ideology. *Philosophical Studies, 2*(1), pp. 11-15.

Quine, W.V.O. (1951b). Two Dogmas of Empirism. *Philosophical Review, 60*(1), pp. 20-43.

Quine, W.V.O. (1969). *Ontological Relativity and Other Essays*. Columbia University Press.

Quine, W.V.O. (1973). *Filosofía de la lógica*. Alianza.

Shapiro, S. (1991). *Foundations without Foundationalism: A Case for Second-order Logic*. Clarendon.

Shoesmith, D. J. y Smiley, T. J. (1978). *Multiple-conclusion Logic*. Cambridge University Press.

Speaks, J. (2021). Theories of Meaning. En E. N. Zalta (Ed.). *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab. Stanford University.

Tarski, A. (1933). The Concept of Truth in Formalized Languages. En J. Corcoran (Ed.) *Logic, Semantics, Metamathematics* (pp. 152-278). Hackett Publishing Company

Tarski, A. (1944). The Semantic Conception of Truth: and the Foundations of Mathematics. *Philosophy and Phenomenology Research, 4*(3), 341-76.

van Dalen, D. (2013). *Logic and Structure*. Springer.

Väänänen, J. (2021). Second-order and Higher-order Logic. En E. N. Zalta (Ed.). *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University.

Väänänen, J. (2023). Model theory of second order logic. En J. Iovino, *Beyond First Order Model Theory,Volume II* (pp. 289-304). Chapman and Hall.

CAPÍTULO V

REVISITING THE CARNAP-QUINE DEBATE ON ANALYTICITY

Siddhant Khamkar[1]

RESUMEN

En el siguiente artículo, intentamos explicar el debate público que tuvo lugar entre Carnap y Quine sobre la naturaleza de la analiticidad y su papel en la filosofía científica, durante mediados del siglo XX. Los famosos argumentos de Quine contra la distinción analítico-sintética en "Dos dogmas del empirismo", que fueron considerados como una refutación del proyecto empirista lógico, constituyen la materia principal de este artículo. Tras ofrecer una breve descripción de la visión recibida del debate, intentamos mostrar, a través de un análisis textual de sus obras, que las posiciones de Carnap y Qui-

1 Siddhant Khamkar es investigador en el Indian Institute of Technology Bombay (IIT), donde se especializa en filosofía de la ciencia, metafísica y epistemología. Su trabajo académico abarca temas como la filosofía de Kant, y la filosofía de la física.

ne no son completamente antitéticas entre sí, una vez realizada, una caracterización de sus diferencias epistemológicas y metodológicas. Concluimos reconsiderando las diferentes propuestas que Quine y Carnap defienden para alcanzar una filosofía de la ciencia fructífera.

Palabras clave: Empirismo lógico, Analítico, Carnap, Quine.

ABSTRACT

In the following paper, we attempt to explicate the public debate that took place between Carnap and Quine on the nature of analyticity and its role in scientific philosophy, during the mid-twentieth century. Quine's famous arguments against the analytic-synthetic distinction in the 'Two Dogmas of Empiricism' which were considered as a refutation of the logical empiricist project form the primary subject matter of this paper. Upon giving a brief account of the received view of the debate, we then attempt to show, through a textual analysis of their works, that Carnap and Quine's positions are not entirely antithetical to each other, followed by a characterisation of their epistemological and methodological differences. We then conclude by reconsidering the different proposals that Quine and Carnap advocate for conducting a fruitful philosophy of science.

Keywords: Logical Empiricism, Analytic, Carnap, Quine.

1. Introduction

The public debate between Rudolf Carnap and W.V Quine, two of the most influential philosophers of the 20th century, has had an enormous impact in shaping theories about scientific knowledge and setting the contours for different conceptions of philosophy of science. This debate, unfolding over multiple papers and discussions between the 1930s to 1960s, covered philosophical issues including the nature of meaning, truth, necessity, the basis of mathematics and logic, the distinction between analytic and synthetic sentences, and the very feasibility of notions like evidence, verification etc. One of the primary tenets of empiricist philosophy was a sharp distinction between logical and mathematical sentences - which can be varyingly

construed as being true based on meanings or conventions, called analytic sentences, and between factual sentences about the world, called synthetic sentences. An earlier notion of analyticity was previously characterised by Kant in his *Critique of Pure Reason* wherein - analytic judgements are those where the predicate is 'contained' in the subject, while synthetic judgements are those where the predicate lies entirely outside the subject. Given its essential role in his philosophical framework, Carnap (1966) uses the example of Einstein's Relativity to argue for the importance of this distinction:

> In my opinion, a sharp analytic-synthetic distinction is of supreme importance for the philosophy of science. The theory of relativity, for example, could not have been developed if Einstein had not realised that the structure of physical space and time cannot be determined without physical tests. He saw clearly the sharp dividing line that must always be kept in mind between pure mathematics, with its many types of logically consistent geometries, and physics, in which only experiment and observation can determine which geometries can be applied most usefully to the physical world. This distinction between analytic truth (which includes logical and mathematical truth) and factual truth is equally important today in quantum theory, as physicists explore the nature of elementary particles and search for a field theory that will bind quantum mechanics to relativity. (p. 257)

Quine famously argues in his works that there cannot be such a principled distinction between analytic and synthetic truths and that giving up such a distinction would amount to a blurring of the distinction between speculative metaphysics and natural science. At the heart of this debate lies Quine's famous critique of the logical positivist doctrine in his seminal 1951 paper "Two Dogmas of Empiricism." Quine aims his arguments to undermine two central tenets of the empiricist program he attributes to Carnap and his Vienna Circle colleagues – (1) The notion that there exists a clear demarcation between truths grounded purely in meanings independent of factual evidence (analytic statements), versus empirical statements about the observed world (synthetic statements). (2) The view that scientific statements have their meanings uniquely

fixed by the method of their empirical verification, and that philosophical elucidation should reduce knowledge to an irrefutable basis grounded in sensory evidence and logic.

While Quine's arguments reshaped the philosophical landscape, his portrayal of Carnap and the Vienna Circle has often been critiqued for oversimplifying their aims. Uebel (1996) highlights that Quine, alongside Ayer in *Language, Truth and Logic*, contributed to the persistent image of the Vienna Circle as pursuing "a particularly hard-headed… reductionist empiricism," often depicted as an extension of British empiricism. Uebel argues that such accounts wrongly attribute foundationalist intentions to the Circle's approach and fail to capture its methodological nuances. Richard Creath (2007a) similarly observes that Quine's persuasiveness relies on shaping the available alternatives for readers, thereby constructing a historical narrative of Carnap's philosophy. It is due to the huge influence of Quine's "histories" that there is a widely held belief that Carnap was driven in his 'Aufbau' to reduce science to an absolutely certain basis of sense data ontologically, or that his conception of analyticity was driven by a quest for certainty and other such caricatures, which are still "endlessly repeated by others." It is in opposition to these representations of Carnap's philosophy that Quine articulates his own positions.

Recent commentators such as Richard Creath, Alan Richardson, and Michael Friedman have emphasized that Quine's historical framing distorts Carnap's broader philosophical motivations and program. Creath (1991, 2007b), for instance, demonstrates that Carnap's position can meet Quine's demand for clarity and intelligibility regarding the analytic-synthetic distinction. Richardson (1997) adds that "the main interest of the dispute is to be found in the lack of genuine engagement between Carnap and Quine," as their disagreements often arose from fundamentally divergent methodological frameworks. In this chapter, we critically examine the Received View of the Carnap-Quine debate, addressing its shortcomings and reassessing Carnap's contributions in light of recent scholarship. Finally, we attempt to highlight the significant

epistemological and methodological differences in their philosophy through a textual analysis of their works.

2. The Received View

One of Quine's primary motivations for advocating a naturalised epistemology is his repudiation of a class of a priori truths and any epistemological foundationalism (Quine, 1969). In his view, the logical empiricist notion of analyticity is an attempt to explain the a priority of certain kinds of statements (such as mathematical and logical truths), and hence arguing against the analytic-synthetic distinction was considered a major motor for his move towards naturalising epistemology. In the 'Two Dogmas' (Quine 1951), he points out that there is no clear boundary drawn between analytic and synthetic sentences, and "that there is such a distinction to be drawn at all is an unempirical dogma of empiricists, a metaphysical article of faith" (p. 34).

Quine's attack on the two 'dogmas' of empiricism was highly influential in the philosophical literature in various areas, which started adopting naturalistic views in epistemology. In a critique of David Chalmer's 'Conscious Mind,' Paul Churchland (2007) makes the comment that:

> Since Quine's landmark 'Two Dogmas of Empiricism,' half a century ago, most philosophers have become chary of expressing any sympathy for the analytic/synthetic distinction itself, or for the equally important second dogma, the dogma that the credibility of some observation statements stands or falls on the immediate character of sensory experience alone, with no dependence on the relations those statements bear to an enveloping web of presumed background knowledge. (pp. 160-161)

In the following, we will attempt to set forth Quine's argumentative strategy against the analytic-synthetic distinction, a central tenet of Carnap's project. In his 'Two Dogmas', Quine undertakes the task to show that the distinction is unclear or vague and needs to be expunged from our philosophical vocabularies, along with providing a

rough sketch of an alternate account of epistemology which would be the basis for his future works. The paper contains two main arguments against the distinction - the argument from circularity and the argument against empiricist reductionism while asserting that both the dogmas - the existence of a clear analytic-synthetic distinction, and the dogma of reductionism, are "at root identical."

2.1 The argument from circularity

Quine puts forth various arguments to demonstrate that the explanation of analyticity in terms of synonymy substitution is fundamentally circular. He considers two classes of analytic statements: the first being logically true statements like "No unmarried man is married," which remain true under any reinterpretation of the non-logical components. The second class contains statements made analytic by substitution of synonyms, such as "No bachelor is married" turning into the logically true statement "No unmarried man is married" when "unmarried man" is substituted for its synonym "bachelor." Quine grants that the first class of logical truths is reasonably well understood. However, he argues that explaining the second class relies on a prior grasp of synonymy, which is just as unclear as analyticity itself.

Quine surveys various attempts to characterize synonymy without circularity and finds them all lacking. First, Quine considers definitions as a supposed basis for synonymy, examining three types: definitions reporting pre-existent usage of terms, explications refining prior concepts, and stipulations introducing new concepts. In the first two cases, a definition "hinges on prior relationships of synonymy" instead of establishing it. Only in the case of stipulative definitions, where concepts are conventionally introduced, does synonymy become 'transparent.' However, such cases are exceptional and irrelevant to the analysis of a general notion of synonymy in natural languages. Next, he looks at characterizing cognitive synonymy in terms of necessary interchangeability *salva veritate* (interchangeability while preserving truth value). However, Quine maintains that interchangeability does not account for

'cognitive synonymy,' as the condition is merely extensional. None of the explanations seem satisfactory because they all implicitly appeal to the unelucidated concept of synonymy, even when attempting to define it.

Following his critique of synonymy-based explanations, Quine delivers the same criticism against Carnap's suggestion to stipulate analyticity in formal languages via explicit "semantical rules." Defining a sentence S as 'analytic-for-language-L' means nothing more than attaching an unelucidated label, without clarifying what 'analytic' or 'analytic-for' is, and instead appeals to a further unclear phrase, 'semantic rules.' Quine's argument is thus that the notion of analyticity within natural languages is vague and unclear, and that explicitly defining this notion within artificial languages requires a prior, non-circular understanding of the term. He argues that such a characterization of analyticity within formal languages would only clarify the notion if the "mental, behavioral, or cultural factors relevant to analyticity" were to be provided beforehand. Thus, Quine requires Carnap to provide empirical or behaviourist criteria for analyticity in order for the notion to be intelligibly employed in artificial languages.

2.2 The Argument Against Empiricist Reductionism

In addition to the circularity critique, Quine's second central argument aims to refute the doctrine of "radical reductionism" he attributes to early empiricists like Carnap. Quine says this is the verificationist theory of meaning: "The meaning of a statement is the method of empirically confirming or infirming it" (p. 35). By serving as a meaning criterion, verificationism allegedly explains analyticity as well - analytic sentences, having no factual content, can be vacuously "confirmed no matter what." So, if the verificationist program fails, it would undermine one proposed elucidation of the analytic/synthetic distinction. Quine points out that the early Carnap in his Aufbau employed the verification theory by holding that every meaningful statement could be translated into a sentence

about sense experience, which he could not sufficiently characterise, leading to a subsequent abandoning of the position.

Furthermore, Quine argues that the dogma of reductionism still survives in the empiricists, in a less radical form, due to their adherence to the view that every isolated synthetic sentence can have a range of possible sensory events which can increase the likelihood of confirming it and disconfirming it. He says that this dogma results from the consideration that an isolated sentence can be confirmed at all, whereas according to Quine, "our statements about the external world face the tribunal of sense experience not individually but only as a corporate body" (p. 38). The only alternative to the thesis that individual sentences can be verified is his account of holism, where the unit of empirical significance is the whole of science, providing an alternate account of epistemology in the last section of the 'Two Dogmas.' In this account, our sentences of theories form an interconnected 'web of beliefs' wherein conflicting evidence can be accommodated, in principle, by a revision of any sentence(s) in the web based on pragmatic concerns such as simplicity and equilibrium. No sentence inherently connects to any particular observation. This web's periphery contains sentences closely related to experiences, making them more prone to revisions. In contrast, the logical and mathematical ones are more central to the web and only remotely related to experience, making them more resistant to revisions due to conflicting evidence.

According to Quine, the doctrine of reductionism survives partly through the appeal of explaining 'necessity' as a lack of empirical content by holding analytic statements to be vacuously confirmed no matter what. However, with confirmation holism, no statement holds independently of experience. So verificationism, in aiming to justify analyticity by linking statements to evidentiary circumstances one-to-one, relies on an overly simplistic theory of confirmation. Furthermore, it provides no resources for drawing the boundary between kinds of statements based on their differential empirical import. Quine concludes that "the two dogmas are, indeed, at root identical" (p. 38) in taking for granted that individ-

ual statements have circumscribed paths of evidential contact. By rejecting this presupposition in favour of holism, he undercuts the verificationist definition of meaning meant to ground analyticity within a broader empirical epistemology.

In his intertwined arguments against the two central tenets of logical empiricism, Quine contends that there is no clear boundary between analytic and synthetic statements. His argument from circularity aims to show the unintelligibility of explaining analyticity via synonymy. Meanwhile, his argument against reductionism tries to undermine the verificationist theory of meaning that allegedly grounds analyticity. According to Quine, these two dogmas of empiricism - the analytic-synthetic distinction and the reductionist project - are rooted in the same mistaken presumption that individual statements can be verified in isolation. With his account of confirmational holism as an alternative, he concludes that the two dogmas are identical. Having discussed the received view originating from Quine's seminal critiques, we now move to assessing whether this standard interpretation accurately captures Carnap's philosophical positions and motivations.

3. Does the Received View hold?

In the current section, we will discuss the following considerations of the Received View: a) Quine's arguments show that Carnap's usage of analyticity within artificially constructed languages is not intelligible since it is not based on a clear notion of analyticity in ordinary languages. b) Furthermore, it shows that Carnap's 'less radical reductionism' is a dogma since he holds a verification theory of meaning where isolated individual sentences are the units of confirmation rather than our 'totality of knowledge.' c) Lastly, in his conception of the web of belief, Quine holds that unlike Carnap's analytic statements which can be vacuously confirmed "come what may ", every statement within the web is subject to revision subject to pragmatic considerations.

3.1 Intelligibility of the analytic-synthetic distinction

Quine argues that a notion is intelligible only if it can be understood through its prior usage in ordinary language, which requires a pre-theoretical linguistic grasp of terms like 'analytic,' grounded in shared behavioral dispositions. His objection that Carnap cannot employ a notion of analyticity defined in terms of semantic rules within an artificial language system without it being previously understood in a natural language emphasises his view that artificial language systems should be grounded in natural languages. In a letter to Carnap, Quine (1943/2020) makes his demand for the empirical criterion as follows:

> It is only by having some general, pragmatically grounded, essentially behaviouristic explanation of what it means in general to say that a given sound- or script-pattern is analytic for a given individual, that we can understand what is intended when you tell us (via semantical rules, say) 'the following are to be analytic in my new language'. Otherwise, your specification of what is analytic for a given language dangles in midair. (p. 338)

Quine's demand is characteristic of his conception of scientific language as a continuation of ordinary language, which is an essential point of departure from Carnap's aims of undertaking scientific reconstructions in artificial languages. In the article "Scope and Language of Science" (1957), he construes linguistic behaviour as an input-output mechanism beginning from sensory stimulations to assimilation and processing of the culture of the community and finally to speech output. According to Quine, our scientific knowledge begins from the layman's understanding and standards of evidence and reality and merely improves on it in terms of careful and systematic reasoning.

Scientific language, then, evolves out of ordinary language usage and maintains intersubjective objectivity by - eliminating egocentric particulars/indicator words (such as 'I', 'you', 'this', 'here' and the like), reframing singular terms to general terms, and preferring an extensional mode of language. Scientific language, thus,

"is in any event a splinter of ordinary language, and not a substitute" (Quine, p. 9). Furthermore, pragmatic values such as theoretical simplicity and clarity can influence scientific language. These also determine the acceptable ontology of the scientific domain arrived at by the current state of scientific enquiry. Quine points out that such considerations and determinations are not arrived at in an a priori manner but by the actual practice of science and, hence, are tentative and revisable.

Carnap's analysis of analyticity, by contrast, is framed within his method of 'explication,' which he defines as "the transformation of an inexact, prescientific concept, the explicandum, into a new exact concept, the explicatum" (1950b). This procedure attempts to clarify and replace a previously understood vague concept in everyday or scientific use into a more formal, explicitly defined concept by incorporating it into a "well-constructed system of scientific either logico-mathematical or empirical concepts." Notably, in the same work, published a year before Quine's *Two Dogmas*, Carnap acknowledges a related concern, arguing that although an exact characterization of the prescientific explicandum may not always be possible, meaningful discussion still requires that we "do all we can to make at least practically clear what is meant as the explicandum.[2]"

In response to Quine's criticisms in the Two Dogmas, Carnap (1963) considers the demand of an empirical criterion for analyticity within natural languages as the requirement of a clear pre-theoretic understanding of analyticity as the basis of his explication into the formal notion of analyticity within artificial languages. For instance, Tarski's notion of 'truth' within a formalised system is an explication of the pre-theoretic and pragmatic notion of 'truth' in

2 Quine frames explication slightly differently, focusing on preserving the clarity of useful contexts within a term's ordinary usage while refining its application elsewhere. As he puts it: "Any word *worth explicating* has some contexts…clear and precise enough to be *useful*" (1951, p. 25). I thank an anonymous reviewer for pointing out that Carnap's notion of practical clarity of the explicandum and Quine's emphasis on its utility are related but distinct concerns.

ordinary language, and it is such a pre-theoretic understanding of 'analyticity' that Carnap says Quine is demanding. Thus, Carnap sees Quine's criticisms as pointing to the lack of a clear pragmatic explicandum of analyticity and not as a criticism of his explication of the notion in terms of semantic rules used to define analyticity. He suggests that such a requirement of a clear priori explicandum is unnecessary and that he "would agree with his [Quine's] basic idea, namely, that a pragmatical concept, based upon an empirical criterion, might serve as an explicandum for a purely semantical reconstruction" (p. 919) but that it is not necessary to justify introducing a concept of pure semantics by providing a corresponding pragmatic concept.

In a paper published as a response to the Quinean demand, Carnap (1955) attempts to show that one can employ crude empirical tests for determining the inexact pragmatic intension concepts (analyticity, synonymy, etc.) for a historically given natural language in order to provide Quine with a pre-theoretic understanding of analyticity. Carnap thus reiterates that his explication of analyticity is only with reference to explicitly defined semantical language systems, not natural languages and comments that it "shares this character with most of the explications of philosophically important concepts given in modern logic, e.g., Tarski's explication of truth" (1952, p. 66).

So far, Quine requires Carnap to provide a behaviouristic criterion for analyticity in pragmatic usage. As a response, Carnap attempts to provide such an explicandum by giving crude empirical tests for determining intensional concepts like analytic in ordinary languages. One must note, however, that these crude tests cannot be employed to provide an accurate account of analyticity since ordinary language is ambiguous and vague, and it is precisely this vagueness of natural languages that leads Carnap to construct artificial languages where the notion is explicitly defined[3]. In order to

3 see Loomis (2006, pp.502-504) for a discussion on the problem of accurately characterising a vague explicandum such as analyticity in natural languages.

understand Carnap's response as to why a pragmatic behaviouristic criterion for analyticity is not necessary for his usage of the notion in semantical systems, we will discuss his conception of linguistic frameworks and their construction.

Language-systems or linguistic-frameworks for Carnap (1939) are hierarchical structures that can be analysed into three distinct components at different levels of abstraction - pragmatics, semantics, and syntax). Pragmatics is the field which deals with empirical concerns such as a speaker's environment, actions, and usage of signs to talk about objects and properties in the given language. Conducting linguistic research and other empirical investigations, such as physiological analyses of the neural process of linguistic behaviour, studying speaking habits sociologically, etc., take place in this domain. Semantics studies the intra-linguistic relations between expressions of a language and their 'designatum' by abstracting away from the speaker, where names, properties, etc, are assigned to objects. The more precise the semantic rules of a language are, the less vague the way of speaking would be. At this first level of abstraction, the language built up in the sphere of pragmatics can be studied as an "object-language", the investigations of which are conducted in a "meta-language." Syntax is the next level of abstraction from semantics, where the relations of designation are disregarded and concerned only with formal properties of expressions and relations between them. This syntactical system or calculus only consists of formal rules which could be used for constructing proofs or derivations.

Carnap distinguishes between pure syntax and descriptive syntax, along with one between pure semantics and descriptive semantics. The study of the syntax and semantics of a historically given natural language (such as English) and its actual usage is an empirical study, and hence, they are respectively called descriptive syntax and descriptive semantics, which are based on the pragmatics of the language. Thus, the field of Linguistics, which contains the empirical investigations of natural languages, belongs to the sphere of pragmatics (1942, p. 13). On the other hand, pure

syntax and pure semantics are independent of pragmatics and are concerned with the explicit formulation of a system consisting of chosen semantic rules and logical frameworks for its syntax. The construction and analysis of such semantical systems and various syntactical calculi for either a historical language or a constructed one takes place within a metalanguage and comprises an explicitly defined language-system.

While the fields of pure syntax and semantics are independent of pragmatics, the choice of the rules may be guided by pragmatic facts of some language. For instance, if one wishes to construct a syntax and semantics for some natural language like English, it is essential to resolve the inconsistencies and vagueness of its actual (pragmatic) usage and to replace it with a formalised system, which, though 'constructed' may still be guided by the need to approximate the actual usage of the language. Carnap offers the following analogy: "the fact that somebody's garden has the shape of a pentagon may induce him to direct his studies in mathematical geometry to pentagons… the shape of his garden guides his interests but does not constitute a basis for the results of his study" (1942, pp. 13-14). Here, Carnap illustrates that the choice of rules in the constructions of artificial languages may be prompted by some pragmatic aspects of an existing language, but their correctness is not based upon the usage of existing language.

One of the fundamental features of Carnap's philosophy is his proposal of the "Principle of Tolerance", according to which one is free to construct and investigate different logical frameworks (or languages) built for various purposes. In the 'Logical Syntax of Language', he frames the principle as follows: "It is not our business to set up prohibitions, but to arrive at conventions" (1937b, p. 51). Carnap advocates that one can pursue the construction of various kinds of artificial languages, which may differ in their expressive strengths or in their practical efficiency for the required purposes, the choice between them being only a matter of pragmatic justification.

The acceptance of various possible languages through the Principle of Tolerance leads him to another distinction relative to the frameworks. There are, for Carnap, a distinct set of questions which he calls "internal questions," which are stated within a specific language and presuppose the linguistic framework, including the rules for answering them, while "external questions" are those beyond the scope of the language where the framework itself is called into question, such as the choice of the language used, and the rules to be specified for the language (1950a, pp. 21-22). The first kind of question is susceptible to theoretical justification, such as its correctness. In contrast, the latter questions are only a matter of practical choice, subject to our goals and purposes but cannot be strictly theoretically justified i.e. a language cannot be right or wrong, but only useful or expedient in various degrees for a given purpose. We will now consider the two methods of language construction that Carnap describes in his 'Foundations of Logic and Mathematics.'

The construction of artificial languages can be intended for the practical purposes of communication or the formulation of scientific theories while not being bound to be in accordance with the use of actually existing languages. Such a construction of a language-system Z involves formulating its semantical rules S and the syntactical rules C. In the first method, one may begin the construction by first laying down the semantic rules of the system, by determining the forms of sentences to be used within the language and then go on to construct the calculus C for the system, which begins with the classification of signs employed and determines the rules of transformation within the language. In such a construction, we are bound by the constraint that the semantic rules S should be a true interpretation of the calculus C[4]. In the second method, we begin the construction by first laying down an uninterpreted calculus for the system consisting of the syntactic rules of logical transformations and then go on to add a semantical sys-

4 "We call S an interpretation of C if the rules of S determine truth criteria for all sentences of C; in other words, if to every formula of C there is a corresponding proposition of S" in Carnap (1939, p. 21).

tem S, in which the purely logical signs are now given a designata and a linguistic interpretation. In this method, we are entirely free to choose the logical calculus of a system, the choice of which determines whether its interpretation can yield a useful language.

Thus, we see that in the construction of artificial languages, one may begin by choosing purely formal syntactical rules, then proceed to providing a semantical interpretation of those rules, and finally to the pragmatic usage of the constructed language. Since such constructions involve a freedom to the choice of syntax and semantics, Carnap points out that one is not required to provide a prior pragmatic understanding of any concept (such as analyticity) for its introduction as a purely semantical concept. Despite such conventional constructions, however, Carnap says that the choice of rules for a language is not irrelevant, and points out that:

> We may find that a calculus we have chosen yields a lan¬guage which is too poor or which in some other respect seems unsuitable for the purpose we have in mind. But there is no question of a calculus being right or wrong, true or false. A true interpretation is possible for any given consistent calculus, however the rules may be chosen (1939, p. 27).

In a later work,"The Roots of Reference" (Quine 1974), published shortly after Carnap's death, Quine, however comes to accept a notion of analyticity within natural languages ("stimulus analytic") defined in behaviouristic terms as follows - A sentence is analytic if all the members of a certain linguistic community assent to it, the truth of which they learn during the process of learning the component words of the language. However, Quine asserts that this characterisation of analyticity does not play the same epistemological role as Carnap's, and we will consider this epistemological difference in the latter part of this paper.

In his account of the Quinean demand, Creath (2007b) concedes that a pragmatic understanding of analyticity would ensure that Carnap's artificial languages are not merely formalisms but can be employed as full-fledged languages. He argues that Quine's definition of "stimulus-analytic" itself meets the requirement for

the required behaviouristic criteria. According to Creath, Carnap is not trying to describe an existing natural language or conduct empirical linguistics but is conducting investigations and constructions of possible artificial languages. Given this aim, once any crude behaviouristic criterion for the pragmatic usage of analyticity in natural languages is provided, the notion becomes intelligible in the required sense. Then Carnap can ignore this pragmatic usage and continue to work on his language constructions, thus vindicating him from the Quinean criticisms.

We have seen that Quine objects to Carnap's introduction of a technical notion of analyticity in formal languages without first properly explaining the ordinary language concept. Importantly, however, Carnap's doctrine of analyticity is not meant as a description of any specific natural or scientific language but as a tool in his proposal to construct artificial languages which contain this distinction by semantic rules defined for the specific language, thus precluding any possible general definition of the notion. We shall now address Quine's criticism of reductionism and the verification theory of meaning.

3.2 Confirmational Holism & The Verification Theory

Quine's argument against the verification theory of meaning makes a case against the consideration that individual isolated sentences are the carriers of meaning and can be verified singularly. He says that our (synthetic) sentences are not confirmed individually but with respect to the totality of our knowledge system. Quine articulates this confirmational holism, which makes it the case that any sentence in our knowledge system can be held despite conflicting evidence by changing some other sentence(s) in our theories. In 'Two Dogmas in Retrospect' (Quine 1991), he reduces this requirement that the unit of empirical significance be the whole of science to a large enough cluster of sentences with a "critical semantic mass" which can imply an 'observational categorical.[5]'

5 An 'observational categorical' is an expression connecting an observation sentence specifying an experimental condition, to another specifying the effects of the condition.

It is interesting to note that around two decades prior to the publishing of 'Two Dogmas', Quine had read Carnap's manuscript of the Logical Syntax of Language "as it issued from Ina Carnap's typewriter." (Quine 1986, p. 12). In the published work, Carnap (1934a) explicitly acknowledges that Duhemian holistic considerations apply to the testing of any singular hypothetical sentence and that the testing procedure confronts the entire system of theories by pointing out that:

> … it is, in general, impossible to test even a singular hypothetical sentence. In the case of a single sentence of this kind, there are in general not suitable L-consequences of the form of protocol-sentences; hence for the deduction of sentences having the form of protocol-sentences the remaining hypotheses must be used. Thus the test applies, at the bottom, not to a single hypothesis but to the whole system of physics as a system of hypotheses. (Duhem, Poincare´,p. 318)

In Carnap's language-systems, L-rules are the logical rules of transformation adopted within the language, and the L-consequences are then the logical inferences in the language. Since no individual hypothetical sentence can have logical consequences in the form of observational protocol-sentences, Carnap does not hold that they (empirical sentences) can be individually tested in isolation from other sentences within the framework, contrary to Quine's claims.

Carnap's conception of 'meaning' and 'meaningfulness' of terms and sentences changed considerably throughout his writings. In his early writings such as 'The Elimination of Metaphysics Through Logical Analysis of Language', we find a verificationist criterion of meaning wherein "the meaning of a word is determined by its criterion of application (in other words: by the relations of deducibility entered into by its elementary sentence-form, by its truth-conditions, by the method of its verification)" (1932, p. 63). Thus, the meaning of a word, for instance, an 'apple' is determined by considering the method of verification for the truth conditions of its elementary sentence - "this is an apple."While this criterion was considered an important tenet of logical positivism,

and a part of Quine's critique, it was dropped in Carnap's later works where the criterion (in the above form) does not appear.

In the article 'Testability and Meaning' published in two parts, Carnap (1936, 1937) decisively replaces the earlier verificationist theory attributed to Wittgenstein and chooses confirmation of sentences over verification while pointing out that neither a universal sentence nor a particular sentence can be conclusively verified, but only confirmed to some degree. He converts the earlier logical positivist thesis of verificationism from an assertion to a proposal of constructing scientifically useful languages in which the synthetic statements can, relative to theories, be confirmed or disconfirmed empirically. He then discusses various such constructions of possible languages with varying degrees of confirmation requirements built in the framework.

Carnap advices against the usage of the terms "meaningful" and "meaningless" in the context of existing natural or scientific languages since the expressions "involve so many rather vague philosophical associations" and proposes to replace and reformulate the question of the meaning[6] of a sentence in terms of the methodological character of the statement within the language-system: "whether or not the sentence (and its negation) is verifiable or completely or incompletely confirmable or completely or incompletely testable and the like, according to what is intended by 'meaningful'" (1937, p. 3). He discusses the example of a sentence such as - "This stone is now thinking about Vienna" and says that "our [the Vienna Circle's] mistake was simply that we did not recognise the question as one of decision concerning the form of the language," and instead of talking of a proposal to disallow such statements in scientific languages, they earlier used to assert that the sentence is not merely false but also meaningless, and points that such "careless use of the word 'meaningless' has its dangers."

6 It is to be noted that Carnap's usage of the term 'meaning' is not in a psychological sense, but in Tarski's formalised semantical sense, wherein it is a technical relation between expressions in a language and their 'designatum.'

Carnap's talk of delineating a class of cognitively meaningful sentences is not the pursuit of a theory of meaning but a proposal to construct a criterion to restrict the language of science such that one can formulate an empirical language where synthetic sentences are admitted based on some empirical adequacy conditions. Implicit within such a proposal is the possibility of various kinds of possible language-systems including that of non-empirical languages, and hence he characterises his notion of empiricism by saying that:

> …it is preferable to formulate the principle of empiricism not in the form of an assertion - "all knowledge is empirical" or "all synthetic sentences that we can know are based on (or connected with) experiences" or the like - but rather in the form of a proposal or requirement. As empiricists, we require the language of science to be restricted in a certain way; we require that descriptive predicates and hence synthetic sentences are not to be admitted unless they have some connection with possible observations, a connection which has to be characterised in a suitable way. (1937, p. 33)

The synthetic sentences within Carnap's language-systems are not necessarily connected to experiences but are only the sentences whose truth values are not determined by the rules of the language system itself. It is in his proposal of constructing empirical languages for science that we require them to be considered meaningful only when they are in some sense connected with observations, the connection which is to be characterised by the language suitably. Carnap's search for such a characterisation of a language-relative criterion of empirical significance (or cognitive meaningfulness) spanned throughout his later publications wherein he attempted various approaches (such as - translatability into logical syntax, 'confirmation' relations relative to bodies of reduction sentences, etc.) to construct an account of the empirical adequacy of terms, as well as sentences which can be useful for his formulation of possible empirical languages for the reconstructions of science[7].

7 see Lutz (2017) for an assessment of Carnap's various attempted criterions for characterising the empirical significance of terms and sentences.

Carnap's criterion of confirmability is, thus, one of his proposals for a possible language of science and not the assertion of a claim. Being proposals, they cannot be assigned truth or falsity, and the only factual questions that may be asked of them would be the logical and empirical consequences of accepting such a proposal. Speaking of Carnap's assertion that philosophical theses are devoid of empirical content, Uebel (2019) points out:

> … the theses of the analyticity of logic and arithmetic and the distinction between facts and values are not empirical hypotheses but philosophical theses about possible meaning relations. Such theses are neither descriptive of actual language use nor in any sense ontologically committing. They represent conventions of philosophical analysis to be judged for their fruitfulness in rendering reasonings about logic, arithmetic and ethics clearer. (p. 19)

It is not empirical facts, then, that compel us to distinguish between analytic-synthetic statements or between cognitively meaningful and meaningless sentences. They are merely conventional proposals in Carnap's philosophical analysis to be adopted for representing and clarifying possible relations between sentences in the language. Regarding his doctrine of analyticity, Carnap says that one cannot assign it an experimental or empirical meaning, and that the doctrine, "even though not false, is 'empty' and 'without experimental significance'" (1963, p. 922). Contrary to attributing a strong verificationist theory of meaning to Carnap, we find that he replaced such a view early on with a more nuanced confirmation-based account. Furthermore, his confirmation criteria were proposed as useful constructs for empirical languages, not assertions about how natural languages or scientific languages in use operate. Beyond issues related to meaning and verification, a fundamental disagreement exists concerning the status of logical/mathematical truths and the significance of revising them.

3.3 Revisability of Sentences

Within the Quinean "web of belief", every sentence is potentially subject to a revision when faced with conflicting evidence. Since any testing procedure is confronted by the entire system as a whole, one can choose to accommodate the new evidence by the modification of any sentence within the web. Furthermore, the adjustment of sentences within the field is subject to pragmatic considerations such as simplicity and "considerations of equilibrium." Quine maintains that the revisability of logical or mathematical sentences is of no distinct epistemological status than any other sentences more closely connected to experiences situated at the periphery of the web. He states, "it becomes folly to seek a boundary between synthetic statements, which hold contingently on experience, and analytic statements which hold come what may" (1951, p. 40).

In the holistic considerations of the 'Logical Syntax', Carnap mentions that any conflicting evidence to a hypothetical system must be accommodated by making a change somewhere in the entire system and presents us with a choice to either modify certain statements or postulates within the system or even change the analytic sentences. Commenting on the system of empirical physics, he says:

> No rule of the language of physics is definitively secured; all rules are laid down with the reservation that they may be changed depending on the circumstances, as soon as it seems expedient. This holds not only for the P-rules, but also the L-rules, including those of mathematics. In this respect, there are only differences in degree; it is more difficult to decide to give up certain rules than others. (Carnap 1934a, p. 318)

While such a position is similar to Quine, he disagrees with Quine's characterisation of analytic statements as "held true come what may" and makes a distinction between the two kinds of readjustment within his system - that of changing the entire language and that of changing certain statements within the language whose truth values are not determined by the conventions of the language, i.e. by the rules of logic, mathematics or physics. Echoing Kuhn, he

says that the "change of the first kind constitutes a radical alteration, sometimes a revolution, and it occurs only at certain historically decisive points in the development of science. On the other hand, changes of the second kind occur every minute" (Carnap 1963, p. 921). Carnap clarifies that his notion of analyticity does not apply to the changes of the first kind where a new language is adopted. The notion being language relative, only refers to a single language and that being analytic within one language is different than being analytic in another language. Thus, the analyticity of sentences refers only to their status within the respective language and not to their apparent apriority in our knowledge systems.

In the above discussion of the Received View, we have seen that Quine's demand for the intelligibility of analyticity in natural languages arises from his continuity thesis between ordinary and scientific discourse. Since Carnap is constructing distinct, formalised languages, he does not find it necessary to meet such a demand as his language-systems are not grounded in natural languages and can be formulated by self-chosen rules. Despite this, Richard Creath has shown that the demand can be met by Quine's own behaviouristic criteria, thus making the notion of analyticity intelligible in the required sense. Secondly, we see that Carnap does not hold a notion that individual sentences can be confirmed in isolation from other sentences in a theory. His verification criterion is not an assertion of a theory of meaning but is instead one of his proposals to characterise certain sentences as empirically significant for admitting them in the formulation of an empirical language. Once formulated appropriately, such a criterion would help Carnap distinguish between sentences employed conventionally and factual in nature, within his reconstructed theories in artificial languages. Finally, we see that Carnap's analytic sentences are not confirmed come what may and are instead revisable, amounting to a change of language, showing that the distinction is language-relative. Thus, the received view of the debate rests on shaky ground, as it fails to appreciate the nuances of Carnap's philosophical posi-

tions. We shall now turn to characterising and contextualising the grounds of the dispute between Quine and Carnap.

4. Characterising the dispute

Quine's criticisms of Carnap's philosophy concern various aspects of his scientific reconstructions, such as - the very intelligibility of the notion of analyticity, the tenability of a distinction between questions of facts and of meaning, the epistemological significance of such a distinction, etc. At a conference at the University of Chicago, Carnap, while responding to Quine's criticisms, himself characterises their disagreement as follows:

> Quine and I really differ, not concerning any matter of fact, nor any question with cognitive content, but rather in our respective estimates of the most fruitful course for science to follow. Quine is impressed by the continuity between scientific thought and that of daily life—between scientific language and the language of ordinary discourse—and sees no philosophical gain, no gain either in clarity or fruitfulness, in the construction of distinct formalised languages for science. I concede the continuity, but, on the contrary, believe that very important gains in clarity and fruitfulness are to be had from the introduction of such formally constructed languages. This is a difference of opinion which, despite the fact that it does not concern (in my own terms) a matter with cognitive content, is nonetheless in principle susceptible to a kind of rational resolution. In my view, both programs—mine of formalised languages, Quine's of a more free-flowing and casual use of language—ought to be pursued; and I think that if Quine and I could live, say, for two hundred years, it would be possible by the end of that time for us to agree on which of the two programs had proved more successful. (Stein 1992, p. 279)

Here, Carnap suggests that the difference between his and Quine's positions does not amount to a difference in the empirical content of their positions, but as a difference in two different methodological programs for the analysis of science - Quine's analysis of science within a casual usage of language, and his own program of

constructing artificial languages. Despite both these methodologies being empirically equivalent, Carnap suggests that there can be a rational arbitration made between them in terms of the "fruitfulness" of one of them and that eventually, one of them can be shown as more successful over time.

Within the Quinean program, the analysis of concepts must be grounded in natural languages, and hence the notion of analyticity, which he accepts in the 'Roots of Reference', is a very limited one wherein analytic sentences are those which a linguistic community can assent to, the truth of which is learnt during language acquisition. Such a behaviouristic criterion cannot lead to a fundamental division between the classes of analytic-synthetic sentences, and hence, the notion is epistemically insignificant for him. Carnap's program involves the construction of distinct artificial languages such that one can conventionally introduce a cleavage between analytic and synthetic sentences within the language, regardless of its tenability within natural languages. His language choices are pragmatic in the sense that they cannot be seen as correct or wrong but that their acceptance or rejection "will finally be decided by their efficiency as instruments, the ratio of the results achieved to the amount and complexity of the efforts required" (1950a, p. 40).

Characterising the differences in Quine and Carnap's concept of analyticity, Boghossian (1996) distinguishes between a metaphysical and an epistemological notion of analyticity. In the former, purely the meaning of the terms of an analytic sentence makes it the case that the sentence is true. While in the epistemological sense, the grasp of the meanings allows one to be justified in holding the sentence as true. While Quine charges Carnap of holding both these notions at different times, Carnap is concerned with only the epistemological notion of analyticity in his works. In the following, we will discuss the differences in their views on the epistemological significance of the analytic-synthetic distinction, such as its justificatory role in their respective theories of knowledge. Furthermore, we will characterise the differences between

the two programs, following Carnap, as two different methodologies for the analysis of science.

4.1 Epistemological differences

Quine construes Carnap's project as a foundationalist enterprise that tries to explain the certainty of logico-mathematical sentences and their necessity through a 'linguistic doctrine of logical truth' by considering them analytic. Speaking about the certainty that the logical empiricists grant to logic and mathematical sentences while repudiating non-empirical theories as metaphysical, he comments that "the Viennese solution of this nice problem was predicated on language. Metaphysics was meaningless through misuse of language; logic was certain through tautologous use of language" (Quine 1963, p. 386). He thus believes that Carnap's notion of analyticity is an attempt to maintain an absolute distinction between the special class of logico-mathematical sentences, the truth of which is customarily considered necessary and a priori. Quine asserts that the motivation of Carnap's project arises out of two questions - how do empiricists accept mathematics as meaningful, and how do they account for the necessity of its truth - and says that that "analyticity was his [Carnap's] answer to both" (1991, p. 269).

Quine provides an alternative explanation of the necessity of mathematical truth through his holism, wherein if one is faced with conflicting evidence, they will choose not to revoke a purely mathematical sentence due to a restraining 'maxim of minimum mutilation.' While his naturalistic view does not accommodate a notion of metaphysical necessity, he says that such apparent necessity only arises from "our determination to make revisions elsewhere instead" (p. 270). Accepting such a view of Carnap's project drives Quine to comment that his notion of analyticity through language acquisition in the 'Roots of Reference' will not play the same epistemic role as Carnap's. Through his behaviouristic criterion of analyticity, one can only accept simple cases such as elementary logic and the bachelor example to be considered as analytic, but not the rest of higher mathematics; hence, it is epistemologically insignificant.

Thus, Quine objects to the fact that one cannot maintain an absolute analytic-synthetic distinction between the special class of all mathematical sentences and the rest through his behaviouristic criterion.

In contrast to Quine's view, however, Carnap's usage of analyticity in his reconstructions of science is not a proposal to explain the apparent necessity of logical and mathematical sentences or to maintain an absolute distinction between them and other sentences. We have noted earlier that every analytic sentence within his language systems is revisable, which amounts to a change of the language, and that one can conventionally introduce postulates in a semantical system as analytic. Carnap's formulation of artificial languages for the reconstructions of science in which logic and mathematics are construed as conventionally analytic is instead due to his consideration that empirical science, as it is practised, presupposes the logical and mathematical theorems as tools for formal deductions and technical expedience. Carnap says that the use of mathematics within empirical science, for instance, is "in providing, first, forms of expression shorter and more efficient than non-mathematical linguistic forms and, second, modes of logical deduction shorter and more efficient than those of elementary logic" (1939, p. 44).

The rational reconstruction of science is not meant to provide a general empiricist justification of the scientific methodology but is intended to clarify, analyse and reconstruct the logical structure of scientific theories, which can be represented as axiomatic systems. The mathematical and logical theorems employed in these theories are then seen as formal devices, which "do not contribute to the factual content of the conclusion; they merely help in transforming the premises into the conclusion" (p. 45). We thus see that Carnap's division of the analytic and synthetic statements is not a fundamental distinction between certain kinds of sentences but relative to the specific language being constructed and is not motivated by any epistemological foundationalism which seeks to ground the certainty of logic and mathematics in terms of a truth-by-convention thesis. Instead, it is due to the prin-

ciple of tolerance that "the point of viewing the statements of logic and mathematics as analytic lies in our freedom to choose which system of logic and mathematics best serves the formal deductive needs of empirical science" (Friedman 2006, p. 40).

In his discussion of the controversies in the foundations of mathematics, Carnap (1939) stresses that one can adopt any form of mathematical system, such as classical or intuitionist mathematics, in their language construction, the choice being one of pragmatic utility and the goals of the language usage. He says that the foundational issues within mathematics arise when one construes it as a system of 'knowledge' or an interpreted system rather than merely regarding it as a formal tool for deduction within empirical science. This emphasis shifts one's concern from the 'truth' of the system to that of 'technical expedience' in engineering their language. Thus, the questions of the correctness/truthfulness of one or the other system of mathematics is now replaced by questions such as "which form of the mathematical system is technically most suitable for the purpose mentioned? which provides the greatest safety?" (p. 48) and comments that depending on the requirements, classical mathematics could be used due to its simplicity and technical efficiency. In contrast, intuitionistic mathematics could be preferred for more safety from contradictions.

The epistemological criteria for the distinction between analytic and synthetic statements are now reduced to pragmatic considerations of language planning, thus marking a break from the epistemological tradition[8]. Carnap's logical analysis of science "is in no way concerned with either explaining or justifying our scientific knowledge by exhibiting its ultimate basis" (Friedman 2006, p. 48) and is instead concerned with a new role for philosophy to contribute to the progress of science, while avoiding traditional metaphysical disputes and obscurities. Friedman comments that

8 see Creath (1992, pp. 144-146) for a discussion on this shift from epistemological to pragmatic considerations.

Carnap's logic of science is "not any epistemological project" (at least not in the traditional sense) and says:

> …Carnap is not concerned, as is Quine, with developing a very general empiricist conception of justification or evidence simultaneously embracing scientific knowledge, common-sense knowledge, and logico-mathematical knowledge. Carnap is specifically concerned with the mathematical physical sciences characteristic of the modern period, which are themselves only possible in the first place if we presuppose a certain amount of sophisticated modern mathematics - arithmetic and analysis - for their precise articulation and empirical testing. (p. 50)

4.2 Methodological Differences

While Carnap's project can be protected from the Quinean arguments in the Two Dogmas by asserting that Carnap's reconstructions of scientific theories are not meant as a description of actual scientific methodology but as proposals to reformulate scientific theories in terms of their axiomatic structures, it is not exempt from Quine's (1969) methodological questioning in his 'Epistemology Naturalised' where he asks: "But why all this creative reconstruction, all this make-believe? […] Why not just see how this construction really proceeds? Why not settle for psychology?" Here we see that Quine is rejecting Carnap's aims of reconstruction as a scientifically meaningless piece of fiction and that epistemology should instead be devoted to understanding how scientific construction proceeds from sensory stimulations to the complex forms of scientific theorisations as a part of psychological research.

For Carnap, however, the analysis of science can be conducted in various ways, such as - studying the historical development of scientific activity, studying the dependence of scientific progress on individual work, situating it in societal conditions or studying the various procedures and instruments involved in scientific work - which can be respectively called as the history, psychology, sociology and methodology of science (1938, p. 42). These studies consider "science as a body of actions carried out by certain per-

sons under certain circumstances" and form a part of the 'theory of science.' However, there is another subset of the theory where one can abstract away from the people conducting science and their psychological and sociological conditions to conduct a logical analysis of the linguistic expressions of science which Carnap calls the 'logic of science.'

This is the area of Carnap's interest where he attempts to focus not on scientific 'activity' but on the structure and assertions of scientific theories, which are considered as given in the analysis. In a view which converges towards Quine, traditional epistemology also becomes for him a special chapter in psychology:

> What one used to call epistemology or theory of knowledge is a mixture of applied logic and psychology (and at times even metaphysics); insofar as this theory is logic it is included in what we call logic of science; insofar, however, as it is psychology, it does not belong to philosophy, but to empirical science. (Carnap 1934, p. 6)

Thus, in his framework, traditional epistemological concerns, such as the justification of beliefs, belong to psychological (and not philosophical) investigations, and hence epistemology is to be replaced by his logic of science[9].

The dispute between Quine and Carnap thus amounts to two different philosophical methodologies, each with different goals. Quine proposes to conduct epistemology in a naturalised fashion by the analysis of language and studying the actual usage of scientific concepts. Carnap proposes to go beyond the analysis of actual concepts to explicate and create newer precise concepts through constructing artificial languages. The aim is to construct artificial languages through conventions with precise concepts which can clarify or solve traditional philosophical questions. As an example of the systematising efficiency of this procedure, Carnap uses the example of the scientific concept - 'temperature' which is an

9 See Uebel (2018) for a discussion of Carnap's replacement of epistemology by his logic of science.

explication of the pre-theoretic concept - 'hotness', and says that replacing a phrase like "very hot" by "such and such temperature" within a sentence, allows one to communicate the same sentence more efficiently and systematically (1963, p. 936). Such conceptual engineering is the basis of Carnap's investigations into forming a precise characterisation of confirmation and his later studies in inductive logic.

As Carnap advocates, conceptual engineering is motivated by the actual practice of science and can be considered methodologically naturalistic, since science relies heavily on introducing and clarifying concepts that can be construed as analytic or conventional (Lutz, 2020). Lutz discusses various empirical investigations of scientific methodology that show that scientific practice involves the formulation of new concepts conventionally introduced to replace a pre-theoretically understood vague concept. For instance, Hasok Chang's case study shows that the concept of temperature was constructed and operationalised to be measured in definite values, for which no independent standards exist to judge their correctness (Chang 2004). Thus, 'temperature' is not a discovered entity or property in the world, but "scientists develop and in effect *choose* this concept" (Lutz, 2020, p. 1113). For Carnap, such a process of explication represents a continuous transition of the development and systematisation of our knowledge. While differing in his emphasis on being tethered to natural languages like Quine, Carnap acknowledges a kind of continuity between ordinary discourse and systematic science. He says that "the process of the acquisition of knowledge begins with common sense knowledge; gradually the methods become more refined and systematic, and thus more scientific" (Carnap 1963, p. 934).

Thus, we see that Carnap and Quine are aiming at two different programs in their philosophical analysis. Given our above characterisation, the programs are not necessarily opposed to each other but emphasise the pursuit of different goals for the analysis of science and hence can be complementarily conducted. Returning to Carnap's characterisation of the dispute, at the beginning of the

section, it seems puzzling how he proposes that a rational resolution between the two different programs is possible. Quine is pursuing an epistemology modelled in the scientific tradition itself to understand how scientific theorisation actually proceeds. Whereas Carnap is pursuing a program motivated by the scientific method of introducing and systematising concepts in order to construct more accurate and efficient concepts for use in philosophical and scientific discourse. Given these two approaches, it is hard to understand how a rational comparison can be made to their fruitfulness, as they seem to be, in Kuhnian terms, incommensurable with each other. Nonetheless, we agree with Carnap's suggestion that both programs should be pursued for a philosophically tolerant approach towards the analysis of science and scientific languages.

5. Conclusions

In the paper, we attempted to discuss the received view of the Carnap-Quine debate on the analytic-synthetic distinction and showed that Quine's criticisms involve a lack of engagement with the philosophical aims of Carnap's program. The paper brings out that Carnap's notion of analyticity is not an attempt to ground the truth of logical and mathematical sentences, which are customarily considered necessary and a priori, but is instead a tool to avoid traditional foundational and metaphysical disputes in his programme of constructing scientifically useful artificial languages. Furthermore, we attempted to show that both philosophers converge on a significant number of themes while differing in their respective methodologies for the analysis of science. Finally, we bring out that Quine intends to pursue the investigation of how scientific theorisation is actually conducted, while Carnap aims to analyse the logical structure of theories and formulate more systematic concepts by constructing artificial languages in an attempt to clarify and solve philosophical problems.

Referencias

Boghossian, P. A. (1996). Analyticity Reconsidered. *Noûs, 30*(3), 360–391.

Carnap, R. (1932/1959). The Elimination of Metaphysics through Logical Analysis of Language. En A. J. Ayer (Ed.). *Logical Positivism* (pp. 59-81). Free Press.

Carnap, R. (1934b). On the Character of Philosophic Problems. *Philosophy of Science, 1*(1), 5–19.

Carnap, R. (1936). Testability and meaning. *Philosophy of Science, 3*, 19–47.

Carnap, R. (1937a). Testability and meaning. *Philosophy of Science, 4*, 1–40.

Carnap, R. (1937b). *The Logical Syntax of Language* (A. Smeaton, Trad.). Kegan Paul.

Carnap, R. (1938). Logical Foundations of the Unity of Science. In Neurath et al. (Eds.), *Encyclopedia and Unified Science*. University of Chicago Press.

Carnap, R. (1939). *Foundations of Logic and Mathematics*. University of Chicago Press.

Carnap, R. (1942). *Introduction to Semantics*. Harvard University Press.

Carnap, R. (1950a). Empiricism, Semantics, and Ontology. *Revue Internationale de Philosophie, 4*(11), 20-40.

Carnap, R. (1950b). *Logical Foundations of Probability*. University of Chicago Press.

Carnap, R. (1952). Meaning Postulates. *Philosophical Studies, 3*(5), 65-73.

Carnap, R. (1955). Meaning and Synonymy in Natural Languages. *Philosophical Studies, 6*(3), 33-47.

Carnap, R. (1963). Replies and Systematic Expositions. En P. A. Schilpp (Ed.). *The Philosophy of Rudolf Carnap* (pp. 859-1013). Open Court.

Carnap, R. (1966). *Philosophical Foundations of Physics: An Introduction to the Philosophy of Science*. Basic Books.

Chang, H. (2004). *Inventing Temperature*. Oxford University Press.

Churchland, P. (2007). The Evolving Fortunes of Eliminative Materialism. En B. P. McLaughlin et al. (Eds.). Contemporary Debates in Philosophy of Mind (pp. 160-174). Wiley-Blackwell.

Creath, R. (1991). Every Dogma Has Its Day. *Erkenntnis, 35*(1/3), 347-389.

Creath, R. (1992). Carnap's Conventionalism. *Synthese*, 93(1-2), 141-165.

Creath, R. (2007a). Vienna, the City of Quine's Dreams. En A. Richardson y T. Uebel (Eds.), *The Cambridge Companion to Logical Empiricism* (pp. 16-38). Cambridge University Press.

Creath, R. (2007b). Quine's Challenge to Carnap. En M. Friedman y R. Creath (Eds.), *The Cambridge Companion to Carnap* (pp. 316-335). Cambridge University Press.

Friedman, M. (2006). Carnap and Quine: Twentieth-Century Echoes of Kant and Hume. *Philosophical Topics, 34*(1/2), 35-59

Kant, I. (1998). *Critique of Pure Reason* (P. Guyer y A. Wood, Trads.). Cambridge University Press.

Loomis, E. (2006). Empirical Equivalence in the Quine-Carnap Debate. *Pacific Philosophical Quarterly, 87*(4), 474-491.

Lutz, S. (2017). Carnap on empirical significance. *Synthese, 194*(1), 217–252.

Lutz, S. (2020). Armchair Philosophy Naturalized. *Synthese, 19*(3)7, 1099–1125.

Quine, W. V. (1951). Two Dogmas of Empiricism. *Philosophical Review, 60*(1), 20–43.

Quine, W. V. (1957). The Scope and Language of Science. *The British Journal for the Philosophy of Science, 8*(29), 1–17.

Quine, W. V. (1963). Carnap and Logical Truth. In P. A. Schilpp (Ed.). *The Philosophy of Rudolf Carnap* (pp. 385-406). Open Court.

Quine, W. V. (1969). Epistemology Naturalised. In *Ontological Relativity and Other Essays* (pp. 69-90). Columbia University Press.

Quine, W. V. (1974). *The Roots of Reference*. Open Court.

Quine, W. V. (1986). *The Philosophy of W.V. Quine* (L. Hahn y P. Schilpp, Eds.). Open Court.

Quine, W. V. (1991). Two Dogmas in Retrospect. *Canadian Journal of Philosophy, 21*(3), 265-274.

Quine, W. V. (2020). *Dear Carnap, Dear Van: The Quine-Carnap Correspondence and Related Work* (R. Creath, Ed.). University of California Press (Trabajo original publicado en 1943).

Richardson, A. (1997). Two Dogmas about Logical Empiricism: Carnap and Quine on Logic, Epistemology, and Empiricism. *Philosophical Topics, 25*(2), 145-167.

Stein, H. (1992). Was Carnap Entirely Wrong, after All? *Synthese, 93*(1–2), 275-295.

Uebel, T. E. (1996). Anti-foundationalism and the Vienna Circle's revolution in philosophy. *The British Journal for the Philosophy of Science, 47*(3), 415–440.

Uebel, T. (2018). Carnap's Transformation of Epistemology and the Development of His Metaphilosophy. *The Monist, 101*(1), 57-75.

Uebel, T. (2019). Verificationism and (Some of) its Discontents. *Journal for the History of Analytical Philosophy, 7*(4).

SEGUNDA SECCIÓN

EPISTEMOLOGÍA ANALÍTICA

CAPÍTULO VI

TEORÍAS DE JUSTIFICACIÓN EPISTÉMICA

Miguel Ángel Fernández-Vargas[1]

RESUMEN

Las teorías de la justificación epistémica constituyen una parte central de la epistemología analítica, que es la parte de la filosofía

1 Instituto de Investigaciones Filosóficas, UNAM. Ciudad de México, México. mafv@filosoficas.unam.mx. Miguel Ángel Fernández Vargas es investigador titular en el Instituto de Investigaciones Filosóficas (IIFs) de la UNAM, México. Es doctor en filosofía por el University College de la Universidad de Londres; su área de especialidad es la epistemología, en particular las teorías de la normatividad epistémica y el escepticismo filosófico. Ha publicado artículos sobre esos temas en revistas arbitradas, entre las que sobresalen *Philosophical Issues, Synthese, Teorema, Análisis Filosófico*; ha compilado volúmenes colectivos para editoriales de la UNAM y Oxford University Press. Desde 2008 es coordinador del seminario de epistemología del IIFs-UNAM. Fue miembro del comité de dirección de *Crítica. Revista Hispanoamericana de Filosofía* de 2008 a 2025. Para más información, puede consultarse: https://www.filosoficas.unam.mx/sitio/miguelfernandez.

Agradecimientos: El autor agradece al proyecto PAPIIT "Escepticismo y racionalidad" (IA 400124), de la DGAPA-UNAM, por el apoyo con el que se realizó la presente investigación.

analítica que se encarga del estudio filosófico de la cognición y su normatividad. Este capítulo expondrá brevemente algunos de los temas más importantes que se discuten en las teorías de la justificación epistémica contemporáneas, con el objetivo de ofrecer al lector un trasfondo básico que le permita adentrarse en un estudio ulterior más detallado o especializado de esos temas. La sección 1, explica qué es la justificación epistémica y qué significa que sea un "estatus normativo"; también explica en qué se distingue la justificación epistémica de otros tipos de justificación y cómo contrasta con las "razones pragmáticas" para creer. Finalmente, expone cómo se distingue la justificación de otros dos estatus epistémicos: el conocimiento y las acreditaciones. La sección 2 explica tres dicotomías centrales en teoría de la justificación: la dicotomía justificación prospectiva / justificación doxástica; la dicotomía justificación inferencial / justificación no-inferencial; y la dicotomía conservadurismo / liberalismo. La sección 3 realiza una exposición básica de las teorías clásicas de la estructura de la justificación: las teorías fundacionistas; las teorías coherentistas y las teorías infinitistas. La sección 4 ofrece una exposición del debate internismo / externismo acerca de la justificación epistémica, concentrándose en un par de internismos clásicos: el accesibilismo y el mentalismo, y en un externismo clásico: el fiabilismo. También expone la diferencia entre las dos grandes familias de epistemologías de virtudes: la epistemología de ejecuciones y el responsabilismo de virtudes. Finalmente, la sección 5 expone algunos temas centrales de las teorías de la revocación epistémica: la distinción entre revocadores psicológicos y revocadores normativos; y la distinción entre revocadores refutadores y revocadores socavadores.

Palabras clave: Justificación epistémica; justificación prospectiva; justificación doxástica; justificación inferencial; justificación no-inferencial; conservadurismo epistémico; liberalismo epistémico; fundacionismo; coherentismo; externismo epistémico; internismo epistémico; epistemologías de virtudes; revocadores epistémicos

ABSTRACT

The theories of epistemic justification constitute a central part of analytic epistemology, the area of analytic philosophy that studies cognition and its normativity. This chapter briefly expounds some of the most important themes that are discussed in contemporary theories of epistemic justification, with the aim of providing the reader with a basic background that allows her to make a more detailed or specialized study of those themes. Section 1 explains what epistemic justification is and what is meant by saying that it is a "normative status"; it also explains what distinguishes epistemic justification from other kinds of justification and how it contrasts with "pragmatic reasons" for belief. Finally, it explains what distinguishes justification from two other epistemic statuses: knowledge and entitlement. Section 2 expounds three central dichotomies made in theorizing about epistemic justification: the dichotomy prospective justification / doxastic justification; the dichotomy inferential justification / non-inferential justification; and the dichotomy conservatism / liberalism. Section 3 presents a basic exposition of the three classical theories of the structure of justification: foundationalist theories, coherentist theories and infinitist theories. Section 4 presents an exposition of the debate internalism/ externalism about epistemic justification, focusing on a couple of classical internalisms: Accessibilism and mentalism, and one classical externalism: reliabilism. It also explains the difference between the two families of virtue epistemologies: performance epistemology and virtue responsibilism. Finally, section 5 presents some central topics in theories of epistemic defeat: the distinction between psychological and normative defeaters, and the distinction between rebutting defeaters and undermining defeaters.

Keywords: Epistemic justification; prospective justificacion; doxastic justification; inferential justification; non-inferential justification; epistemic conservatism; epistemic liberalism; foundationalism; coherentism; epistemic externalism; epistemic internalism; virtue epistemologies; epistemic defeaters.

1. ¿Qué es la justificación epistémica?

1.1. La justificación epistémica y otros tipos de justificación

La justificación es una característica que puede predicarse de múltiples manifestaciones de la capacidad que tiene una persona para ser un agente, es decir, de la capacidad que tiene para *hacer cosas*. Las personas podemos realizar acciones, llegar a ciertas conclusiones y formar ciertas creencias, fraguar intenciones, diseñar planes, razonar según ciertos patrones. Al hacer todas esas cosas nos manifestamos como agentes, y de todas esas *manifestaciones agentivas* puede predicarse *justificación*. Los epistemólogos clasifican esta característica como "normativa", lo cual significa que al atribuirla a una manifestación de nuestra agencia estamos haciendo una *evaluación* positiva o negativa de ella, y esto tiene implicaciones acerca de *cómo debemos* normar, conducir o gobernar nuestra capacidad agentiva. Por ejemplo, al decir que la acción de Juan de demandar a la compañía de viajes por incumplimiento está justificada, implicamos que Juan procedió de manera *correcta* al demandarla o que actuó como *debería*; y cuando decimos que su creencia de que se ha contagiado de gripe está injustificada, implicamos que *no debería* creer eso o que lo *correcto* sería que no lo creyera.

Existen distintos tipos de justificación, es decir, distintos tipos de estatus normativos que nuestras manifestaciones agentivas pueden tener; por ejemplo, existe la justificación moral, la justificación epistémica, la justificación legal, la justificación pragmática. Estos tipos de justificación son, en principio, independientes, pues señalan *dominios de evaluación* que son *independientes*. De esto se sigue que una misma manifestación agentiva puede recibir buena calificación con respecto a un tipo de justificación, pero mala con respecto a otro. Por ejemplo, las maniobras de un empresario rapaz para evadir impuestos en un país acosado por la pobreza pueden estar *legalmente* justificadas, pero son *moralmente* reprobables. La creencia de Juan de que su pareja amada cometió el crimen puede estar *epistémicamente* justificada, porque toda la evidencia

disponible apoya esa creencia, pero ser *pragmáticamente* contraproducente, pues hundirá a Juan en la depresión.

Las *teorías de la justificación epistémica* se ocupan exclusivamente del estatus normativo que nuestras manifestaciones agentivas pueden tener en *el dominio de la evaluación epistémica*. Históricamente, las teorías de la justificación epistémica se han concentrado, primordialmente, en la justificación epistémica de un tipo de nuestras manifestaciones agentivas: las creencias. De modo que la mayoría de las teorías de la justificación epistémica que se han formulado son, de hecho, teorías de la justificación epistémica *de las creencias*; de estas teorías nos ocuparemos en el presente capítulo[2].

Si las creencias pueden estar justificadas pragmáticamente, moralmente o de alguna otra manera, las teorías de la justificación epistémica no teorizan sobre esos tipos de justificación, sino sólo sobre la justificación *epistémica* de las creencias. Surge entonces la pregunta: ¿Qué distingue a la justificación epistémica de otros tipos de justificación? La respuesta ortodoxa sostiene que la característica distintiva de la justificación epistémica es su conexión con la *verdad*, entendida como "meta cognitiva". Aquí están algunas enunciaciones de esta concepción:

> La evaluación epistémica se realiza desde lo que podríamos llamar "el punto de vista epistémico". Ese punto de vista se define a partir de la meta [*aim*] de maximizar verdad y minimizar falsedad dentro de un cuerpo de creencias amplio.... Para que una creencia esté epistémica-

2 En la epistemología contemporánea, y en particular en las teorías de la justificación epistémica que vamos a discutir en este capítulo, se entiende por el término "creencia" la actitud que consiste en tomar o aceptar algo como verdadero. En el habla ordinaria el término "creencia" puede usarse con otras connotaciones, por ejemplo, puede usarse para expresar cierto nivel de inseguridad, como cuando alguien afirma: "Creo que todavía hay leche en el frigorífico", tratando de expresar que no está seguro de que quede algo de leche. También hay un uso de "creencia" asociado a las creencias religiosas, como cuando alguien dice: "Los creyentes suelen omitir esas partes de las Escrituras". El uso del término creencia en la epistemología contemporánea no tiene estas connotaciones, y sólo significa la actitud de tomar algo como verdadero. Véase Schwitzgebel, 2023, secc. 1, para una discusión de este y otros aspectos de la noción de creencia, tal como se le usa en la epistemología contemporánea.

> mente justificada tiene que recibir calificaciones altas relativamente a esa meta.... (Alston, 1985, pp. 83-4)
> ... la característica distintiva de esta especie particular de justificación [i.e. la epistémica] es su relación esencial o interna con la meta cognitiva de la verdad. Según esta concepción, las actividades cognitivas están justificadas sólo si y en la medida que están orientadas hacia esa meta.... (Bonjour, 1985, p. 8)
> Para que las evaluaciones sean epistémicas, tiene que ocurrir que las creencias de las personas se evalúen con el estándar de qué tan buen trabajo hacen en cuanto a realizar la meta de creer verdades y no creer falsedades.... (Foley, 1987, 124)

Esta concepción ortodoxa se conoce como "veritista", por el lugar central que otorga a la verdad en la caracterización de aquello que es distintivo de la justificación epistémica. De acuerdo con esta concepción veritista, la conexión de la justifciación epistémica con la verdad consiste en que una creencia obtiene la propiedad de estar justificada sólo si "recibe buenas calificaciones relativamente a", "está orientada hacia" o "hace un buen trabajo con respecto a" la meta de maximizar la formación de creencias verdaderas. Pero estas son metáforas que cubren varias maneras distintas, más específicas y menos metafóricas, de entender esa conexión con la verdad. Por ejemplo, de acuerdo con un tipo de teoría evidencialista de la justificación, la creencia de S de que p está justificada sólo si la evidencia que S tiene apoya a p[3]. En contraste, las teorías fiabilistas de la justificación sostienen que la creenica de S de que p está justificada sólo si es el resultado de un mecanismo fiable de formación de creencias[4]. El hecho de que una creencia se apoye en evidencia que S tenga y el hecho de que sea el resultado de algún proceso fiable de formación de creencias, son hechos diferentes, pero a pesar de ello ambos hechos se conciben como maneras diferentes que aseguran que la creencia "reciba buenas calificaciones" desde "el punto de vista de maximizar la verdad".

3 Véase Feldman & Conee 1985 quienes articulan y defienden este tipo de evidencialismo.
4 Véase en Goldman 1979 la formulación clásica de este tipo de teoría fiabilista.

El veritismo permite trazar un contraste muy claro entre la justificación epistémica y los otros tipos de justificación. Por ejemplo, pensemos en el siguiente caso. Juan cree que una persona acusada no es culpable, pero su creencia se origina no en la evidencia disponible, sino en el enorme apego emocional que Juan tiene con esa persona (quien es su pareja); si Juan no creyera en la inocencia de su pareja, el impacto psicológico devastaría su vida entera. De modo que quizá Juan tenga algún tipo de razón/justificación pragmática para creer en la inocencia del acusado. Pero hay un sentido diferente en el que claramente su creencia no está justificada, y las teorías veritistas, como el evidencialismo o el fiabilismo, permiten validar ese sentido: la creencia de Juan no está justificada en el sentido de que la evidencia disponible no apoya *la verdad* de esa creencia; o bien, no está justificada porque formar creencias a partir de apegos emocionales no constituye un mecanismo que de manera fiable conduzca a la formación de *creencias verdaderas*. La creencia de Juan no "recibe buenas calificaciones" desde "el punto de vista de maximizar la verdad". En un caso como este, es totalmente claro que las razones pragmáticas que el agente pueda tener para creer algo, son irrelevantes para la cuestión de si está *epistémicamente* justificado en creerlo.

Sin embargo, la relación entre la justificación epistémica y los factores pragmáticos y morales puede ser más compleja en otros tipos de casos. En décadas recientes se han estudiado diferentes tipos de "intrusiones" que estos factores pueden tener en la determinación de la justificación epistémica, lo cual ha venido a poner en duda que la evaluación epistémica pueda siempre ser "pura" o independiente de aquellos factores que tradicionalmente se había pensado que eran relevantes sólo para la justificación pragmática o la moral, pero no para la epistémica.

Consideremos un caso típico de "intrusión pragmática" [*pragmatic encroachment*][5]. Imaginemos el siguiente par de casos. En el

5 Caso modificado a partir del caso en Staley, 2005, p. 4. Las discusiones estándar sobre intrusión pragmática, incluida la discusión seminal de Stanley 2005, consideran casos

Caso 1, Juan y su pareja Luisa desea depositar sus cheques quincenales, es viernes en la tarde y si no lo depositan antes del lunes su cuenta se quedará sin fondos. Sin embargo, *nada importante depende* de que tengan fondos en su cuenta. Este hecho, aunado a que están exhaustos por una semana laboral pesada, los hace tomar la decisión de depositar sus cheques el sábado en la mañana, cuando vayan camino al gym; después de todo a Juan le parece recordar (aunque a Luisa no) haber ido hace algún tiempo a ese banco un día sábado. De modo que, con base en ese tenue recuerdo, Juan dice a Luisa: "Los depositaremos mañana, el banco estará abierto".

El Caso 2 es exactamente como el 1, excepto que ahora es *extremadamente importante* que el lunes la cuenta de Juan tenga fondos, pues una cuestión de vida o muerte depende de ello. En este caso, Juan tiene la misma evidencia endeble (su incierto recuerdo al respecto) de que el banco estará abierto el sábado e insiste en concluir: "Depositemos los cheques mañana, el banco estará abierto".

En ambos casos Juan cree que el banco estará abierto el sábado, pero existe la intuición de que mientras en el Caso 1 su creencia está epistémicamente justificada, en el Caso 2 no lo está, a pesar de que en ambos casos tiene la misma evidencia a favor de ella. Puesto que la única diferencia entre los casos es que en el 2 es muy importante por razones *prácticas* que su creencia sea verdadera, mientras que en el caso 1 no lo es, se concluye que este tipo de casos ilustran el hecho de que factores prácticos pueden ser relevantes, de hecho determinantes, de si una creencia está epistémicamente justificada o no. Ocurre aquí una "intrusión" de los factores pragmáticos en la justificación epistémica.

Una manera de explicar en qué consiste la relevancia para la justificación epistémica de los factores pragmáticos en casos de intrusión pragmática, consiste en sostener que esos factores pueden *modificar el estándar* epistémico que el agente tiene que satisfacer para que esté justificado en creer; pero los factores relevantes para

---

en los que factores pragmáticos afectan si un agente *sabe* o no, se trata pues de una intrusión pragmática en el *conocimiento*. En la exposición del texto principal discuto el caso paralelo de intrusión pragmática en la *justificación* epistémica.

determinar si el agente *satisface el estándar* o no, siguen siendo puramente epistémicos (no pragmáticos); por ello, si el agente de hecho satisface el estándar está *epistémicamente* justificado[6]. Volvamos a los casos de Juan y el banco. En el Caso 1 el hecho (que es un factor práctico) de que nada importante dependa de que la creencia de Juan sea verdadera, hace que el estándar que Juan tiene que satisfacer para que su creencia de que el banco estará abierto esté justificada sea un estándar muy bajo, de modo que la endeble evidencia que tiene (que es un factor puramente epistémico) resulta ser suficiente para satisfacer ese estándar bajo y por ello su creencia está epistémicamente justificada. En cambio, en el Caso 2, el hecho (que es un factor práctico) de que hay algo de suma importancia que depende de que su creencia sea verdadera, hace que el estándar que Juan tiene que satisfacer sea muy alto, de modo que la endeble evidencia que tiene (que es un factor puramente epistémico) resulta insuficiente para satisfacer ese estándar alto y por ello su creencia no está epistémicamente justificada.

1.2. Justificación epistémica y otros estatus normativos epistémicos

La justificación epistémica se relaciona con otros estatus normativos que también son epistémicos y que nuestras creencias pueden tener. De esos estatus el que más se ha discutido es el *conocimiento*: así como una creencia puede tener el estatus de estar justificada, también puede tener el de ser conocimiento. Tradicionalmente, la justificación se distingue del conocimiento diciendo que mientras éste es un estatus normativo *fáctico*, aquella no lo es. Esto significa que tener conocimiento de que p *implica la verdad* de p, mientras que estar justificado en creer que p no implica la vedad de p. Por ejemplo, si *sé* que fue Carlos quien comió las galletas, entonces es verdad que fue Carlos; saber que fue él es inconsistente con que

6 Véase Fantl & McGrath 2012 y Roeber 2016.

no haya sido él. En cambio, si sólo estoy *justificado* en creer que fue Carlos, esto es consistente con que pueda haber sido alguien más[7].

En décadas recientes se han discutido otros estatus epistémicos de las creencias, distintos del conocimiento y de la justificación, que se han denominado *acreditaciones* [*entitlements*]. Aunque distintos autores han articulado y defendido distintos tipos de acreditaciones, una característica común a todos ellos es que se trata de estatus que no *son evidenciales*; es decir, cuando un agente está acreditado para creer que p, lo que constituye su acreditación no es que tenga evidencia a favor de p, sino otros factores. Los autores que introducen estas acreditaciones, suelen reservar el término "justificación" para referirse a un estatus normativo que, a diferencia de las acreditaciones, está constituido por tener evidencia a favor de una creencia. Un tipo de acreditaciones que se ha propuesto son las perceptuales, las que se basan en nuestra percepción sensorial del mundo. Éstas no están constituidas porque tengamos evidencia, sino por la fiabilidad de nuestros sistemas perceptivos, en tanto ella es una consecuencia de ciertas funciones representacionales que nuestra percepción ha evolucionado para cumplir[8]. Otro tipo de acreditaciones que se han propuesto son las que nos permiten dar por sentado, de manera epistémicamente correcta, una serie de presuposiciones de nuestras investigaciones disciplinares. Esas presuposiciones son indispensables para conducir las investigaciones en cuestión, pero nunca son ellas mismas objeto de estas mismas investigaciones. Por ejemplo, un historiador que investiga el desarrollo del ejército británico durante el siglo XIX, da por sentado que ese fragmento del pasado realmente existió; o un químico que estudia las propiedades de los compuestos alcalinos, da por sentado que estas sustancias no alteran sus propiedades caprichosa-

7 Recientemente algunos filósofos han cuestionado esta distinción y han argumentado a favor de la posición revisionista según la cual las únicas creencias genuinamente justificadas son aquellas que tienen el estatus de conocimiento, véase Williamson 2000; Sutton 2007, Ichikawa 2014.

8 Véase en Burge 2003 y 2020 una articulación sumamente detallada de este tipo de acreditación.

mente, sino que las mantienen uniformemente en la naturaleza. En esas disciplinas es epistémicamente correcto que sus practicantes hagan esas presuposiciones, no porque hayan descubierto evidencia adecuada en su favor, sino porque ellas *hacen posible* las investigaciones que se proponen realizar y el eventual descubrimiento de verdades al interior de esas disciplinas; por ello, los practicantes de esas disciplinas están *acreditados* para hacer esas presuposiciones[9].

El reconocimiento de estatus epistémicos diferentes de la justificación y el conocimiento, como las acreditaciones recién mencionadas, hace que la terminología se vuelva un asunto complicado. Durante mucho tiempo los filósofos usaron el término "justificación epistémica" para referirse indiscriminadamente a estatus parecidos a los que ahora se identifican como tipos de acreditación; y de hecho muchos filósofos siguen usando el término "justificación" para referirse, por ejemplo, al estatus normativo de nuestras creencias basadas en percepción. No obstante, es importante enfatizar que el reconocimiento de estatus epistémicos diferentes de la justificación constituye un *progreso* filosófico, pues las diferencias entre los distintos tipos de estatus reconocidos son genuinas y epistemológicamente significativas. El uso continuado del término monolítico "justificación", produce una falsa ilusión de uniformidad; que encubre aquellas diferencias epistemológicamente importantes. Una mejor práctica terminológica sería usar tantos términos como resulte necesario con el fin de respetar las diferencias epistémicas sustantivas que es teóricamente fructífero marcar. Esto no significa que una vez que distingamos diversos estatus normativos, antaño cubiertos por el término "justificación", éste terminará por ser una etiqueta vacía, no, no es así. Como veremos en la sección 3 de este capítulo, el término "justificación epistémica" sigue refiriendo a un fenómeno sumamente robusto, sobre el cual se han propuesto teorías complejas y de enorme influencia en la filosofía[10].

9 Véase Wright 2004; 2014, quien ha articulado y defendido este tipo de acreditaciones.

10 Sin embargo, véase Alston 2005, quien aboga por eliminar el término "justificación" del vocabulario filosófico y favorece el uso de etiquetas más precisas, acordes a la diversidad de estatus epistémicos que de hecho existen.

2. Algunas dicotomías usadas en teorías de la justificación epistémica

En la teorización sobre la justificación epistémica de las creencias, se utilizan con mucha frecuencia algunas dicotomías básicas que es necesario comprender. Esas distinciones conforman algo así como un "kit" elemental que cualquiera que se adentre en el estudio de la epistemología contemporánea tiene que llevar consigo para poder entender argumentos complejos en los cuales esas dicotomías se presuponen. En esta sección explicaremos brevemente tres de esas fundamentales distinciones.

2. 1. Justificación prospectiva vs doxástica

Hay una diferencia entre el estado en el que uno *tiene razones* que lo justifican en creer que p, sin que uno de hecho forme la creencia de que p, y el estado en el que uno de hecho *cree que p sobre la base de las razones* que justifican esta creencia; el primer estado se describe diciendo que uno tiene justificación prospectiva, proposicional o *ex ante* para creer que p, y el segundo estado diciendo que uno está doxásticamente justificado en creer que p[11]. Por ejemplo, imaginemos un caso en el que Juan recibe un diagnóstico fiable y correcto acerca de que no se ha contagiado de COVID, pero su inveterada hipocondría lo lleva a desestimar esta evidencia y a no formar la creencia de que no tiene COVID. En este caso Juan tiene justificación prospectiva para creer que no tiene COVID, pero no está doxásticamente justificado en creer que no tiene COVID, pues ni siquiera forma esta creencia; para que estuviera doxásticamente justificado tendría que formar esa creencia sobre la base de la buena evidencia que tiene en su favor.

Esta distinción es de suma importancia en numerosas discusiones. Un ejemplo es el papel que desempeña en comprender el

11 La mayoría de los filósofos que discuten esta distinción todavía usan el término "justificación proposicional" para referirse al primer lado de la distinción. Por buenas razones otros prefieren usar el término "justificación prospectiva" (véase Pryor 2014) o "justificación *ex ante*" (véase Goldman, 1979, p. 21, y Comesaña, 2020, p. 74)

fenómeno de la revocación de la justificación[12]. Un revocador es una razón que socava la justificación que un sujeto tiene para creer que p. Sin embargo, algunos filósofos piensan que estados mentales *irracionales* pueden socavar la justificación de un agente[13]. El caso de Juan del párrafo anterior es un caso de este tipo: su hipocondría produce en él la *duda irracional* de que la prueba de COVID podría estar equivocada. Los teóricos de la revocación que desean reconocer los revocadores irracionales tienen que explicar cuál es la justificación de Juan que queda socavada. Como se sugirió en el párrafo anterior, *prima facie*, la justificación prospectiva de Juan no es afectada por sus dudas irracionales. De modo que si en este caso hay revocación genuina, los defensores de los revocadores irracionales tendrían que concentrarse en argumentar que es la justificación doxástica la que queda revocada. Es controversial si hay genuinos revocadores irracionales, pero aun concediendo que existan, la distinción prospectiva/doxástica, permite circunscribir el daño epistémico que esos (supuestos) revocadores podrían causar, poniendo a la justificación prospectiva fuera de su alcance.

Aunque la distinción justificación prospectiva / doxástica es fácilmente aprehensible a través de ejemplos como los que he señalado, ella misma es objeto de teorización en las teorías de la justificación. Los dos conceptos centrales en esta distinción son, por un lado, el concepto de *tener razones* sin usarlas para formar una creencia en particular, y, por otro, está el concepto de *basar* la creencia sobre las razones que se tengan. La elucidación satisfactoria de ambos conceptos se ha revelado una tarea difícil. Por ejemplo, el concepto de que una creencia C esté basada en una razón R se ha denominado el concepto de "basamento" [*basing*]. A primera vista éste parece ser un concepto puramente *causal*, en el sentido de que cuando la creencia C está basada en la razón R, la razón R juega un papel en mantener a través del tiempo la creencia C. Sin embargo, R puede jugar ese tipo de papel causal y *no* hacer que C esté doxás-

12 La sección 5 de este capítulo se ocupa de explicar este fenómeno.

13 Véase Lackey, 2008, Apéndice A.3.

ticamente justificada; esto puede suceder cuando, a pesar de que R sí juegue un papel causal determinante en mantener C a través del tiempo, esto ocurra gracias a que el agente esté procediendo de manera incompetente o inadecuada. Por ejemplo, Juan puede tener evidencia excelente para creer que el acusado es inocente y de hecho basar su creencia en esa evidencia, en el sentido de que es esa evidencia el factor que sostiene su creencia a través del tiempo. Pero de hecho esa evidencia desempeña este papel causal en la psicología de Juan no porque él aprecie su pertinencia y fuerza sino porque el gurú lector de cartas, a quien Juan siempre consulta en casos como este, le ha ordenado que confíe ciegamente en la evidencia que se presente, sea cual sea. En una situación así es claro que Juan no está justificado en creer que el acusado es inocente, a pasar de que de hecho tiene buenas razones para creerlo y basa su creencia en esas razones, en el sentido causal de basamento. Ejemplos como este hicieron que se añadiera una "condición aretaica" a la relación de basamento, la cual requiere que el agente debe basar de *una manera competente* o *virtuosa* su creencia en las buenas razones que tenga. En el ejemplo anterior, Juan no cumple con esta condición porque basa su creencia en la evidencia de que dispone pero de manera incompetente, siguiendo ciegamente las órdenes de un charlatán[14].

2. 2. Justificación inferencial vs no-inferencial

El sentido común reconoce una distinción entre estar justificado de manera *directa* y estarlo de manera *indirecta*. Por ejemplo, cuando veo que un pedazo de papel se está quemando sobre la mesa, estoy justificado de manera directa en creer eso, pues tengo contacto cognitivo directo con el hecho en cuestión, lo tengo ahí frente a mis ojos. En contraste, cuando veo humo saliendo de detrás de la pila

14 En Turri 2010 se encuentra la propuesta aretaica original sobre la relación de basamento. El volumen Silva y Oliveira 2022 contiene ensayos recientes que discuten los fundamentos de la distinción justificación prospectiva / justificación doxástica; varios de los ensayos exploran la importancia de esta distinción para otros debates en epistemología contemporánea.

de libros que está sobre la mesa y a partir de ello concluyo que un pedazo de papel se está quemando sobre la mesa, estoy justificado en creer esto, pero de un modo indirecto a través de una *inferencia* a partir de otras cosas que estoy justificado en creer, como que aquí hay humo y que esta clase de humo generalmente es producido por fuego consumiendo papel. Esta distinción intuitiva es de suma importancia en la epistemología y se conoce como la dicotomía justificación inferencial / no-inferencial.

La manera más simple de entender esta dicotomía es utilizando casos de *inferencia explícita* y contrastándolos con casos de justificación basada en memoria o percepción en los que la inferencia está claramente ausente. Cuando alguien me pregunta dónde estacioné mi automóvil, mi recuerdo de que lo estacioné en el cajón "A29" viene a mí de manera automática; en este caso mi recuerdo justifica mi creencia de manera inmediata, sin que medie ningún razonamiento. En cambio, cuando un detective reúne información sobre el caso que está investigando, utiliza esa información como premisas a partir de las cuales, mediante un razonamiento, llega a la conclusión de que el sospechoso realmente no pudo estar en la escena del crimen a la hora en la que éste ocurrió. Lo que lo justifica en creer la conclusión a la que llega no es simplemente que tenga los fragmentos de información que de hecho tiene, sino el *razonamiento* que lleva a cabo con esos fragmentos de información.

Al reflexionar sobre el contraste entre los tipos de casos mencionados en el párrafo anterior, se descubre que la característica epistemológica más significativa que los distingue es que en los casos de razonamiento explícito, la justificación de la creencia en la conclusión *depende* de dos cosas: en primer lugar, depende de que las premisas del razonamiento estén *ellas mismas justificadas*, y en segundo lugar, de que la conclusión esté *basada correctamente* en las premisas, por medio de unos "pasos" o un "tránsito" que sea epistémicamente correcto. Por ejemplo, si las premisas que usa el detective en su razonamiento no estuviesen justificadas y fueran disparates que se le ocurren sentado en su sillón, no podrían funcionar como evidencia alguna a favor de la conclusión a la que lle-

ga; asimismo, aunque sus premisas sí estuvieran justificadas, pero la conclusión que extrae de ellas fuera totalmente extravagante, de modo que, no estuviera basada correctamente en ellas, la conclusión tampoco estaría justificada. En contraste, estos dos tipos de dependencia no se presentan en casos de justificación no-inferencial: mi justificación para creer que estacioné mi auto en el cajón "A29" no depende de que otras creencias mías estén justificadas, ni de que haya dado pasos inferenciales para poder basar correctamente esa creencia en aquellas otras, simplemente porque no estoy usando otras creencias mías para llegar a la creencia de que el auto está en el cajón "A29", y, por lo tanto, al formar esa creencia tampoco estoy dando ningún "paso" inferencial de unas creencias a otras. En este tipo de caso la justificación de mi creencia depende de otra cosa, a saber, de mi memoria.

Sin embargo, el tipo de dependencias epistémicas que hemos mencionado como características de la justificación inferencial, no se restringen a casos donde se realiza un razonamiento *explícito*. Por ejemplo, la justificación inductiva es un paradigma de justificación inferencial, se trata de la justificación que obtenemos a partir de haber observado una regularidad en el pasado y esto nos justifica para creer que en casos que aún no hemos observado esa regularidad se mantendrá. Pero las creencias que formamos sobre casos no observados no tienen que ser el resultado de un razonamiento explícito, muchas de ellas ocurren de manera automática; una vez que hemos observado la regularidad relevante en el pasado, las creencias sobre casos aun no observados pueden ocurrir sin que medie un *razonamiento*[15]. Por ejemplo, si hasta ahora he observado una cantidad enorme de diversos cisnes y todos ellos han sido blancos, ello me justifica en creer que el cisne que veré mañana en el zoológico será también blanco. Pero esta creencia no es el resultado de un razonamiento *explícito*, como el del detective que describimos arriba, quien infiere una conclusión sobre el acusado ponderando la evidencia de que dispone; mi creencia sobre el cisne ocurre sin que medie un ra-

15 Véase Hume, 1748/1995, Secc. 5.

zonamiento. No obstante, esta creencia exhibe *el mismo tipo* de *dependencias epistémicas* que exhiben las creencias que son resultado de razonamientos explícitos: la justificación de mi creencia de que el cisne que veré mañana será blanco, depende de que mis creencias de que todos los cisnes que he visto en el pasado han sido blancos estén justificadas, y depende también de que mi creencia sobre el cisne que veré mañana esté basada correctamente en mis creencias sobre los cisnes que he visto en el pasado; aunque este "estar basada" no consiste en el caso presente en un paso inferencial explícito de unas creencias a otras, sino más bien en que la creencia sobre el caso no observado se "apoye" correctamente en las creencias sobre los casos ya observados. Si todas mis creencias pasadas sobre el color de los cisnes que he visto fueran totalmente dudosas (quizá porque las formé en condiciones no-optimas de iluminación), o si, aun estando justificadas, no apoyaran la conclusión de que los cisnes son blancos (quizá porque la mitad de los cisnes que he observado no han sido blancos), entonces yo no estaría justificado en creer que el cisne de mañana en el zoológico será blanco. De modo que lo que define a la justificación inferencial de una creencia son las dos dependencias epistémicas que hemos descrito, sea o no que se realice un razonamiento explícito.

2.3. Liberalismo vs conservadurismo

Algunos filósofos han propuesto que un tipo de dependencia epistémica emparentada con una de las dependencias que caracterizan a la justificación inferencial, se presenta incluso en casos de percepción y memoria, es decir, en casos prototípicos de justificación *no*-inferencial. Ya vimos que cuando la creencia de que p está justificada inferencialmente, esta justificación depende de que las otras creencias en las que *se basa* la creencia de que p estén ellas también justificadas. Pues bien, estos filósofos piensan que hay casos de justificación por percepción y memoria en los que la justificación para creer que p depende de que el agente *tenga justificación* para otras creencias, aun cuando ni siquiera haya formado esas otras creencias y, por lo tanto, aun cuando su creencia de que p no esté *basada* en

ellas. Por ejemplo, sostienen que cuando uno observa una cebra en la jaula de un zoológico, y en consecuencia cree que ahí hay una cebra, la justificación de esta creencia depende en parte de que uno tenga justificación para creer que la situación en la que se encuentra es conducente a formar creencias verdaderas sobre el entorno circundante; por ejemplo, piensan que la justificación de la creencia sobre la cebra depende de que se tenga justificación para creer que no es una mula disfrazada de cebra. Ellos *no* están sosteniendo que la creencia de que hay una cebra tiene que estar *basada* (ni siquiera de manera implícita) en una creencia justificada de que ese animal no es una mula disfrazada de cebra, pues aceptan que la mayoría de nosotros ni siquiera ha formado jamás esta última creencia. De modo que *no* están diciendo que la justificación de la creencia de que ahí hay una cebra sea *inferencial*, en el sentido explicado en la sección anterior. Lo que están señalado es que a pesar de que la justificación de una creencia sea no-inferencial, porque no depende de estar basada en otras creencias justificadas, sin embargo, la justificación de esa creencia puede seguir *dependiendo* de que *haya justificación* para creer otras cosas, y sin estas justificaciones adyacentes la creencia no estaría justificada. En el caso de la creencia sobre la cebra, ordinariamente contamos con información de trasfondo que constituye una justificación para creer que no nos están engañando con un animal disfrazado como cebra; por ejemplo, sabemos que las autoridades de zoológicos generalmente no actúan de esa forma. Lo que ocurre es que no se ha presentado la ocasión apropiada que nos haga formar la creencia que esa justificación justificaría; pero la justificación está ahí, como información de trasfondo. Utilizando vocabulario que hemos aprendido en secciones previas, podemos decir que tenemos justificación *prospectiva* para creer que ese animal que vemos no es una mula disfrazada de cebra (aun cuando no formemos esta creencia), y esto es lo que requiere la posición que estamos considerando para que la creencia de que ahí hay una cebra logre estar justificada mediante nuestra percepción.

La dependencia epistémica que hemos descrito como emparentada, pero importantemente distinta de la dependencia que caracte-

riza a la justificación inferencial, constituye el eje de un debate entre "conservadores" y "liberales" en teoría de la justificación, debate que tiene consecuencias de largo alcance en la epistemología. Los "conservadores" defienden que la dependencia con respecto a justificaciones prospectivas adyacentes se extiende a casos de creencias basadas en percepción y memoria, los "liberales" lo niegan y sostienen que la justificación que proviene de la percepción y la memoria es *inmediata*, pues no exhibe ni las dependencias que caracterizan a la justificación inferencial, ni la dependencia respecto de justificaciones prospectivas que defienden los conservadores[16].

El debate entre estos conservadores y liberales tiene consecuencias importantes para múltiples temas de la epistemología, uno de ellos es el escepticismo filosófico. Una de las premisas típicas que usa el escéptico radical dice que para estar justificado en creer cualquier cosa ordinaria, como que uno tiene manos, es necesario tener justificación para creer que uno no es víctima de un engaño radical e indetectable en que el que parece que uno tiene manos sin tenerlas; por ejemplo, sostiene que es necesario tener justificación para creer que uno no es un cerebro en una cubeta (CEC) estimulado artificialmente para creer falsamente que tiene manos. Nótese que esta premisa escéptica ejemplifica un caso de conservadurismo, o al menos comparte el espíritu de esta posición: sostener que para estar justificado en creer que tengo manos es necesario tener justificación para creer que no soy un CEC, es como sostener que para estar justificado en creer que ese animal es una cebra es necesario tener justificación para creer que no es una mula disfrazada de cebra. De modo que la estructura de la postura conservadora no es en principio hostil a esa premisa del escepticismo radical, lo cual condiciona de maneras muy importantes qué tipo de respuestas al escepticismo radical puede ofrecer un filósofo que abrace el conservadurismo en teoría de la justificación. Por ejemplo, un conservador no puede simplemente rechazar la

16 Uno de los principales defensores del liberalismo en teoría de la justificación es Pryor, 2000; 2004; véase también Huemer, 2007. Wright es uno de los principales defensores del conservadurismo epistémico, véase sus 2000; 2002; véase también a Davies 2004.

condición que impone el escéptico para estar justificado en creer que tengo manos y proceder a explicar por qué parece (falsamente) que ese tipo de condición sea correcta, ya que el conservador postula condiciones de ese mismo tipo, estructuralmente análogas. Así que su respuesta al escéptico más probablemente tendrá que consistir en aceptar la condición que impone el escéptico, o al menos una versión de ella, y proceder a explicar cómo podemos cumplirla[17]. En cambio, un liberal en teoría de la justificación no tiene que ofrecer este tipo de respuesta "directa" al escéptico, pues su postura implica rechazar precisamente *el tipo* de condición que el escéptico impone a nuestra justificación para creer que tenemos manos. Así, el liberal puede tomar ese rechazo como punto de partida para ofrecer un tipo de respuesta no-directa al escepticismo[18].

3. Teorías sobre la estructura de la justificación

Probablemente las teorías de la justificación más discutidas son teorías acerca de cómo las razones deben *estructurarse* para resultar en la justificación de las creencias. Estas teorías suelen clasificarse dentro de un esquema tripartito clásico según el cual hay tres tipos: las teorías fundacionistas, las coherentistas y las infinitistas. Estas tres categorías corresponden a tres respuestas posibles a un problema que fue originalmente planteado por los escépticos de la Grecia Antigua, y que de hecho lleva el nombre de uno de ellos, se trata de "El Trilema de Agripa"[19].

Una forma de presentar este problema echa mano de nuestra práctica común de pedir y ofrecer razones a favor de nuestras creencias. Pensemos en cualquier caso en el que buscamos averi-

17 Esta es la estrategia que desarrolla el conservador Wright 2004.

18 Esta es la estrategia del liberal Pryor 2000, por ejemplo.

19 Agripa fue un escéptico de la Grecia Antigua que vivió durante el siglo I d.c. El Trilema que lleva su nombre es solo uno de varios argumentos escépticos que la tradición le atribuye. Puede verse una de las exposiciones originales del Trilema en Sexto Empírico 1993, Libro I, cap. XV; véase también Aristóteles 1995, Libro I, caps. 1 y 2. Existen innumerables exposiciones contemporáneas del problema, véanse por ejemplo, Zalabardo 2014, secc. 2; Aikin 2011, cap. I, secc. 4.

guar cuál es la justificación que alguien tiene para creer algo. Por ejemplo, digamos que Juan cree que el tratamiento médico que le fue prescrito no lo curará (llamemos a esta "creencia 1"), y al preguntarle por qué cree eso, nos dice que cree que el virus de influenza que tiene es inmune a los medicamentos en cuestión (llamemos a esta creencia "creencia 2"). Volvemos a preguntar por qué cree eso y entonces descubrimos que lo cree porque tuvo un sueño la noche anterior en el que su antiguo médico de la niñez (ya fallecido) le aseguraba que así era. Puesto que su creencia 1 se basa en su creencia 2, al descubrir que la 2 está injustificada, concluimos que la 1 tampoco lo está. Esto sugiere que para que una creencia en p esté justificada sobre la base de alguna consideración q, la creencia en q tiene que estar también justificada. Podemos generalizar esta sugerencia como el "Principio inferencial": La justificación de cualquier creencia proviene del apoyo que le den otras creencias justificadas. Utilizando terminología que explicamos en la sección 2.2, podemos afirmar que de acuerdo con este principio toda justificación de una creencia es *inferencial*.

Si aceptamos el Principio inferencial, entonces se genera un trilema, es decir, un conjunto de tres opciones cada una de las cuales parece inaceptable. Si la creencia en p se justifica sobre la base de la creencia en q, entonces la creencia en q tiene que estar justificada; pero dado el principio inferencial la justificación de la creencia en q tiene que provenir de otra creencia, digamos que de r. Pero entonces la creencia en r también tiene que estar justificada en otra creencia, la cual a su vez también tendrá que estar justificada en otra creencia, y así *ad infinitum*. Este es el primer cuerno del Trilema. Una segunda opción es que la cadena de creencias justificadas no prosiga indefinidamente sino que se convierta en un *círculo*; por ejemplo, si la creencia en r se justifica en la creencia en s y la creencia en s se apoya en alguna de las creencias que ya aparecieron antes en la cadena de justificación. Finalmente, una tercera opción es que la cadena se frene en una creencia particular que no se justifique en ninguna otra creencia.

Los tres cuernos del Trilema parecen inaceptables porque intuitivamente ninguno de ellos puede generar justificación para la creencia con la que inicia una cadena de justificación; es decir, intuitivamente una creencia no puede justificarse mediante una cadena de razones que se prolongue *ad infinitum*, ni por una cadena circular, ni por una cadena que culmine en un mero supuesto (i.e. en una creencia que no esté ella misma justificada). Pero si ninguna de las tres opciones disponibles puede generar justificación para nuestras creencias, entonces obtenemos la conclusión escéptica de que ninguna creencia está jamás justificada.

El razonamiento que conduce a este resultado escéptico es un ejemplo de un tipo de problema que en filosofía se denomina "paradoja". Una paradoja filosófica consiste en un conjunto de premisas que son individualmente verosímiles y parecen estar bien sustentadas, pero que, conjuntamente, conducen a una conclusión totalmente inaceptable, quizá incluso absurda. En el caso del Trilema de Agripa, las premisas verosímiles son el "Principio inferencial" y la tesis de que ninguno de los tres cuernos del trilema genera justificación para nuestras creencias; mientras que la conclusión inaceptable es, precisamente, la conclusión escéptica de que ninguna creencia está justificada. Esta conclusión es "paradójica" en el sentido de que contradice la convicción ordinaria y científica de que tenemos justificación para creer muchas cosas, y sin embargo es el resultado de premisas que también encuentran apoyo en nuestras convicciones ordinarias y científicas.

Una de las tareas de la filosofía a la largo de su historia ha sido enfrentarse a paradojas de este tipo e intentar resolverlas. Los intentos por solucionar paradojas filosóficas generalmente dan lugar a distintas teorías filosóficas, y esto es exactamente lo que ha ocurrido con los intentos por resolver el Trilema de Agripa; los intentos por resolverlo han dado lugar a las *teorías clásicas de la estructura de la justificación*. Cada una de esas teorías tiene como punto de partida el rechazo de alguna de las premisas que genera la paradoja. Las teorías *fundacionistas* rechazan el principio inferencial y articulan distintos enfoques según los cuales la justificación de una

creencia no necesariamente tiene que provenir de otras creencias. Las teorías *coherentistas* rechazan que una cadena de justificación que contenga un círculo no pueda generar justificación; y las teorías *infinitistas* rechazan que una cadena que se prolonga *ad infinitum* sea inaceptable. A continuación, describiré brevemente algunos de los aspectos más importantes de estas tres teorías.

3.1. Fundacionismos

Las teorías fundacionistas de la justificación surgen de rechazar el "principio inferencial" que se usa en la formulación del Trilema de Agripa; según estas teorías la justificación de algunas creencias no deriva de otras creencias. Las teorías fundacionistas sostienen dos tesis centrales. En primer lugar, sostienen que el universo de creencias de un agente se divide en creencias *básicas* y *no-básicas*; la justificación de las creencias no-básicas es, efectivamente, *inferencial*, i.e. depende de que estén basadas en otras creencias justificadas. Pero la justificación de las creencias básicas es *no-inferencial*: no depende de que estén basadas en otras creencias justificadas. En segundo lugar, sostienen que la justificación de toda creencia no-básica depende, directa o indirectamente, de creencias básicas. En general, las teorías fundacionistas no dicen mucho sobre los tipos de inferencia mediante los cuales se justifican las creencias no-básicas a partir de las básicas, ya sea de manera directa o indirecta. La atención de las teorías fundacionistas se ha concentrado históricamente en explicar con gran detalle el concepto de creencia básica. ¿Qué es lo que hace que una creencia sea básica? ¿Qué es lo que hace que esté justificada, si su justificación no deriva de que se base en otras creencias? En la historia del fundacionismo se han ofrecido dos tipos de respuesta a estas preguntas; el primer tipo sostiene que lo que hace que una creencia básica esté justificada es una *característica intrínseca* de la creencia, el segundo tipo sostiene que lo que justifica a una creencia básica es una *característica relacional* de la creencia, es decir una característica que relaciona a las creencias básicas con otras cosas, típicamente con otros estados mentales o con algunas propiedades de la fuente que dio origen a esas creencias. En lo que sigue llama-

ré al primer tipo de fundacionismo "fundacionismo intrínseco" y al segundo "fundacionismo extrínseco".

El fundacionismo intrínseco fue el primero en desarrollarse y por ello con frecuencia se le denomina "fundacionismo clásico"[20]. Un ejemplo paradigmático de esta postura es la concepción de Descartes según la cual lo que dota a una creencia de un grado alto de justificación es su "claridad y distinción"[21], y estas son características intrínsecas de las creencias: para descubrir si una creencia las posee debemos examinar y concentrarnos en la creencia misma, no es necesario atender o considerar nada más. Versiones posteriores de fundacionismo intrínseco postularon otras características emparentadas con la "claridad y distinción" cartesianas, como la *incorregibilidad* o la *indubitabilidad*[22]. Estas características también son poseídas por una creencia en sí misma, y para descubrir que la tiene sólo debemos concentrarnos en la creencia misma. Por ejemplo, la creencia de que *me parece visualmente que hay un vaso con agua frente a mi* es, según este tipo de fundacionismo, indubitable, y para descubrir que lo es sólo tengo que reflexionar sobre esta creencia en sí misma y hacer el intento de ponerla en duda, tratando de imaginar una razón que me hiciera pensar que podría ser falsa.

Uno de los problemas más señalados del fundacionismo intrínseco se refiere a la amplitud del universo de creencias básicas que es capaz de admitir. Nótese que los criterios para que una creencia sea básica son sumamente estrictos; por ejemplo, mi creencia de que *hay un vaso con agua frente a mí* no cuenta como básica porque es *corregible* (por ejemplo, cuando estoy alucinando el vaso y alguien que no está alucinando me informa de mi error), y porque también es *dudable* (por ejemplo, aun cuando sea verdad que hay un vaso frente a mí, puedo recibir información engañosa que me haría dudar que lo hay). Estos criterios tan estrictos tienen como consecuencia que el conjunto de creencias que pueden satisfacerlos

20 Véase, por ejemplo, la discusión de Fumerton 2001.

21 Véase Descartes 1642/1977

22 Véanse las versiones de fundacionismo que defienden Lewis, 1946, cap. VII; Ayer 1950; Chisholm 1982, secc. 1.

sea muy estrecho. En general, tradicionalmente se asume que sólo creencias acerca de las sensaciones propias (como la ya mencionada de que "*me parece* que hay un vaso frente a mí") y acerca de algunos otros estados mentales propios, contarían como básicas, pues sólo tales creencias satisfacen los criterios que están en juego. Esta consecuencia es muy grave, pues de ella se desprende que ninguna creencia sobre *el mundo físico circundante* es básica. El hecho de que el conjunto de creencias básicas resulte ser estrecho o restringido de esta manera es problemático para el fundacionismo intrínseco, porque recuérdese que las creencias básicas tienen un trabajo que cumplir: Son los cimientos de los cuales depende inferencialmente, ya sea directa o indirectamente, la justificación de todas las otras creencias que tengamos; y no es claro cómo es que podrían cumplir satisfactoriamente con este trabajo. A este respecto hay dos problemas más específicos:

1. Es suficiente una breve reflexión para percatarnos de que en nuestras prácticas habituales de formación de creencias muy rara vez formamos creencias sobre nuestros propios estados mentales. En general, la posición por *de fault* es formar creencias sobre el mundo que es independiente la mente propia; por ejemplo, al recibir la estimulación visual correspondiente formo la creencia de que hay un vaso frente a mí, no la creencia de que *me parece* que hay un vaso frente a mí. Desde luego que *podemos* formar este último tipo de creencias acerca de cómo nos parecen sensorialmente las cosas. Pero formar este tipo de creencias requiere que adoptemos una posición reflexiva y nos preguntemos, no *qué hay* frente a mí, sino, *qué me parece* que hay frente a mí. El problema es que, de hecho, en general no adoptamos esa posición reflexiva y, por tanto, no formamos las creencias correspondientes. Pero el fundacionista necesita que esas creencias reflexivas estén disponibles *siempre*, pues sobre ellas descansa todo el edificio de la justificación[23]. ¿Debemos decir, entonces, que las creencias no-básicas de un sujeto

23 Pollock & Cruz 1999, pp. 61-65; Steup 1996, pp. 107-108, discuten este problema para el fundacionismo intrínseco.

no-reflexivo no están justificadas hasta que no realice los ejercicios reflexivos que le lleven a formar las creencias básicas que necesita? Esta parece una consecuencia sumamente revisionista con respecto a nuestras prácticas reales de justificación.

2. Aunque asumiéramos que cada persona forma todas las creencias básicas que le fueran asequibles, mediante los ejercicios reflexivos correspondientes, el fundacionista intrínseco seguiría enfrentando un problema mayúsculo: ¿Cuáles son los vínculos inferencias que permiten justificar las creencias no-básicas a partir de las básicas? Para los casos en los que el contenido de las creencias no-básicas es muy cercano al contenido de las básicas, el fundacionista podría postular reglas de transición que son intrínsecamente racionales; como la regla que permite concluir que *hay un vaso frente* a mí a partir de que *me parece que hay un vaso frente a mí*. Pero conforme los contenidos de las creencias no-básicas se alejan más y más del magro contenido de las creencias básicas, el fundacionista se verá obligado a introducir reglas de inferencia cada vez más complejas hasta que llegue un punto en el que simplemente no es comprensible cómo la justificación de nuestras creencias teóricas más avanzadas puede surgir, en última instancia, del magro contenido de nuestras creencias básicas. Por ejemplo, la creencia en la teoría de la evolución de las especies es una de las creencias mejor confirmadas en las ciencias empíricas. ¿Cómo podemos justificarla si nuestros puntos de partida son únicamente creencias acerca de cómo se nos presentan sensorialmente las cosas? Ningún fundacionista intrínseco consiguió jamás dar una respuesta completa y detallada acerca de cuáles son los vínculos inferenciales entre creencias básicas y no-básicas que permitirían dar respuesta a preguntas de este tipo. Desde luego, si el fundacionista es al mismo tiempo un escéptico que piensa que muy pocas (o en el caso más radical, ninguna) de nuestras creencias no-básicas están justificadas, entonces la incapacidad de dar respuestas positivas a ese tipo de preguntas no es una objeción a su posición, sino una consecuencia de ésta.

Pero la mayoría de los (pocos) fundacionistas que aún existen no son escépticos[24].

Al contemplar estos problemas del fundacionismo intrínseco, es difícil no preguntarse cómo fue que los partidarios de esta posición terminaron acorralándose a sí mismos en una concepción tan empobrecida de creencia básica. En los orígenes del fundacionismo la presión proveniente del escepticismo desempeñó un papel en ese empobrecimiento. Como indiqué más arriba, Descartes fue el padre del fundacionismo en la Época Moderna, y su caracterización de los rasgos que hacen que una creencia sea básica ocurre en el mismo contexto en el que articula su proyecto de la "duda metódica", el cual procreó al escepticismo más formidable de la filosofía occidental. La identificación de la claridad y la distinción como los rasgos que hacen que una creencia esté intrínsecamente justificada ocurre dentro de una secuencia argumental en la que Descartes busca alguna creencia de la cual no pueda dudar[25]. Este vínculo de la noción de creencia básica con la indubitabilidad, o alguna característica similar, persistió en muchos de sus herederos fundacionistas, a pesar de que ellos no perseguían el proyecto cartesiano de la duda metódica. Otro factor que acorraló a los fundacionistas intrínsecos en la pobreza teórica que estamos discutiendo, fue su incapacidad dogmática de concebir la posibilidad de que la justificación de una creencia pudiera provenir de algo que no sea una creencia. Como indicamos al inicio de la presente sección, los fundacionistas rechazan el "Principio inferencial", que dice que la justificación de una creencia sólo puede provenir de *otras* creencias, pues las creencias básicas que postulan contraejemplifican este principio. Sin embargo, rara vez se nota que los fundacionistas intrínsecos siguen asumiendo que la justificación de una creencia básica proviene de *alguna* creencia: es verdad que no proviene de *otras creencias*, pero proviene de *sí misma*, es decir, de cualquiera que sea la característica intrínseca que la convierte en una creencia

24 Véase Hasan y Fumerton 2022.

25 Véase la "Segunda Meditación" de sus *Meditaciones Metafísicas*.

básica[26]. Este dogma, según el cual toda justificación de creencias (básicas y no-básicas) proviene de características pertenecientes a creencias, y que podríamos llamar la "Obsesión doxástica", obligó a estos fundacionistas a buscar el origen de la justificación de las creencias básicas en características intrínsecas de ellas mismas. La Obsesión doxástica les impedía imaginar que la justificación de una creencia básica pudiese provenir de algo que no fuese sus características intrínsecas.

El fundacionismo extrínseco surgió cuando los filósofos se liberaron de la obsesión doxástica y pudieron contemplar la posibilidad de que la justificación de una creencia (básica) podía provenir de algo que no sea una creencia. El fundacionismo extrínseco subraya que lo que es esencial de una creencia básica es el hecho de que su justificación sea *no-inferencial*, es decir, que su justificación no provenga de *otras* creencias, y para asegurar que posea esta característica no es necesario sostener que su justificación debe surgir, de alguna manera, a partir de sus *características intrínsecas*. Para asegurar que la justificación de una creencia básica sea no-inferencial hay al menos otra opción en el espacio de posibilidades, a saber, que provenga de otros ítems externos a la creencia misma, siempre y cuando esos ítems no sean ellos mismos creencias.

Históricamente ha habido dos tipos de fundacionismo extrínseco. El primer tipo sostiene que la justificación de las creencias básicas proviene de estados generados por facultades cognitivas como la percepción, la memoria o la intuición racional. Estos estados *no son creencias*; por ejemplo, mi percepción visual de que hay un vaso con agua frente a mí no es la creencia de que hay tal vaso ahí, pues puedo tener la percepción sin formar la creencia correspondiente (por ejemplo, cuando dudo que mi percepción sea verídica). Lo mismo ocurre con los recuerdos y las intuiciones racionales, no son creencias, pero, de algún modo, tienen el poder de justificar creencias. Algunos proponentes de este tipo de teoría explí-

26 Véase la noción de Alston 1989 de "auto-respaldo", la de Chisholm 1989 de "auto-presentación" y los comentarios de Lehrer 1990, p. 61 sobre el vínculo entre el concepto de creencia básica y el de "auto-justificación".

citamente la clasifican como un tipo de teoría fundacionista[27]. Sin embargo, con el paso de los años han surgido posiciones que han hecho suya esta noción de justificación "básica", "no-inferencial" o "inmediata", pero desligándola de otros compromisos del fundacionismo tradicional. De este modo, han surgido las posiciones denominadas "dogmatismo"[28] y "conservadurismo fenoménico"[29], que adoptan una noción de creencia básica, en el sentido del fundacionismo extrínseco que estamos discutiendo, pero que guardan silencio acerca de otras cuestiones sobre las que el fundacionismo tradicionalmente adoptaba una posición, como por ejemplo la idea de que todas las creencias no-básica debían justificarse, en última instancia, en creencias básicas. Estas nuevas posiciones emparentadas con el fundacionismo extrínseco se interesan más en otras cuestiones, como la elucidación precisa de cuáles son las características de los estados cognitivos perceptuales o intuitivos que los dotan del poder para conferir justificación inmediata a las creencias; o la cuestión acerca del impacto que la noción de justificación "inmediata" que defienden tiene en otros debates de gran escala, como el escepticismo filosófico.

El segundo tipo de fundacionismo extrínseco sostiene que las creencias básicas son justificadas por la fiabilidad del proceso o la etiología mediante la que se formaron. A diferencia del fundacionismo extrínseco anterior, este fundacionismo no busca la justificación de mi creencia de que hay un vaso frente a mí simplemente en mi *estado perceptual* correspondiente, sino en un hecho más complejo del cual ese estado perceptual es una parte, a saber, el hecho "histórico" de que mi creencia es el producto de un proceso de formación de creencias que tiene como insumo mi estado perceptual y que ha mostrado a través del tiempo resultar generalmente en creencias verdaderas. Los defensores de este tipo de teoría reconocen que la fiabilidad también es decisiva para la justificación de las creencias *in-*

27 Por ejemplo, Pollock & Cruz 1999, cap. 2, sec. 4.2; Pollock 2001; Feldman 2003, pp. 70-80.

28 Véase Pryor 2000 y 2004.

29 Véase Huemer 2007.

ferenciales, no sólo para las básicas. Por ello explican la justificación de las creencias inferenciales diciendo que su justificación proviene de procesos que toman como insumos, no estados cognitivos como percepciones o recuerdos, sino *creencias*. Por ejemplo, la justificación de mi creencia de que los vasos de cristal son frágiles es inferencial porque deriva de un proceso de formación de creencias que toma como insumos *otras creencias* (por ejemplo, mis creencias pasadas de que distintos vasos particulares han sido frágiles), y que me lleva fiablemente a la creencia verdadera de que los vasos de cristal en general son frágiles, siempre y cuando aquellas otras creencias individuales hayan sido verdaderas. De este modo, este tipo de teoría tiene la *estructura* tradicional del fundacionismo: Una explicación de la justificación de las creencias básicas (en términos de procesos fiables de formación de creencias que tienen como insumos estados cognitivos que no son creencias), y una explicación de la justificación de creencias inferenciales (en términos de procesos "condicionalmente fiables" que toman como insumos a otras creencias, algunas de las cuales pueden ser básicas)[30].

Nótese que cualquiera de las dos versiones de fundacionismo extrínseco evita el problema que enfrenta el fundacionismo intrínseco acerca de que los contenidos de las creencias básicas que puede admitir son demasiado magros para servir de base al resto del edificio del conocimiento. En efecto, los fundacionismos extrínsecos pueden admitir innumerables creencias acerca *del mundo físico* como creencias básicas. De hecho, cualquier creencia sobre el mundo que pueda ser justificada directamente por la percepción o la memoria o que sea el resultado de un proceso fiable de formación de creencias, puede contar como básica. De este modo, el proyecto de explicar cómo el resto de las creencias sobre el mundo puede justificarse sobre la base de las creencias básicas se vuelve más viable.

30 Goldman 1979 es la presentación clásica de este tipo de fundacionismo extrínseco; véase también su 2009, donde argumenta que este tipo de posición es superior al primer tipo de fundacionismo extrínseco que revisamos arriba.

3.2. Coherentismo e infinitismo

Las teorías coherentistas sobre la justificación pueden verse como si surgieran de rechazar la premisa en la formulación del Trilema de Agripa que dice que cualquier cadena de justificación que sea circular no produce justificación. Detrás de la condena a la circularidad está una idea más básica que podemos llamar el "Principio de la justificación asimétrica", según el cual la justificación "fluye" en una sola dirección: si p justifica a q, entonces q no puede contribuir a la justificación de p. Los razonamientos burdamente circulares violan este principio de manera muy clara. Consideremos el siguiente caso. El juez A le dice al juez B, "El testigo X es fiable y declaró que p, por lo tanto debemos creer que p", y B pregunta "Pero, ¿qué razones tienes para creer que X es fiable?" y A responde: "X siempre dice la verdad, por ejemplo, cuando declaró que p". En este caso la creencia de que X es fiable ayuda a justificar la creencia de que p, y luego la creencia de que p se usa directamente, y sin apoyo independiente, sino solo basada en el testimonio del mismo X, para justificar la creencia de que X es fiable. Los coherentistas están de acuerdo con sus adversarios en que esta forma tan burda de violar el Principio de justificación asimétrica es inaceptable. Pero ese principio puede violarse de maneras mucho más sutiles, y en esos casos los coherentistas sostienen que violarlo no constituye un obstáculo para la justificación. Imaginemos que la creencia de que p se usa para apoyar la creencia de que q, y la creencia de que q para justificar la creencia de que r, y la creencia de que r para justificar la creencia de que m y de esta manera, por conexiones de apoyo que quizá estén mediadas por teorías y presuposiciones de fondo, la creencia de que m resulta que, eventualmente, también ofrece cierto apoyo a p. En este caso el círculo de apoyo no es burdo, al contrario, es muy amplio y al recorrerlo nuestra *comprensión* de las *relaciones probabilísticas* y *explicativas* entre las creencias involucradas mejora, y esto es lo que, de acuerdo con los coherentistas, hace que las creencias que forman parte de ese círculo coherente de apoyo mutuo obtengan su justificación. En otras palabras, cada creencia

deriva su justificación de su membresía en un cuerpo coherente de creencias que *pueden apoyarse simétricamente entre sí*[31].

El compromiso de los fundacionistas con las creencias básicas muestra su aceptación del Principio de justificación asimétrica, pues las creencias básicas mantienen relaciones *asimétricas* de justificación con otras creencias: las básicas contribuyen a la justificación de otras creencias, pero no necesitan recibir su justificación de otras creencias. Por ello los coherentistas rechazan que haya creencias básicas, pues sostienen que *todas* las creencias derivan su justificación de las relaciones de apoyo potencialmente simétricas que tienen con otras creencias[32].

Las teorías coherentistas han enfrentado varios problemas, quizá el más discutido y grave es el que se conoce como "El problema del aislamiento", que consiste en lo siguiente[33]. Las teorías coherentistas no consideran que la *experiencia perceptual* sea una fuente de justificación, la fuente de la justificación es las relaciones explicativas, probabilísticas y de coherencia en general *entre creencias*. Pero un sistema de creencias que es altamente coherente puede estar totalmente desligado del modo como es el mundo. Por ejemplo, imaginemos a un cerebro en una cubeta, que es estimulado artificialmente para que forme un sistema de creencias altamente coherente, pero en el cual todas sus creencias son *falsas*. Según el coherentismo las creencias que son miembros de ese sistema de creencias del cerebro en la cubeta están justificadas, pero este no parece ser un dictamen correcto, sobre todo cuando recordamos un tema que explicamos en la sección 1 de este capítulo. Allí vimos que el rasgo distintivo de la justificación *epistémica* es que nos "aproxima a la verdad", si una creencia está epistémicamente

31 Véase Bonjour, 1985, cap. 5. En Lehrer 2000, caps. 5 y 6 y Thagard 2000 pueden verse concepciones diferentes de coherencia.

32 Es importante notar que los fundacionistas extrínsecos no niegan que las relaciones de coherencia con otras creencias *mejoren* la justificación de una creencia básica; lo que niegan es que *toda* justificación provenga de relaciones con otras creencias: hay un tipo de justificación no-inferencial o "inmediata" que proviene de otras fuentes.

33 Pueden verse discusiones del "Problema del aislamiento" en Bonjour, 1985, caps. 6 y 7; Lehrer 2000, caps. 7 y 8. Véase también Klein y Warfield 1994.

justificada entonces, en algún sentido, está más cercana de la verdad. Ahora bien, como acabamos de notar, un sistema de creencias puede contar con las características de coherencia que postula el coherentista y aun así no acercarnos a la verdad en ningún sentido. Por lo tanto, está en duda que la naturaleza de la "justificación" que emana de cumplir con las condiciones coherentistas sea *epistémica*. El problema es que un sistema de creencias (como el del cerebro en la cubeta que acabamos de describir) puede ser altamente coherente y estar "aislado" del mundo que da contenido a las creencias y determina si son verdaderas o falsas. Una forma de evadir este problema sería concediendo a la experiencia perceptual un papel como fuente de justificación: La percepción constituye un canal de conexión entre nuestra mente y el mundo a través del cual el mundo puede determinar el contenido de nuestras creencias y, de esta manera, contribuir a determinar en qué medida estas creencias están justificadas en tanto representaciones de ese mundo independiente. Sin embargo, conceder este papel "normativo" a la experiencia significaría reconocer una fuente no-inferencial de justificación, y así hacer una concesión clave al fundacionismo.

Las teorías infinitistas de la justificación pueden verse como si surgieran de rechazar la premisa en la formulación del Trilema de Agripa que dice que cualquier cadena de justificación que se prolongue *ad infinitum* es inaceptable. Desde la Antigüedad clásica, el principal obstáculo para la verosimilitud del infinitismo ha sido la idea de que está comprometido con la consecuencia de que para que una creencia de un sujeto esté justificada, este sujeto tendría que "recorrer" una cadena infinita de razones. Puesto que esto es imposible, no parece que el infinitismo pueda ofrecer una teoría positiva de la justificación y parece más bien que conduce a la conclusión escéptica de que ninguna creencia está justificada. Los defensores contemporáneos del infinitismo han intentado desligarlo de esa idea, insistiendo en que lo que es esencial para el infinitismo es la idea de que las cadenas de justificación sean *potencialmente* infinitas, y que ninguna de ellas tiene que de hecho "actualizarse" para que alguna creencia esté justificada. Algunos infinitistas argu-

mentan que este compromiso real de su posición les permite dar cuenta de algunos aspectos de nuestras prácticas reales de justificación. Por ejemplo, pensamos que la justificación es una cuestión de grado, es decir, una creencia puede estar mejor o peor justificada. El infinitismo puede explicar el grado de justificación de una creencia dada C en términos de la extensión del *segmento* de la cadena *potencialmente* infinita de razones sobre el que la creencia se apoye. Este tipo de razón a favor del infinitismo no es muy fuerte, pues no es claro que sus adversarios fundacionistas y coherentista no puedan dar cuanta de los aspectos de nuestras prácticas de justificación en los que el infinitista se concentra[34].

4. El debate internismo/externismo sobre la justificación epistémica

La mayoría de los filósofos utilizan el término "justificación epistémica" de manera equivalente al término "razones epistémicas"; yo mismo he procedido así en las páginas previas. De modo que una teoría de la justificación epistémica viene a ser una teoría de las *razones epistémicas para creer*. Sin embargo, el concepto de *razón para creer* parece tener su nicho en las *prácticas dialécticas* en las que hacemos explícitas nuestras razones para creer, defendemos nuestras creencias mediante esas razones y tratamos de convencer a nuestros interlocutores. En contraste, el concepto de justificación no parece estar *necesariamente* ligado a esas prácticas, por ejemplo, uno puede estar justificado en creer cosas que nunca ha tenido que exponer y menos defender ante nadie. Muchas creencias son así, la mayoría de nuestras creencias basadas en la percepción y en la memoria son así; estamos justificados en tenerlas, pero no hemos tenido que formular y ofrecer nuestras razones a favor de ellas ante ningún adversario u objetor. La actividad de reflexionar sobre esas fuentes de creencias preguntándonos qué nos justifica en cada caso, con el fin de for-

34 Discusiones contemporaneas sobre distintos aspectos del infinitismo pueden ecntrarse en la colección de ensayos de Turri & Klein 2014; véase también el debate entre Klein y Ginet 2014.

mular explícitamente las razones por las que creemos, así sea para nosotros mismos en *foro interno*, no es la regla, sino la excepción.

Desde luego que podríamos sostener que el concepto de *razón para creer* tampoco está *necesariamente* ligado a las prácticas dialécticas. Utilizando terminología que explicamos en la sección 2. 1., podríamos sostener que podemos tener razones *prospectivas* para creer que p, aun cuando de hecho no formemos esta creencia y, por lo tanto, aun cuando no podamos usarla como premisa en una interlocución con una audiencia. Sin embargo, la diferencia entre los casos en los que el estatus epistémico de nuestras creencias deriva de que nos involucremos en el tipo de actividades dialécticas señaladas y los casos en los que no tiene este origen, es una diferencia real e importante que tiene que marcarse de alguna manera. Algunos filósofos tratan de capturarla mediante la distinción entre *el estado* de justificación y *la actividad* de justificar[35]. Aunque las actividades dialécticas son una manera de lograr que una creencia esté justificada, no es la única; como señalé en el párrafo previo, los humanos contamos con facultades y competencias que pueden ser la fuente de la justificación de nuestras creencias sin que intervenga actividad dialéctica alguna. Esta diferencia es un elemento importante en el debate entre teorías *internistas* y *externitas* de la justificación.

4.1. Internismo vs externismo en epistemología

De acuerdo con las teorías internistas de la justificación, la justificación de las creencias está constituida o determinada por factores "internos" al sujeto de la creencia. Distintas teorías internistas conciben de manera diferente esos "factores internos". Hay dos tipos principales de teorías internista. Por un lado, están las *teorías mentalistas*, que conciben los factores internos como estados, procesos y condiciones mentales de los sujetos. Por otro lado, están las *teorías accesibilistas*, que conciben los factores internos como aquellos factores a los que el sujeto de la creencia tiene algún tipo de acceso especial, por ejemplo, aquellos factores a los que un sujeto tiene

35 Véase Audi 1988 y Steup 1996, p. 10.

acceso por "mera reflexión". Así, el internismo mentalista sostiene que los únicos factores relevantes para la justificación de las creencias de S son los estados y condiciones mentales de S, mientras que el internismo accesibilista sostiene que los únicos factores relevantes son aquellos a los que S tiene acceso por mera reflexión[36]. Estas dos posiciones no son equivalentes, porque el conjunto de estados y condiciones mentales de un sujeto no coincide con el conjunto de hechos y condiciones a los que tiene acceso por mera reflexión. Puede haber condiciones mentales que no sean accesibles al agente por mera reflexión; por ejemplo, la operación de un sesgo racista que inclina a una persona a formar ciertas creencias sesgadas no le resulta accesible por mera reflexión. Inversamente, existen hechos que son accesibles para un sujeto por mera reflexión, pero que no son estados ni condiciones mentales en absoluto; por ejemplo, para un agente con competencia aritmética básica el hecho de que 10 x 5 = 50 es accesible por mera reflexión, pero esa relación de identidad no es un estado o condición mental del agente (ni de nadie) sino un hecho matemático abstracto.

Las teorías externistas se oponen a las internistas en la medida que sostienen que *no todos* los factores relevantes para la justificación son internos, en cualquiera de los sentidos distinguidos arriba; es decir, sostienen que al menos algunos factores determinantes de la justificación son externos a los sujetos. El fiabilismo de procesos, que mencionamos en la sección 3 como un ejemplo de fundacionismo, es también un ejemplo de teoría externista en este sentido. En efecto, el fiabilismo sostiene que la fiabilidad de un proceso de formación de creencias es un factor que determina la justificación de las creencias que produce, pero la fiabilidad de un proceso no es un factor interno en ninguno de los sentidos que utilizan las teorías internistas: no es un estado mental del sujeto y tampoco es algo a lo que el agente tenga acceso por mera re-

36 Pryor, 2001, p. 109 y Kornblith, 2005, p. 5 discuten tipos de internismo accesibilista. El evidencialismo de Conee & Feldman 1985 es un tipo de internismo mentalista.

flexión; es un hecho en el mundo que debe descubrirse mediante investigación empírica[37].

Las teorías internistas son más afines que las externistas a la idea de otorgar a la actividad de justificar un papel importante. Aunque ambos tipos de teoría reconocen que una creencia puede estar justificada sin que haya sido justificada por la actividad dialéctica del sujeto que cree, sin embargo, la satisfacción de las condiciones de las cuales el internismo sostiene que depende la justificación de una creencia, coloca al sujeto en una mejor posición para *citar* aquello que justifica su creencia y así para poder *defenderla* en una interlocución; mientras que esto no ocurre con la satisfacción de las condiciones externistas. Por ejemplo, supongamos que Juan cree que el diagnóstico médico que acaba de recibir es correcto, y el diagnóstico se origina en una fuente fiable, a saber, las pruebas clínicas que le fueron practicadas; pero Juan no tiene idea alguna acerca de la fiabilidad del diagnóstico, ni siquiera ha pensado sobre ese asunto. Un tipo de internismo diría que la creencia de Juan no está justificada porque no satisface la condición interna de que Juan tenga una *creencia racional acerca de la fiabilidad* de la fuente de su creencia; en cambio, un externista diría que la creencia de Juan sí está justificada, porque cumple con la condición externa de que su creencia *se origine en una fuente fiable*. Nótese que desde la perspectiva del externista, la creencia de Juan está justificada a pesar de que él está en una posición muy pobre para participar en una interlocución dialéctica acerca de las credenciales epistémicas de su creencia. Por ejemplo, como no sabe que la fuente de su creencia es fiable, tiene pocos materiales para convencer a alguien más de que lo sea, o para defenderla ante alguien que la pone en duda. En cambio, si Juan satisficiera las condiciones del internista para estar justificado, sí podría participar en ese tipo de interlocuciones, pues podría echar mano de su creencia racional de que las pruebas clíni-

37 Además de la teoría de Goldman 1979; 1986 y de las teorías de otros fiabilistas, como Lyons 2009, otras teorías externistas de la justificación pueden encontrarse en la epistemología de virtudes de Sosa 2017, caps. 12 y 13; 2021, cap. 10 y en la de Greco 2014. Véase la sección 4.2 del presente capítulo sobre epistemología de virtudes.

cas en cuestión son fiables y usar en su argumentación todas aquellas razones sobre las que su creencia se base. Por este tipo de razón, la conexión entre el estatus epistémico de nuestras creencias y la habilidad de participar en interlocuciones dialécticas acerca de ellas parece mucho más íntima desde la perspectiva internista que desde la externista.

Tanto los enfoques externistas como los internistas han enfrentado fuertes objeciones. Un grupo de objeciones al externismo rechaza que las condiciones que señala el externista como determinantes de la justificación sean *suficientes*. Por ejemplo, imagínese que un agente posee el poder de la clarividencia, que es perfectamente fiable, pero que opera en su mente de tal manera que no le ofrece nada que pueda citar como razón o evidencia a favor de sus creencias. El veredicto intuitivo es que las creencias clarividentes de tal sujeto no están justificadas, que sus creencias necesitan algo más, además de un origen fiable, para estar justificadas[38]. Otras objeciones contra el externismo sostienen que las condiciones que postula no son *necesarias* para la justificación. Por ejemplo, imagínese a un sujeto que es presa de un escenario como el que se representa en la película *Matrix*: su cerebro es artificialmente estimulado por una supercomputadora y tiene una vida mental interna idéntica a la que tendría si estuviera en un escenario normal, pero todas las creencias que le introyecta la Matrix son *falsas*. La intuición da el veredicto de que las creencias de este sujeto estarían justificadas, a pesar de que los procesos de formación de creencias que las producen no sean fiables. Por lo tanto, la fiabilidad no parece necesaria para la justificación[39].

Por su parte, los internistas también se han enfrentado a serias objeciones, una de las más fuertes consiste en que algunos factores que son incuestionablemente relevantes para la justificación no cumplen con el criterio de ser "internos" en ninguno de los sentidos que reconocen los distintos enfoques internistas. Por ejemplo, las *rela-*

38 Véase Bonjour, 1985, cap. 3.
39 Véase Cohen 1984 y Fumerton 1995.

ciones de apoyo probabilístico entre distintas proposiciones es uno de los factores determinantes de la justificación de muchas creencias empíricas, pero ninguna de estas relaciones son estados mentales de nadie, y la mayoría de ellas no son accesibles por mera reflexión. Por ejemplo, la proposición que describe un peculiar salpullido rojizo de Juan constituye apoyo probabilístico para la proposición que atribuye una rara enfermedad a Juan; pero la relación probabilística entre el salpullido y la enfermedad no es un estado mental de nadie (es una relación que se da objetivamente entre esos estados del mundo, independientemente de que haya mentes que la capten), y tampoco puede descubrirse por mera reflexión (el descubrimiento de esa relación entre salpullido y enfermedad pudo haber requerido realizar una investigación empírica sustantiva)[40].

4.2 Epistemologías de virtudes

Una reacción natural al debate internismo/externismo es que cada posición acierta en algo importante, pero erra en otras cosas que su contrincante se encuentra en mejor posición para explicar. Parecería que, si pudiera lograrse una síntesis de ambos enfoques, que combinara sus aciertos y prescindiera de sus defectos, esto nos daría una teoría superior. Una familia de teorías conocidas como "Epistemologías de virtudes" puede verse como si ofrecieran distintas versiones de síntesis entre enfoques internistas y externistas.

Existen dos grandes familias de epistemologías de virtudes, la *epistemología de ejecuciones* y el *responsabilismo de virtudes*. La epistemología de ejecuciones busca definir los estatus normativos clásicos como la justificación y el conocimiento a partir de la noción de *competencia epistémica*, que se entiende como una disposición para formar de manera fiable creencias verdaderas cuando se ejercita en condiciones adecuadas. Una creencia posee *justificación apta* si es el resultado de una competencia epistémica de este tipo[41]. Podemos ver que esta es una condición externista, porque la fiabilidad

40 Véase Goldman 1999; 2009.
41 Véase Sosa 2003, p. 157.

de una competencia epistémica es un factor externo, en el sentido que ya explicamos en la sección previa de este capítulo. Sin embargo, los epistemólogos de ejecuciones no rechazan los factores internos, sino que buscan darles el lugar que les corresponde. Por ejemplo, aunque los factores internos no figuran en la noción de justificación apta que acabamos de citar, sí figuran en la definición de otros estatus epistémicos, como el de *conocimiento reflexivo*, el cual consiste en tener un tipo de *acceso* apto al hecho de que una creencia particular posee justificación apta[42], y este tipo de acceso apto a la aptitud es una forma de rescatar la insistencia de los internistas en la importancia del acceso a las condiciones que determinan el estatus normativo de las creencias[43].

Por su parte, el responsabilismo de virtudes rescata elementos internistas de manera diferente. Este enfoque utiliza el concepto de *virtud intelectual o epistémica* entre otras cosas para caracterizar los estatus epistémicos valiosos clásicos, *i.e.* el conocimiento y la justificación, pero también otros a los que la epistemología tradicional no ha dedicado suficiente atención, como la comprensión y la sabiduría. Dentro de este tipo de teorías, se conciben las virtudes intelectuales como *rasgos del carácter* intelectual de los agentes, tales como la meticulosidad, la mente abierta y la integridad intelectuales. Los responsabilistas no niegan que estos rasgos de carácter puedan contribuir a la adquisición fiable de creencias verdaderas, pero se concentran en otras de sus características, por ejemplo, en su elemento *motivacional*[44]. De acuerdo con ellos, poseer una de esas virtudes significa, en parte, estar motivado a comportarse de ciertas maneras en ciertos contextos. Por ejemplo, tener la virtud de la meticulosidad significa estar motivado a examinar escrupulosamente las razones que se presenten en una discusión. Poseer virtudes intelectuales es esencial para lograr diversas metas que, según los responsabilistas, son constitutivas de una vida humana

42 Véase Sosa 2007, cap. 2.

43 En Fernández 2016 pueden verse ensayos contemporáneos que discute los fundamentos y las aplicaciones de la epistemología de ejecuciones.

44 Véase, por ejemplo, Zagzebski, 1996, caps. 2, 3 y 4.

plena, y para lograr esas metas la motivación que es parte de las virtudes intelectuales es crucial. Pero ese componente motivacional es un factor interno de las virtudes, y parece serlo en los dos sentidos que reconocen las teorías internistas que revisamos en la sección previa: es una condición mental de los agentes motivados y es una condición a la que *prima facie*, parecen poder tener acceso por mera reflexión. De este modo, el responsabilismo de virtudes conjuga elementos internistas y externistas, pero de un modo diferente del modo como lo hace la epistemología de ejecuciones.

5. Teorías de la revocación de la justificación

Las teorías y enfoques sobre la justificación epistémica que hemos presentado en las secciones previas estudian la justificación de las creencias tomando como situaciones paradigmáticas aquellas en las que una creencia *ya está justificada*, o bien un agente *obtiene* o *logra* justificación para creer algo. Pero la vida cognitiva de las personas contiene no sólo este tipo de episodios *positivos* en los que *obtenemos* justificación para nuestras opiniones, también tiene episodios negativos en los que *perdemos* esa justificación. Por ejemplo, al llegar al lugar donde trabajo, veo el auto de Juan estacionado e infiero que Juan ya está en el edificio, pues, generalmente, cuando su auto se encuentra en el estacionamiento eso indica que Juan ha llegado a trabajar. Con base en esto, mi creencia de que Juan ya está en el edifico está justificada. Sin embargo, unos minutos después de mi llegada, una colega de ambos, María, me comunica que Juan no llegó a trabajar pues está enfermo, y quien llegó en su auto fue su esposa, quien se presentó a recoger unos documentos de la oficina de Juan. Este testimonio de María *revoca* mi justificación inicial, de manera que al recibirlo *dejo de estar justificado* en creer que Juan está en el edificio.

Los filósofos también han teorizado sobre la pérdida de la justificación, concentrándose en el fenómeno que denominan "revocación de la justificación", el cual ocurre cuando la adquisición de información ocasiona que una creencia que estaba justificada, deje de estarlo. Entonces se dice que la información adquirida *revoca* nues-

tra justificación. Esto es lo que ocurre en el ejemplo que describimos en el párrafo anterior. Las teorías de la revocación epistémica se han desarrollado mucho en las últimas dos décadas, en el breve espacio disponible que resta en este capítulo explicaré sólo un par de cuestiones fundamentales que son parte de cualquier teoría de la revocación[45].

Un primer asunto fundamental que abordan las teorías de la revocación es la cuestión de *qué tipos de cosas* pueden causar la revocación de la justificación de una creencia. De acuerdo con una concepción tradicional existen dos tipos de revocadores, los *psicológicos* y los *normativos*[46]. Los revocadores psicológicos se llaman así porque son estados psicológicos como creencias o dudas del sujeto cuya creencia justificada es revocada. En el ejemplo que dimos arriba, la *creencia* que formo cuando acepto el testimonio de mi colega en el sentido de que Juan no está en el edificio y fue su esposa quien condujo su auto, es un revocador psicológico en este sentido. Sin embargo, muchos entre quienes aceptan esta taxonomía de revocadores, usan la categoría de revocador psicológico de un modo más amplio o liberal. Nótese que, en el ejemplo anterior, la creencia que formo sobre el verdadero paradero de Juan, y que actúa como revocador, es una creencia que intuitivamente está *justificada*, pues la formo sobre la base del testimonio de mi colega, quien es en general una persona fiable. Pero muchos teóricos de la revocación piensan que cualquier creencia o duda, *aunque no esté justificada y sea irracional*, puede revocar la justificación de otra creencia. Por ejemplo, si en lugar de formar la creencia sobre el paradero de Juan a partir del testimonio de mi colega, la formara por *wishful thinking*, empujado por la aprensión enorme que tengo al hecho de que Juan asista ese día, pues sé que tan pronto vaya anunciará mi despido. En este caso la creencia de que Juan no está en el edificio no estría justificada y sería irracional, pero mu-

45 En Grundmann 2011 puede verse una revisión general de los temas que abordan las teorías de la revocación. En Brown y Simion 2021 puede encontrarse un conjunto de ensayos sobre temas contemporáneos en teoría de la revocación.

46 Véase, por ejemplo, Lackey 2008, Secc. 2.2 y Apéndices A.2 y A.3.

chos teóricos de la revocación piensan que aún así puede revocar mi creencia justificada inicial[47]. De modo que la categoría de revocadores psicológicos generalmente incluye creencias y dudas tanto justificadas como injustificadas.

La otra categoría tradicional de revocadores es la de "revocadores normativos". De hecho, hay dos categorías diferentes de revocadores normativos que son utilizadas por diferentes autores. De acuerdo con una categoría, un revocador normativo es una razón o evidencia de la cual el sujeto que cree *no es consciente, pero debería ser consiente de ella*[48], y si fuera consciente de ella se convertiría en un revocador psicológico de su justificación. Por ejemplo, un médico está justificado, sobre la base de la mejora que el paciente reporta, en creer que el tratamiento que ha indicado al paciente está funcionando. Sin embargo, los resultados de sus exámenes clínicos más recientes demuestran que de hecho el problema ha continuado avanzando; estos resultados se encuentran en el escritorio del médico, pero él ha omitido revisarlos (en parte porque asume que el tratamiento está funcionando). En este caso los resultados de los exámenes clínicos más recientes son evidencia de la cual el médico *no es consciente, pero debería ser consciente de ella* (pues es parte de su responsabilidad profesional), y si fuera consciente de ella se convertiría en un revocador psicológico de su justificación para creer que el paciente se está recuperando. Esa evidencia proporcionada por los resultados de los más recientes exámenes clínicos es entonces un revocador normativo del primer tipo.

Otros autores han identificado un segundo tipo de revocadores normativos que consisten en evidencia de la cual el sujeto que cree *sí es consciente pero no cree en ella y sin embargo debería creer en ella*, y si creyera en ella se convertiría en un revocador psicológico[49]. Podemos modificar el ejemplo anterior para obtener un caso de este tipo. Supongamos que el médico sí es consciente de los resultados

47 Véase, por ejemplo, Plantinga, 2000, cap. 11, sec. 1; Lackey, 2008, sec. 2.2; Bergman, 2006, cap. 6, sec.2.

48 Véase Goldberg 2017 y 2018, cap.6.

49 Véase Lackey, 2008, p. 45.

de los exámenes clínicos más recientes del paciente, pero se rehúsa a aceptarlos sin una buena razón, sino simplemente por terquedad y envanecimiento: piensa que esos resultados tienen que estar equivocados, pues contradicen su "infalible" juicio. En este caso, el agente es consciente de la evidencia revocadora, se rehúsa a aceptarla, pero *debería aceptarla* porque es correcta y él no tiene una buena razón para rechazarla. Esa evidencia constituye un segundo tipo de revocador normativo.

Un segundo asunto fundamental que abordan las teorías de la revocación se refiere a la cuestión de cuál es *el mecanismo de la revocación*, es decir, *cómo* es que un revocador, sea psicológico o normativo, consigue revocar. De acuerdo con una clasificación que es ya clásica, existen dos tipos fundamentales de revocadores según su mecanismo acción, por un lado, están los "refutadores" y por otro los "socavadores"[50]. Un refutador es un revocador que va en contra de *la verdad* de la creencia cuya justificación revoca. En el ejemplo con el que comenzamos esta sección, la creencia que formo a partir del testimonio de mi colega es un *refutador* de mi justificación para creer que Juan ya se encuentra trabajando en el edificio, porque constituye una razón para pensar que esta creencia *es falsa* y que Juan está enfermo en su casa. En contraste, un *socavador* no va necesariamente en contra de la verdad de la creencia cuya justificación revoca, sino que más bien constituye una razón que ataca la conexión entre la evidencia y la creencia, surgiendo que esa conexión *no es fiable*. Podemos modificar el ejemplo anterior para obtener uno donde se ilustre la acción de un socavador. Supongamos que el testimonio de mi colega María, en lugar de afirmar que Juan está enfermo en su casa y su esposa fue quien condujo su auto, afirmara que Luis, otro colega de la oficina, acaba de comprar un auto que es idéntico al auto de Juan y que hoy fue el primer día que lo ha llevado al trabajo. Este nuevo testimonio *no* me da una razón para pensar que mi creencia de que Juan está en el edifico sea falsa:

50 Esta clasificación fue introducida por Pollock 1970. Véase también Pollock & Cruz 1999, cap. 7, secc. 2.

el auto que vi en el estacionamiento podría haber sido el de Luis y aun así Juan podría también estar ya trabajando en el edificio; es decir, el nuevo testimonio no es un *refutador* de la justificación de mi creencia de que Juan ya está en el edificio. El nuevo testimonio me da más bien una razón para pensar que mi evidencia para creer que Juan ya está en el edificio (que consiste en haber observado lo que parecía ser su auto en el estacionamiento) no constituye una base fiable para creer que Juan ya llegó, pues el testimonio en cuestión introduce la posibilidad de que el auto que realmente observé haya sido el nuevo auto de Luis. El nuevo testimonio *socava* la conexión entre mi evidencia y mi creencia de tal modo que ésta última deja de estar justificada.

Las teorías de la justificación epistémica constituyen probablemente el área más basta de la epistemología contemporánea. El objetivo de este capítulo ha sido presentar de una manera básica algunos de los temas más fundamental que se estudian en esas teorías, y de esta manera ofrecer al lector herramientas que le permitan poder comprender discusiones más avanzadas ya sea sobre esos mismos temas, o bien sobre otros que se encuentran con aquellos íntimamente conectado.

Referencias

Aikin, S. (2011). *Epistemology and the Regress Problem*. Oxford y NuevaYork: Routledge.

Alston, W. (1985). Concepts of Epistemic Justification. En W. Alston, *Epistemic Justification. Essays in the Theory of Knowledge*. Ithaca: Cornell University Press, 1989, pp. 81-114.

_____ (2005). *Beyond "Justification": Dimensions of Epistemic Evaluation*. Ithaca: Cornell University Press.

Aristóteles (1995). Analíticos Segundos. En *Tratados de Lógica*, Trad. Miguel Candel Sanmartín. Madrid: Gredos.

Audi, R. (1988). Justification, Truth and Reliability. En R. Audi, *The Structure of Justification*. Cambridge, England: Cambridge University Press, 1988, pp. 299-331.

Ayer, A.J. (1950). Basic propositions. En M. Black (comp.) *Philosophical Analysis: A Collection of Essays*. Ithaca: Cornell University Press, 1950.

Bergmann, M. (2006). *Justification Without Awareness. A Defense of Epistemic Externalism*. Oxford University Press.

Bonjour, L. (1985). *The Structure of Empirical Knowledge*. Cambridge, Mass.: Cambridge University Press.

Brown, J. y Simion, M. (comps.) (2021). *Reasons, Justification and Defeat*. Oxford: Oxford University Press.

Burge, T. (2003) Perceptual entitlement. *Philosophy and Phenomenological Research*, 67 (3), pp. 503-48.

_____ (2020). Entitlement: The Basis for Empirical Epistemic Warrant. En P. J. Graham y N. J. L. L. Pedersen (comps.), *Epistemic Entitlement*. Oxford University Press, pp. 37-142.

Cohen, S. (1984) Justification and Truth. *Philosophical Studies*, 46, pp. 279-95.

Comesaña, J. (2020). *Being Rational and Being Right*. Oxford: Oxford University Press.

Chisholm, R. (1982) *The Foundations of Knowing*. Minneapolis: University of Minnesota Press.

_____ (1989). *Theory of Knowledge*. Tercera Edición. Englewood Cliffs, N. J.: Prentice Hall.

Davies, M. (2004). Epistemic Entitlement, Warrant Transmission and Easy Knowledge. *Aristotelian Society*, Supplementary Volume, 78 (1), pp. 213-245.

Descartes, R. (1642/1977). *Meditaciones Metafísicas, con objeciones y respuestas*. Trad. Vidal Peña. Madrid: Alfaguara.

Fantl, J. y McGrath, M. (2012). Pragmatic Encroachment: it's not only about Knowledge. *Episteme*, 9 (1), pp. 27-42.

Feldman, R. (2003). *Epistemology*. Upper Saddle River, Nueva Jersey: Prentice Hall.

Feldman, R. y Conee, E. (1985). Evidentialism. *Philosophical Studies*, 48, pp. 15-34.

Fernández Vargas, M. A. (comp.) (2016). *Performance Epistemology: Foundations and Applications*, Oxford: Oxford University Press.

Fumerton, R. (1995). *Metaepistemology and Skepticism*. Lanham: Rowman & Littlefield.

_____ (2001). Classical Foundationalism. En M. R. DePaul (comp.) *Resurrecting Old-Fashioned Foundationalism*. Lanham, Maryland: Rowman & Littlefield, pp. 3-20.

Foley, R. (1987). *The Theory of Epistemic Rationality*. Cambridge, Mass.: Harvard University Press.

Goldberg, S. (2017). Should have known. *Synthese*, 194 (8), pp. 2863–94.

_____ (2018). *To the Best of Our Knowledge: Social Expectations and Epistemic Normativity*. Oxford: Oxford University Press.

Goldman, A. (1979). What is Justified Belief? En G. Pappas (comp.) *Justification and Knowledge: New Essays in Epistemology*. Boston: D. Reidel. pp. 1-25.

_____ (1986). *Knowledge and Cognition*. Cambridge, MA.: Harvard University Press,

_____ (1999). Internalism Exposed. *The Journal of Philosophy*, 96 (6), pp. 271-293.

_____ (2009). Internalism, Externalism and the Architecture of Reasons. *The Journal of Philosophy*, 106 (6), pp. 309-38.

Greco, J. (2014). Justification is not internal. En M. Steup, J. Turri y E. Sosa (comps.) *Contemporary Debates in Epistemology*, segunda edición, Oxford: Blackwell, pp. 325-36.

Grundmann, T. (2011) Defeasibility Theory. En S. Bernecker y D. Pritchard (comps.) *The Routledge Companion to Epistemology*. Oxford y Nueva York: Routledge, pp. 156-66.

Hasan, A. y Fumerton, R. (2022) Foundationalist Theories of Epistemic Justification. *Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/entries/justep-foundational/

Huemer, M. (2007). Compassionate Phenomenal Conservatism. *Philosophy and Phenomenological Research*, 74 (1), pp. 30–55.

Hume, D. (1748/1995). *Investigación sobre el conocimiento humano*. Trad. de Jaime de Salas Ortueta. Madrid: Alianza Editorial.

Ichikawa, J. J. (2014). Justification is Potential Knowledge. *Canadian Journal of Philosophy*, 44 (2), pp. 184-206.

Klein, P. y Ginet, C. (2014). Is Infinitism the Solution to the Regress Problem? En M. Steup, J. Turri y E. Sosa (comps.) *Contemporary Debates in Epistemology*, segunda edición, Oxford: Blackwell. pp. 274-297.

Klein, P. y Warfield, T. (1994). What Price Coherence? *Analysis*, 54, pp. 129-32.

Kornblith, H. (2004). Conditions on Cognitive Sanity and the Death of Internalism. En R. Schantz, (comp.), *The Externalist Challenge: New Studies on Cognition and Intentionality*. New York: Walter de Gruyter, pp. 77-88.

Lackey, J. (2008). *Learning From Words. Testimony as a Source of Knowledge*. Oxford: Oxford University Press.

Lehrer, K. (1990). *Theory of Knowledge*. Boulder, Colo.: Westview Press.

_____ (2000). *Theory of Knowledge*. Segunda edición. Boulder, Colo.: Westview Press

Lewis, C. I. (1946). *An Analysis of Knowledge and Valuation*. LaSalle, Ill.: Open Court.

Lyons, J. C. (2009). *Perception and Basic Beliefs. Zombies, Modules and the Problem of the External World*. Oxford: Oxford University Press.

Plantinga, A. (2000). *Warranted Christian Belief*. Nueva York: oxford University Press.

Pollock, J. L. (1970). The structure of epistemic justification, *American Philosophical Quarterly*, 4, pp. 62-78.

_____ (2001). Non-doxastic Foundationalism. En M. R. DePaul (comp.) *Resurrecting Old-Fashioned Foundationalism*. Lanham, Maryland: Rowman & Littlefield, pp. 41-58.

Pollock, J. L. y Cruz, J. (1999). *Contemporary Theories of Knowledge*. Segunda edición. Lanham: Rowman and Littlefield.

Pryor, J. (2000). The Skeptic and the Dogmatist. *Nous*, 34 (4), pp. 517-549.

_____ (2001). Highlights of Recent Epistemology. *British Journal for the Philosophy of Science*, 52 (1), pp. 95-124.

_____ (2004). What's Wrong with Moore's Argument? *Philosophical Issues*, 14, pp. 349–378.

_____ (2018). The Merits of Incoherence. *Analytic Philosophy*, 59 (1), pp.112–141.

Roeber, B. (2016). The Pragmatic Encroachment Debate. *Nous*, 52 (1), pp. 171-195.

Schwitzgebel, E. (2023). Belief. *Stanford Encyclopedia of Philosophy*: https://plato.stanford.edu/entries/belief/

Sosa, E. (2003). Beyond Internal Foundations to External Virtues. En E. Sosa y L. Bonjour, *Epistemic Justification: Internalism vs. Externalism, Foundations vs.Virtues*, Malden: Blackwell.

_____ (2007). *A Virtue Epistemology. Apt Belief and Reflective Knowledge*, Vol. I. Nueva York: Oxford University Press.

_____ (2017). *Epistemology*. Princeton: Princeton University Press.

_____ (2021). *Epistemic Explanations. A Theory of Telic Normativity and What it Explains*. Oxford: Oxford University Press.

Sexto Empírico (1993). *Esbozos Pirrónicos*. Trad. Antonio Gallego y Teresa Muñoz Diego. Madrid: Gredos.

Silva, P. y Oliveira L. R. G. (2022). *Propositional and Doxastic Justification: New Essays on their Nature and Significance*. Nueva York: Routledge.

Stanley, J. (2005). *Knowledge and Practical Interests*. Nueva York: Oxford University Press.

Steup, M. (1996). *An Introduction to Contemporary Epistemology*. New Jersey: Prentice Hall.

Sutton, J. (2007). *Without Justification*. Cambridge, Mass.: MIT Press.

Thagard, P. (2000). *Coherence in Thought and Action*. Cambridge, MA: The MIT Press.

Turri, J. (2010). On the Relationship Between Propositional and Doxastic Justification. *Philosophy and Phenomenological Research*, 80 (2), pp. 312-326.

Turri, J. y Klein, P. (2014). *Ad Infinitum: New Essays on Epistemological Infinitism*, Oxford: Oxford University Press.

Williamson, T. (2000). *Knowledge and its Limits*. Oxford: Oxford University Press.

Wright, C. (2000). Cogency and Question-Begging. Some Reflections on McKinsey's Paradox and Putnam's proof, *Philosophical Issues*, 10, pp. 140-63.

_____ (2002). (Anti)-Sceptics, Simple and Subtle: G. E. Moore and J. McDowell. *Philosophy and Phenomenological Research*, 65, 2, pp. 330-348.

_____ (2004). Warrant for Nothing (and Foundations for Free)? *Aristotelian Society Supplementary* Volume, 78 (1), pp. 167–212.

_____ (2014). On Epistemic Entitlement (II): Welfare State Epistemology. En D. Dodd & E. Zardini (comps.), *Scepticism and Perceptual Justification*. Oxford University Press, pp. 213-247.

Zagzebski, L. (1996). *Virtues of the Mind: An Inquiry into the Nature of Virtue and the Ethical Foundations of Knowledge*. Cambridge: Cambridge University Press,.

Zalabardo, J. (2014). *Conocimiento y escepticismo. Ensayos de epistemología contemporánea*. México: IIFs-UNAM.

CAPÍTULO VII

LAS PARADOJAS ESCÉPTICAS

Santiago Echeverri[1]

RESUMEN

En la vida diaria, solemos entender el escepticismo como una posición que podemos rechazar por considerarla absurda o incoherente.

---

1 Santiago Echeverri es Investigador Titular A en el Instituto de Investigaciones Filosóficas de la UNAM. Realizó un doctorado en Filosofía y Ciencias Cognitivas en el Instituto Jean Nicod de París (2010). Además, tiene una maestría en Ciencias Cognitivas por la Escuela de Altos Estudios en Ciencias Sociales (EHESS) (2005) y una maestría en Historia de la Filosofía por la Universidad de París IV – Sorbona (2004). Obtuvo el pregrado en Filosofía y Literatura en la Universidad de Antioquia (2002). Sus investigaciones se enfocan en la epistemología, la filosofía de la mente y la filosofía de las ciencias cognitivas. En trabajos recientes, ha examinado las paradojas escépticas, las representaciones perceptuales de objetos y la referencia en primera persona. Ha publicado artículos sobre estos temas en revistas internacionales como *Journal of Philosophy, Philosophy and Phenomenological Research, British Journal for the Philosophy of Science, Philosophy of Science y Philosophical Studies*. Fue vicepresidente de la *Sociedad Francoparlante de Filosofía Analítica* (*SOPHA*) (2018–2022) y miembro del comité editorial de la revista *dialectica* (2011–2015). Desde 2024 es el director de *Crítica. Revista Hispanoamericana de Filosofía*.

En la filosofía griega, el escepticismo era más bien una actitud inquisitiva que llevaba a la suspensión del juicio y, en consecuencia, a la tranquilidad del alma. En la filosofía moderna, muchos filósofos veían el escepticismo como una fase del pensamiento que debía superarse antes de sentar las bases firmes para las ciencias. Estas concepciones contrastan con la visión dominante en la epistemología analítica actual, donde el escepticismo suele entenderse como una serie de paradojas que surgen al reflexionar sobre el funcionamiento de conceptos epistémicos ordinarios como CONOCIMIENTO y JUSTIFICACIÓN EPISTÉMICA. Una paradoja escéptica emerge cuando identificamos cierta condición que parece ser necesaria para tener conocimientos o justificación epistémica, pero también parece que dicha condición no se puede cumplir. Este capítulo explica por qué la concepción del escepticismo como una serie de paradojas es central para la epistemología analítica contemporánea e identifica algunos rasgos centrales de las paradojas escépticas sobre el mundo externo. Después de identificar dos tipos de respuestas, se presentan dos paradojas escépticas influyentes basadas en los principios de cierre y de subdeterminación, se señalan algunas diferencias importantes entre ambas paradojas y se examinan críticamente algunas tentativas para resolverlas.

Palabras clave. Paradojas escépticas; principio de cierre del conocimiento; principio de subdeterminación; externismo epistemológico; análisis modales del conocimiento; teorías de las alternativas relevantes; contextualismo epistemológico; autorización epistémica; disyuntivismo epistemológico.

ABSTRACT

In everyday life, we often understand skepticism as a position that one may set aside because of its absurdity or incoherence. In Greek philosophy, skepticism was rather an inquiring attitude that led to the suspension of judgment and, as a result, the freedom from distress. In Modern philosophy, many philosophers viewed skepticism as a phase of thought that had to be overcome before laying firm foundations for the sciences. These pictures differ from the outlook

that dominates current analytic epistemology, where skepticism has normally been understood as a series of paradoxes that arise when we reflect on the workings of ordinary epistemic concepts like KNOWLEDGE and EPISTEMIC JUSTIFICATION. A skeptical paradox emerges when we identify a condition that seems to be necessary to have knowledge or epistemic justification, while it also seems that that condition cannot be met. This chapter explains why the picture of skepticism as a series of paradoxes is central to contemporary analytic epistemology and identifies some of the main features of skeptical paradoxes about the external world. After identifying two types of replies, it presents two influential paradoxes based on the closure and underdetermination principles, identifies some relevant differences between the two paradoxes, and critically examines some attempts at solving them.

Keywords. Skeptical paradoxes; knowledge closure principle; underdetermination principle; epistemological externalism; modal analyses of knowledge; relevant alternatives theories; epistemological contextualism; epistemic entitlement; epistemological disjunctivism.

1. Introducción

En la vida diaria, solemos entender el escepticismo como la posición que adopta una persona caprichosa que opera con estándares epistémicos más exigentes que los demás. En consecuencia, el escéptico es alguien que defiende ideas extravagantes que contradicen las opiniones de la mayoría. Así, algunos escépticos niegan la realidad del cambio climático y otros ponen en duda la autenticidad de las grabaciones del primer alunizaje.

En la filosofía griega, el escepticismo tenía un significado diferente. En griego clásico, 'escéptico' (*skeptikós*) significa 'investigador'. En lugar de adoptar una posición contraria a la de la mayoría, el escéptico adopta una *actitud inquisitiva* de búsqueda de la verdad. Para ello, el escéptico emplea una serie de técnicas conocidas como 'modos', las cuales le permiten examinar distintas opiniones. Al emplear dichas técnicas, el escéptico se percata de que carece

de buenas razones para creer distintas proposiciones acerca de la naturaleza de las cosas. Por tanto, su investigación le conduce a un estado de indecisión, también conocido como 'suspensión del juicio' (*epoché*). Según nos informa Sexto Empírico en sus *Esbozos Pirrónicos*, dicho estado induce en el escéptico la tranquilidad del alma (*ataraxía*).

Los argumentos escépticos desempeñaron un papel importante en Europa durante la Edad Moderna, cuando la Reforma dio lugar a desacuerdos religiosos generalizados y el desarrollo de la nueva ciencia física produjo incertidumbre sobre los poderes cognitivos de la mente humana (Popkin 2003). Durante ese período, algunos filósofos radicalizaron formas previas de duda. Por ejemplo, Descartes (1641) insistió en que no hay indicios seguros de que no estemos soñando y sugirió que un Dios todopoderoso podría engañarlo sobre afirmaciones aritméticas aparentemente simples. Hume (1739–1740) extendió el escepticismo a cualquier cuestión de hecho, incluida la afirmación inductiva aparentemente obvia de que el sol sale todas las mañanas. Muchos filósofos de ese período pensaban que el escepticismo era una fase del pensamiento que debía superarse antes de poder sentar las bases firmes para las ciencias (Descartes 1641; Kant 1781/1787).

Aunque la filosofía analítica se ha inspirado del escepticismo antiguo y moderno, también le ha dado un significado diferente. En lugar de caracterizar el escepticismo como una posición, una actitud, o una fase del pensamiento, los filósofos analíticos han entendido el escepticismo como una serie de paradojas que arrojan conclusiones absurdas tales como 'No tenemos ningún conocimiento' o 'No tenemos conocimientos acerca de entidades externas a nuestras mentes'. Dado que tales paradojas son el resultado de una reflexión filosófica sobre las normas que gobiernan implícitamente nuestras prácticas epistémicas ordinarias, el estudio de las paradojas escépticas ofrece una metodología idónea para comprender el conocimiento y otros estatus epistémicos relacionados. Por tanto, cuando un filósofo analítico discute cierto argumento escéptico, es poco probable que esté defendiendo una posición ex-

céntrica, recomendando una forma de vida o intentando sentar las bases firmes para las ciencias. Es más probable que esté sugiriendo que cierta teoría del conocimiento u otro estatus epistémico relacionado tiene un defecto importante que merece nuestra atención. Este capítulo ofrece una introducción a algunos de los debates más importantes en torno a las paradojas escépticas.

2. Tesis escépticas

Una *tesis escéptica* es una proposición que contradice una o varias proposiciones que gozan de amplia aceptación en una comunidad. Cuando la comunidad en cuestión incluye a la mayoría de los humanos adultos, se dice que la tesis escéptica contradice al *sentido común* (Moore 1925, p.116). En filosofía, las tesis escépticas más discutidas contradicen proposiciones relacionadas con nuestra posesión de distintos tipos de *conocimientos* o *justificación* para tener ciertas creencias. Comencemos con nuestros conocimientos. Una tesis escéptica *universal* es aquella que afirma que no tenemos ningún conocimiento. Además, existen tesis escépticas *radicales* que contradicen nuestra posesión de conocimientos acerca de grandes *clases* de proposiciones. Pese a no ser universales, tales tesis escépticas tienen un alcance suficientemente amplio para despertar la curiosidad filosófica. A modo de ilustración:

- ESCEPTICISMO SOBRE EL MUNDO EXTERNO[2]. No tenemos conocimientos acerca de entidades externas a nuestras mentes. Por ejemplo, no sabemos que hay montañas, sillas, rocas, edificios, sombras y otros seres humanos.
- ESCEPTICISMO SOBRE LAS OTRAS MENTES. No tenemos conocimientos acerca de los estados mentales de otros sujetos. Por ejemplo, no sabemos que otras personas tienen deseos, emociones o creencias.
- ESCEPTICISMO MORAL. No tenemos conocimientos acerca de los valores y las normas morales. Por ejemplo,

2 En adelante, utilizaré mayúsculas para nombrar tesis y conceptos.

> no sabemos que está mal torturar a otros seres humanos, ni que tenemos ciertas obligaciones y derechos.

Algunas formas de escepticismo son más abarcadoras que otras. Supongamos que los conocimientos acerca de los estados mentales de otros sujetos descansan sobre conocimientos acerca de entidades externas a nuestras mentes, como emisiones lingüísticas, gestos o acciones (Carnap, 1928). Dada esta concepción, el escepticismo sobre el mundo externo implica el escepticismo sobre las otras mentes. En cambio, si entidades como las montañas o las sillas no son mentes, el escepticismo sobre las otras mentes no implica el escepticismo sobre el mundo externo.

La clasificación precedente se enfoca en formas de escepticismo acerca del *conocimiento proposicional*, es decir, el tipo de conocimiento que atribuimos mediante oraciones de la forma '*S* sabe que *p*', donde '*S*' denota un sujeto y '*p*' es una variable proposicional[3]. Suele admitirse que dicho conocimiento debe satisfacer una serie de condiciones. Aunque no hay un consenso pleno al respecto, muchos filósofos piensan que el conocimiento debe satisfacer al menos dos condiciones. Si *S* sabe que *p*, entonces:

> CONDICIÓN DE VERDAD. Es verdad que *p*.
> CONDICIÓN DE JUSTIFICACIÓN. *S* tiene justificación para creer que *p*.

Estas dos condiciones nos permiten introducir otras dos tesis escépticas:

- ESCEPTICISMO SOBRE LA VERDAD. Las proposiciones de cierta clase son falsas.
- ESCEPTICISMO SOBRE LA JUSTIFICACIÓN. Carecemos de justificación para creer proposiciones de cierta clase.

3 En adelante, siempre que hablemos de conocimiento nos referiremos al conocimiento proposicional.

Una concepción influyente sostiene que una proposición es verdadera cuando 'corresponde' a los hechos. Y es falsa cuando 'no corresponde' a los hechos. La concepción de la verdad como correspondencia implica que una proposición puede ser falsa, incluso si todas las personas la creen. Por ejemplo, en la época de Ptolomeo, era verdad que la tierra gira alrededor del sol, pese a que la mayoría de la gente creyera que el sol gira alrededor de la tierra. El escepticismo sobre la verdad suele radicalizar situaciones como ésta; señala que cierta clase de proposiciones no corresponden a los hechos, pese a que el sentido común las tenga por verdaderas.

Cuando se afirma que el conocimiento requiere justificación, la palabra 'justificación' no denota una *actividad*, sino un *estado* (Alston 1991; Pryor 2014). Mientras que las acciones tienen lugar en el tiempo, los estados persisten a través de lapsos de tiempo. Luego, uno puede estar en un estado sin efectuar ninguna acción. Cuando un sujeto, *S*, tiene justificación para creer que *p*, *S* está en un estado en el cual es 'apropiado' o 'le está permitido' creer que *p* desde un punto de vista epistémico. La relativización a un 'punto de vista epistémico' indica un tipo de evaluación que apunta al cumplimiento de la CONDICIÓN DE VERDAD. Eso no significa que sólo las proposiciones verdaderas puedan estar justificadas; nos dice que quien tiene justificación para creer que *p* está en un estado conducente a la verdad de *p*. Un sujeto puede tener justificación en ese sentido incluso si no ha formado la creencia de que *p* o ha formado dicha creencia de manera incorrecta. Cuando un sujeto, *S*, tiene justificación para creer que *p*, *S* reúne todos los requisitos necesarios desde el punto de vista de la verdad para creer que *p* (Firth 1978; Goldman 1979). Una tesis escéptica sobre la justificación afirma que un sujeto no puede reunir los requisitos necesarios para que sea apropiado o le esté permitido creer cierta clase de proposiciones.

Estos tres tipos de escepticismo—sobre el conocimiento, la verdad y la justificación—mantienen relaciones interesantes entre sí. Para ilustrarlas, nos enfocaremos en el escepticismo sobre el mundo externo.

Una tesis escéptica sobre la verdad afirma que no existen entidades externas a nuestras mentes. Dado su enfoque en una cuestión de existencia, algunos filósofos piensan que este tipo de escepticismo le compete a la metafísica (Chalmers 2005; Wright 2004). No obstante, sería un error separar este tipo de escepticismo de la epistemología. Si no existen entidades externas a nuestras mentes, no podemos tener conocimientos sobre el mundo externo, pues la verdad es necesaria para el conocimiento. Hay un sentido en el cual el escepticismo sobre la verdad es 'más fuerte' que el escepticismo sobre el conocimiento. Después de todo, la falta de conocimientos sobre el mundo externo es compatible con que las proposiciones sobre el mundo externo sean verdaderas. Por el contrario, la falsedad de las proposiciones sobre el mundo externo es incompatible con la posesión de conocimientos sobre el mundo externo. Sin embargo, defender la afirmación según la cual muchas proposiciones sobre el mundo externo son verdaderas no es suficiente para escapar del escepticismo acerca del conocimiento sobre el mundo externo. Después de todo, la verdad no es suficiente para el conocimiento.

Hay un sentido en el cual el escepticismo sobre la justificación es 'más fuerte' que el escepticismo sobre el conocimiento. Al fin y al cabo, la falta de conocimientos sobre el mundo externo es compatible con que uno tenga (algo de) justificación para creer proposiciones sobre el mundo externo. Por esta razón, muchos autores creen que el escepticismo sobre la justificación es más 'profundo' que el escepticismo sobre el conocimiento (Cohen 1999; Conee y Feldman 2004; Greco 2000, 2007; Williams 2008; Wright 1991). No obstante, defender la afirmación según la cual tenemos (algo de) justificación para creer proposiciones sobre el mundo externo no es suficiente para concluir que tenemos conocimientos sobre el mundo externo. Esto se explica porque la posesión de (algo de) justificación para creer una proposición es compatible con el incumplimiento de una o varias condiciones para el conocimiento.

3. Respuestas anti-escépticas

Una *respuesta anti-escéptica* es un conjunto de consideraciones que buscan defender la legitimidad de las creencias cuestionadas por determinada tesis escéptica. Pese a que existen múltiples respuestas anti-escépticas, podemos reconocer dos grandes categorías: las *respuestas consecuentes* y las *respuestas antecedentes*.

Las respuestas consecuentes critican directamente las tesis escépticas sin tomar en cuenta los argumentos que conducen a tales tesis escépticas. Al criticar directamente las tesis escépticas, resulta natural pensar que éstas caracterizan la posición que adopta alguien más: un escéptico (real o hipotético) que está en desacuerdo con el sentido común. Luego, la tarea filosófica central es ofrecer razones para rechazar la posición escéptica sin presuponer la verdad de ninguna proposición cuestionada por la tesis escéptica. De lo contrario, habremos cometido una petición de principio en contra del escéptico. Muchas respuestas consecuentes buscan alcanzar su cometido argumentando que no es posible adoptar la posición escéptica sin incurrir ciertos costos. Por ello, quienes desarrollan respuestas consecuentes suelen señalar que las posiciones escépticas son incoherentes o que tienen implicaciones prácticas indeseables.

Las respuestas antecedentes se enfocan en examinar los argumentos que conducen a las tesis escépticas. Hay al menos dos razones para preferir las respuestas antecedentes a las respuestas consecuentes. Por una parte, la historia de la filosofía sugiere que las respuestas consecuentes tienen un alcance muy limitado (Strawson 1985; Stroud 1968). Por ello, en vez de persistir en dicho proyecto, muchos filósofos ulteriores han adoptado estrategias 'defensivas'. En lugar de probar que distintas tesis escépticas son falsas, han atacado las razones que podrían esgrimirse a favor de las tesis escépticas (Bergmann 2021; Byrne 2004; Pryor 2000). Por otra parte, enfocarse en los argumentos subyacentes a las tesis escépticas nos permitirá alcanzar una mejor comprensión del origen de tales tesis escépticas. Si tales argumentos invocan conceptos como CONOCIMIENTO o JUSTIFICACIÓN, el examen de tales argumentos nos permitirá entender mejor el funcionamiento de tales

conceptos, incluso si suponemos de entrada que las tesis escépticas son falsas.

Esta última idea ha motivado la formulación del escepticismo como un conjunto de paradojas. En lo que resta de este capítulo, introduciremos esta concepción del escepticismo, expondremos una familia de paradojas escépticas y discutiremos algunos intentos de solución.

4. El escepticismo como un conjunto de paradojas

Supongamos que dos personas debaten sobre una tesis escéptica. El escéptico da razones para defender la tesis según la cual no tenemos conocimientos sobre el mundo externo. Por su parte, el dogmático intenta mostrar que sí tenemos tales conocimientos. Desafortunadamente, después de invertir mucha energía, ninguno ha logrado convencer a su interlocutor; el escéptico y el dogmático se encuentran en un 'empate técnico'. Al constatar el tiempo perdido, parece que lo único que les queda es encogerse de hombros y lamentarse por su falta de ingenio argumentativo. Este resultado desalentador parece señalar el fin del interés del escepticismo para la filosofía.

Este diagnóstico descansa sobre un supuesto cuestionable: que el escepticismo sólo puede entenderse como una posición que adopta un escéptico (real o hipotético) que desafía al partidario del sentido común. Sin embargo, existe una manera diferente de entender el escepticismo. Según esta concepción, una tesis escéptica es una conclusión de un argumento cuyas premisas apelan a *nuestros* propios compromisos de sentido común. Por tanto, el germen del escepticismo ya está en nosotros. Luego, el estudio del escepticismo no es la crónica de un debate infructuoso con alguien más, sino un intento por comprender *nuestros* propios compromisos de sentido común. La manera canónica de desarrollar esta visión consiste en construir el escepticismo como una serie de *paradojas* que emergen cuando reflexionamos sobre nuestros conceptos epistémicos ordinarios.

Una analogía permitirá comprender mejor este punto. La creencia en la existencia del movimiento es parte del sentido co-

mún. Sin embargo, el filósofo griego Zenón de Elea ofreció argumentos para probar que el movimiento no existe. Aunque pocos se tomarían en serio dicha conclusión, los argumentos de Zenón deben su interés teórico a que sus premisas están respaldadas por una reflexión previa sobre nuestros conceptos ESPACIO y MOVIMIENTO. Para ilustrar, la célebre paradoja de la dicotomía nos invita a suponer que Homero desea recorrer un camino. Parece plausible que, antes de llegar a la meta, Homero debe primero recorrer la mitad del camino. Sin embargo, también parece plausible que, antes de llegar a la mitad del camino, Homero debe primero recorrer la mitad de la mitad del camino. Ahora bien, también parece plausible que, antes de llegar a la mitad de la mitad del camino, Homero debe primero recorrer la mitad de la mitad de la mitad del camino. Y así sucesivamente. El argumento de la dicotomía lleva a la conclusión sorprendente de que Homero *nunca* puede recorrer un camino. Como el argumento se puede generalizar a otros casos, el movimiento no existe. Aunque pocos se tomarían en serio esta conclusión, ello no implica que el argumento de la dicotomía carezca de interés filosófico. Como muchos otros problemas teóricos, el argumento de Zenón nos presenta una dificultad para la creencia ordinaria de que existe el movimiento. Dado que la paradoja surge de la reflexión sobre nuestros conceptos ESPACIO y MOVIMIENTO, sería inadecuado ignorarla como si se tratase de una posición excéntrica. Un tratamiento filosófico de este problema exige explicar cómo es posible que exista el movimiento, dado que parece haber obstáculos conceptuales para que Homero pueda recorrer un camino (Stroud 1984, p. 5; 1991, pp. 114–115).

Al igual que ocurre con el argumento de la dicotomía, hay argumentos que apelan a premisas aparentemente plausibles que, tomadas en conjunto, conducen a conclusiones escépticas. Algunas de esas premisas apelan a condiciones que supuestamente deben cumplir ciertos estatus epistémicos, como el conocimiento o la posesión de justificación. Las paradojas surgen porque es difícil ver cómo un sujeto humano normal podría cumplir tales condiciones. Luego, el examen de las paradojas escépticas nos ofrece una vía

expedita para alcanzar una mejor comprensión del funcionamiento de ciertos estatus epistémicos. Es por eso por lo que muchas teorías epistemológicas pueden evaluarse por su capacidad para resolver paradojas escépticas. En la próxima sección, examinaremos un problema que enfrenta la formulación de las paradojas acerca del conocimiento del mundo externo[4].

5. El problema de la aplicación general

Si bien es cierto que algunas tesis escépticas tienen mayor alcance que otras, las tesis escépticas más interesantes buscan tener aplicación general. Así, aunque el escepticismo sobre el mundo externo no sea estrictamente universal, su interés teórico se debe a que pone en duda todos o muchos de nuestros conocimientos sobre el mundo externo. Esto plantea el problema de la aplicación general: ¿Cómo puede un argumento escéptico poner en duda todos o muchos de nuestros conocimientos sobre el mundo externo?

En sus *Meditaciones Metafísicas*, Descartes plantea una serie de dudas acerca de los sentidos con el fin de sentar bases firmes para las ciencias. Su primera duda toma como punto de partida la observación según la cual *algunas veces* cometemos errores perceptuales. Al mirar una superficie, uno podría pensar que está vacía, cuando en realidad está ocupada por millones de microorganismos. Al ver a un caminante acercarse a lo lejos, uno podría pensar que allá viene un amigo, cuando en realidad se trata de un desconocido. Estas observaciones le sugieren a Descartes que es posible que los sentidos lo engañen *siempre*: "[E]s prudente no fiarse por entero de quienes nos han engañado una vez" (Descartes 1641, p. 16).

4 En la tradición analítica, Barry Stroud es el primer defensor de una formulación del escepticismo como una serie de paradojas. Un antecedente importante se encuentra en Clarke (1972). Hoy en día, la mayoría de los filósofos analíticos entienden el escepticismo como una paradoja. Véase Avnur (2012), Bergmann (2021), Byrne (2004, 2014), Greco (2000, 2007), Pryor (2000), Schiffer (1996), Vogel (2004), Wright (1985, 1991) y Pritchard (2012, 2016), entre otros. La formulación de los problemas filosóficos en términos de 'obstáculos conceptuales' se encuentra en Cassam (2007).

Este razonamiento no carece de méritos. Si alguna vez le prestaste dinero a un amigo y éste nunca te lo pagó, quizá sería prudente no volver a prestarle dinero a nadie más. No obstante, el razonamiento anterior tiene dos defectos importantes. Primero, las premisas no respaldan la conclusión. Aunque los sentidos nos engañen *algunas veces*, de allí no se sigue que los sentidos puedan engañarnos siempre. De hecho, los casos de errores perceptuales antes mencionados parecen tener *peculiaridades* que no se aplican a todos los casos de percepción. Cometemos errores perceptuales cuando estamos ante cosas muy pequeñas o ante cosas muy alejadas. Sin embargo, es mucho más difícil equivocarnos cuando vemos cosas suficientemente grandes o cosas que se encuentran a una distancia apropiada. Estas observaciones nos llevan a identificar un segundo defecto en el razonamiento anterior. Es fácil identificar las situaciones en las cuales cometemos errores perceptuales, razón por la cual no es muy plausible asumir que los sentidos puedan engañarnos siempre (Austin, 1962). Si queremos desarrollar argumentos escépticos de aplicación general, no podremos enfocarnos en errores perceptuales ordinarios.

Una estrategia más adecuada para formular argumentos escépticos de aplicación general consiste en centrarnos en casos *representativos* de conocimiento perceptual (Cavell 1979, pp. 52– 53; Conant 2004, pp. 107–108). Un geómetra puede probar que los ángulos de *todos* los triángulos suman 180 grados si se enfoca en las propiedades que todos los triángulos tienen en común y no en las características peculiares de determinado triángulo dibujado en un pizarrón. De manera análoga, un argumento escéptico puede alcanzar una conclusión escéptica general si se enfoca en aquellos casos de conocimiento del mundo externo mediante los sentidos que no se diferencian de manera relevante de los demás casos. Como los casos que impiden alcanzar una conclusión general son aquellos donde las condiciones no son aptas para adquirir conocimientos (porque estamos ante entidades muy pequeñas o entidades muy alejadas), los casos representativos de conocimiento del mundo externo mediante los sentidos son aquellos casos donde no podría-

mos citar tales condiciones para explicar por qué podríamos estar equivocados. Así, un caso representativo de conocimiento perceptual es una situación que parece *ideal* para alcanzar conocimientos sobre el mundo externo mediante los sentidos. Si uno pudiese estar equivocado en un caso *ideal* para alcanzar conocimientos sobre el mundo externo mediante los sentidos, será mucho más razonable pensar que también puede estar equivocado en *todos* los demás casos (Stroud 1991, pp. 21–22).

El mismo Descartes sigue esta metodología. Adoptando la perspectiva de la primera persona, Descartes considera las siguientes creencias perceptuales: "[E]stoy aquí, sentado junto al fuego, con una bata puesta y este papel en mis manos". También piensa que "estas manos y este cuerpo son míos" (1641, p.16). Esta vez, la situación elegida parece ideal para alcanzar conocimientos sobre el mundo externo mediante los sentidos, pues "no parece haber ninguna razón para negar que existan estas manos y este cuerpo mío" (p. 16). De plantear una duda semejante, su actitud sería análoga a la de ciertos desquiciados (pp. 16–17).

Una vez adoptado este punto de partida, Descartes introduce una posibilidad de error mucho más radical que los errores perceptuales antes mencionados: quizá podría estar soñando que está ahí, sentado junto al fuego, con una bata puesta y un papel en sus manos. Por supuesto, a Descartes *le parece* que no está soñando en ese momento. No obstante, piensa que podría tener ese mismo parecer dentro de un sueño. Esta observación da lugar a lo que denominaremos una 'hipótesis escéptica extraordinaria'[5].

Utilizaremos la expresión 'hipótesis escéptica' (*he*) para referirnos a una descripción de un escenario en el cual 1) un sujeto, *S*, no sabe que *p*, y 2) *S* no puede detectar que está en ese escenario escéptico. Luego, si una *he* fuese verdadera, *S* no sabría que *p* y *S* no podría detectar que está en el escenario escéptico descrito por *he*. Algunos errores perceptuales ordinarios pueden describirse como

5 Mi concepción de las hipótesis escépticas extraordinarias se inspira de algunas observaciones de Maddy (2017) sobre el argumento cartesiano del sueño.

hipótesis escépticas. Supongamos que mi colega Ricardo tiene un hermano gemelo idéntico. Supongamos además que veo a Ricardo a una buena distancia y formo la creencia <Allá va Ricardo>[6]. Una hipótesis escéptica diría que, en ese caso, estoy viendo al hermano gemelo de Ricardo y que, dado que Ricardo y su hermano gemelo son idénticos, no puedo detectar que estoy viendo al hermano gemelo de Ricardo.

Las hipótesis escépticas extraordinarias radicalizan los dos rasgos característicos de las hipótesis escépticas. Primero, describen escenarios que se aplican, no a una única proposición, sino a una clase de proposiciones. Segundo, suelen reforzar la incapacidad del sujeto para detectar que está en un escenario escéptico. Mientras que yo podría utilizar métodos diferentes para detectar si estoy viendo al hermano gemelo de Ricardo (como plantearle ciertas preguntas o acercarme para observarlo con detenimiento), las hipótesis escépticas extraordinarias establecen una imposibilidad de principio para detectar, mediante el empleo de otros métodos, que uno está en un escenario escéptico. Así, un sueño extraordinario sería un sueño que involucra experiencias indiscriminables introspectivamente de las experiencias que tenemos en la vigilia y del cual nunca podemos despertarnos. Esto lo diferencia de los sueños ordinarios, los cuales suelen ser fragmentarios e incoherentes y se interrumpen al despertar.

En adelante, cuando hablemos de 'hipótesis escépticas', nos referiremos a hipótesis escépticas extraordinarias. Utilizaremos la expresión 'caso bueno' para denotar un caso que permite adquirir conocimientos sobre el mundo externo mediante los sentidos y 'caso malo' para denotar un escenario descrito por una hipótesis escéptica[7].

¿A qué deben su fuerza las hipótesis escépticas? Todas ellas apelan al siguiente principio:

6 Utilizaré paréntesis angulares para referir a proposiciones.

7 La distinción entre 'casos buenos' y 'casos malos' proviene de Williamson (2000).

> PARIDAD. Al menos en principio, todo lo que *parece* ocurrir en un caso bueno también puede *parecer* ocurrir en un caso malo[8].

PARIDAD implica que, desde la perspectiva del sujeto, estar en un caso bueno es compatible con estar en un caso malo. Podemos precisar esta idea si recordamos que saber que *p* exige cumplir ciertas condiciones. Luego, existen tantas formas en las que PARIDAD puede amenazar la posesión de conocimientos como hay condiciones del conocimiento que pueden no cumplirse mientras que al sujeto le parece que sí se cumplen.

Consideremos la CONDICIÓN DE VERDAD. Si un sujeto, *S*, sabe que *p*, entonces es verdad que *p* (Sección 2). Las hipótesis escépticas más comunes describen escenarios en los cuales un gran número de creencias del sujeto son falsas y el sujeto no puede detectar que son falsas. Dada su conexión con la verdad, podemos bautizar a estas hipótesis escépticas 'hipótesis aléticas' (del griego *alétheia*, verdad). La hipótesis cartesiana del genio maligno es una hipótesis alética, pues dice que ese ser superior pudo haber procedido de tal modo que todas las creencias del meditador cartesiano son falsas. Un ejemplo más reciente nos lo ofrece la hipótesis del cerebro en una cubeta (CC) (Putnam, 1981). Una variación de esta hipótesis señala que unos científicos malvados capturaron a Ricardo cuando éste se encontraba durmiendo, lo drogaron y le quitaron su cerebro, el cual mantuvieron vivo en una cubeta utilizando tecnología avanzada. A continuación, conectaron el cerebro de Ricardo a una computadora que reproduce fielmente el patrón de activación neuronal que se produciría si Ricardo estuviese percibiendo entidades externas a su mente. En consecuencia, Ricardo ahora tiene alucinaciones no-verídicas que no puede discriminar

8 Los filósofos griegos pensaban que había criterios para distinguir las experiencias oníricas y las experiencias de la vigilia. Luego, ninguno de ellos habría aceptado PARIDAD. Por el contrario, algunas de las hipótesis escépticas cartesianas presuponen PARIDAD. Para análisis iluminadores del carácter único de las hipótesis escépticas cartesianas en la historia de la filosofía, véase Burnyeat (1982) y Williams (2010).

introspectivamente de las experiencias verídicas que tendría si estuviese percibiendo entidades externas a su mente.

Descartes y muchos otros filósofos entendieron el sueño como una hipótesis alética. Después de todo, muchos sueños tienen contenidos no-verídicos. Sin embargo, algunos sueños tienen contenidos verídicos. Por ejemplo, Descartes podría haberse quedado dormido mientras estaba sentado junto al fuego, con una bata puesta y un papel en sus manos y soñar, verídicamente, que estaba sentado junto al fuego, con una bata puesta y un papel en sus manos (Moore, 1944, p. 189). Pese a tener un contenido verídico, el sueño de Descartes es incompatible con que Descartes sepa que está sentado junto al fuego, con una bata puesta y un papel en sus manos. En estas circunstancias, parece que no se cumple una tercera condición para el conocimiento que no hemos mencionado todavía. Si *S* sabe que *p*, entonces debe cumplirse una CONDICIÓN ANTI-SUERTE:

> CONDICIÓN ANTI-SUERTE. La creencia de *S* de que *p* no es verdadera por pura suerte.

Si un sujeto forma una creencia verdadera con base en un sueño verídico, su creencia es verdadera por pura suerte (Stroud, 1991, p. 25). Después de todo, es muy fácil que ese mismo sueño dé lugar a creencias no-verídicas. Luego, los sueños verídicos también cuentan como hipótesis escépticas. Llamaremos a las hipótesis escépticas que violan condiciones para el conocimiento diferentes de la CONDICIÓN DE VERDAD 'hipótesis no-aléticas'. En adelante, sólo discutiremos formas de escepticismo que apelan a hipótesis aléticas.

Para alcanzar una conclusión escéptica de aplicación general, debemos enfocarnos en casos representativos y formular hipótesis escépticas extraordinarias. Sin embargo, esta estrategia no garantiza que alcanzaremos una conclusión escéptica de aplicación general. Recordemos que el escepticismo sobre el mundo externo señala que *no tenemos* conocimientos sobre el mundo externo. Las formas de escepticismo que apelan a hipótesis aléticas buscan de-

fender dicha tesis al recrear escenarios en los cuales *todas* nuestras creencias sobre el mundo externo son falsas. No obstante, la hipótesis del CC no permite recrear un escenario semejante. Incluso si la hipótesis del CC fuese verdadera, ésta sería compatible con que Ricardo hubiese adquirido conocimientos sobre el mundo externo *antes* de que su cerebro fuese puesto en una cubeta. Además, dicho escenario es inteligible sólo si suponemos que *hay* un ser humano, Ricardo, que vivió una vida más o menos normal hasta que fue capturado. En consecuencia, dicho escenario sólo puede ser actual si es verdad que *hay* científicos malvados que capturaron a Ricardo cuando éste se encontraba durmiendo, lo drogaron y le quitaron su cerebro. Y *hay* al menos un cerebro, drogas, una cubeta y una computadora. Todas estas proposiciones son acerca de entidades externas a nuestras mentes. Finalmente, los científicos malvados supuestamente sabían muchas cosas acerca de Ricardo, las drogas, el funcionamiento de los cerebros y la relación de los cerebros con el resto del mundo externo. En suma, para generar este tipo de hipótesis, parece necesario que muchas proposiciones sobre el mundo externo sean verdaderas y que algunos sujetos tengan conocimientos sobre el mundo externo (Clarke, 1972; Echeverri, 2017; Levin, 2000; Schönbaumsfeld, 2016)[9].

Una solución a este problema es conformarnos con formular hipótesis escépticas que respalden conclusiones escépticas menos generales. Así, podemos admitir que la hipótesis del CC no recrea un escenario donde *no tenemos ningún* conocimiento sobre el mundo externo. Sin embargo, sí es una hipótesis que cuestiona muchos de los conocimientos sobre el mundo externo que Ricardo cree poseer. Si la hipótesis del CC fuese verdadera, Ricardo no podría

9 El mismo Descartes (1641, p. 17) señala que los sueños no-verídicos presuponen la existencia de 'cosas más generales' o 'más simples' como ojos, cabeza, manos y cuerpo entero, al igual que ideas más generales como la idea de naturaleza corpórea en general y su extensión, el concepto de figura de las cosas extensas, los conceptos de cantidad y de magnitud y los conceptos de número, lugar y tiempo. Esto lo lleva a introducir la hipótesis escéptica más radical de un ser superior engañador. Si ese ser interactúa con sujetos ubicados en el espacio, tal hipótesis enfrenta dificultades análogas a las que enfrenta la hipótesis del CC.

adquirir conocimientos acerca del mundo externo después de que pusieron su cerebro en una cubeta y lo conectaron a una computadora (Bergmann, 2021; Pritchard, 2005a, 2016). Aunque esta hipótesis está centrada en Ricardo y no en todos los seres humanos, no es difícil formular hipótesis análogas para cualquier otro ser humano. Luego, dichas hipótesis pueden amenazar una gran parte de los conocimientos sobre el mundo externo de cualquier ser humano. Hipótesis escépticas de este tipo pueden atacar todos los conocimientos que cualquier ser humano normalmente obtendría mediante los sentidos después de un momento dado.

6. Estructura de las paradojas escépticas

Una hipótesis escéptica describe un escenario en el cual no se cumple una condición necesaria para tener cierto tipo de conocimientos y el sujeto no puede detectar que se encuentra en dicho escenario. Formular una hipótesis escéptica no basta, empero, para concluir que un sujeto carece de ese tipo de conocimientos. Si el sujeto no se encuentra en el escenario escéptico, todavía no tenemos razones para pensar que carece de ese tipo de conocimientos. Para alcanzar una conclusión escéptica se necesita algo más: mostrar que existe una tensión entre ciertas hipótesis escépticas y la posesión de cierto tipo de conocimientos. La manera canónica de introducir dicha tensión es sostener que saber que *p* exige *poder descartar* hipótesis contrarias a la posesión de conocimientos. Si nos enfocamos en la CONDICIÓN DE VERDAD, las hipótesis en cuestión serán hipótesis aléticas contrarias[10]. En adelante, nos referiremos a tales hipótesis como 'alternativas'. Esto nos lleva a formular un principio básico para muchos argumentos escépticos:

10 Dos proposiciones *p* y *q* son contrarias si y sólo si la verdad de *p* implica la falsedad de *q* o viceversa. Por ejemplo, la proposición <Sara sólo tiene dos hermanos> es contraria a la proposición <Sara tiene tres hermanos>. Podría ocurrir, empero, que ambas proposiciones son falsas.

> PEA: PRINCIPIO DE ELIMINACIÓN DE ALTERNATIVAS. Si un sujeto, *S*, sabe que *p*, y además *S* sabe que *p* es incompatible con *q*, entonces *S* puede descartar que *q*.

Como veremos más adelante, hay distintas maneras de entender la exigencia de que alguien pueda 'descartar' una proposición. Por lo pronto, procederemos con una comprensión intuitiva. PEA nos dice que, si un sujeto, *S*, sabe que *p*, entonces *S* puede descartar todas las alternativas a *p* conocidas por *S*. Algunos ejemplos de la vida diaria parecen validar este principio. Supongamos que Ricardo ve una cebra en un corral y forma la creencia de que hay una cebra en el corral. Supongamos además que Ricardo sabe que la proposición <Hay una cebra en el corral> es incompatible con la proposición <Hay un cocodrilo en el corral> (pues sólo hay un animal en el corral). Para que la creencia de Ricardo de que hay una cebra en el corral sea un conocimiento, parece necesario que Ricardo pueda descartar la proposición incompatible <Hay un cocodrilo en el corral>. De lo contrario, no diríamos que Ricardo *sabe* que hay una cebra en el corral (Stroud, 1991, p. 32).

Supongamos ahora que Ricardo sabe que <Hay una cebra en el corral> es incompatible con <Soy un CC>. Dado PEA, para que Ricardo sepa que hay una cebra en el corral, debe poder descartar la hipótesis del CC. Por desgracia, las hipótesis escépticas están diseñadas para que el sujeto no pueda detectar que se encuentra en los escenarios escépticos descritos por tales hipótesis. De allí parece seguirse que Ricardo no puede descartar la hipótesis del CC. Por tanto, Ricardo no sabe que hay una cebra en el corral.

En suma, la conclusión escéptica se impone si aceptamos que tener ciertos conocimientos exige el cumplimiento de la condición impuesta por PEA y además aceptamos que dicha condición no se puede cumplir en ciertos casos.

Con estos elementos a disposición, podemos construir el argumento escéptico como una *reducción al absurdo*:

Premisa 1. Ricardo sabe que hay una cebra en el corral [proposición de sentido común].
Premisa 2. Si Ricardo sabe que hay una cebra en el corral, y además Ricardo sabe que la proposición <Hay una cebra en el corral> es incompatible con la proposición <Soy un CC>, entonces Ricardo sabe que (él mismo) no es un CC [PEA].
Premisa 3. Aunque Ricardo sabe que la proposición <Hay una cebra en el corral> es incompatible con la proposición <Soy un CC>, Ricardo no sabe que (él mismo) no es un CC [razonamiento a partir de la hipótesis escéptica].
Conclusión. Ricardo no sabe que hay una cebra en el corral.

Si nos enfocamos en un caso representativo de conocimiento, la Conclusión se aplica no sólo al conocimiento de Ricardo de que hay una cebra en el corral, sino a cualquier otro supuesto conocimiento cuyo contenido proposicional sea incompatible con que Ricardo sea un CC. Como podemos formular versiones de la hipótesis del CC para cualquier otro sujeto humano, parece que hemos alcanzado una conclusión escéptica de gran generalidad.

Existen otras maneras de presentar el problema escéptico. Si eliminamos la Conclusión, tendremos tres proposiciones que nos parecen individualmente plausibles, pero que, en conjunto, nos llevan a una contradicción. Generalizando a partir del caso anterior, tenemos la siguiente tríada:

PROPOSICIÓN DE SENTIDO COMÚN. *S* sabe que *p*.
PROPOSICIÓN CONECTIVA. Si *S* sabe que *p*, y además *S* sabe que *p* es incompatible con una hipótesis escéptica *he*, entonces *S* puede descartar *he*[11].
PROPOSICIÓN SEMI-ESCÉPTICA. Aunque *S* sabe que *p* es incompatible con una hipótesis escéptica *he*, *S* no puede descartar *he*.

11 Tomo la expresión 'proposición conectiva' de Avnur (2012).

El primer elemento de la paradoja es una proposición que expresa la creencia de sentido común según la cual tenemos cierto tipo de conocimientos. Denominaremos a este primer elemento PROPOSICIÓN DE SENTIDO COMÚN. El segundo elemento es una proposición que establece una conexión entre la PROPOSICIÓN DE SENTIDO COMÚN y las hipótesis escépticas. A dicha proposición la denominaremos PROPOSICIÓN CONECTIVA. Como la PROPOSICIÓN CONECTIVA se deriva de un principio más fundamental, nos referiremos a este último como el PRINCIPIO CONECTIVO. El tercer elemento es una proposición que señala que la condición formulada en la PROPOSICIÓN CONECTIVA no se puede cumplir. Nos referiremos a esta proposición como PROPOSICIÓN SEMI-ESCÉPTICA, ya que expresa una conclusión escéptica limitada basada en un razonamiento previo acerca de las hipótesis escépticas.

Si aceptamos la conjunción de la PROPOSICIÓN CONECTIVA y la PROPOSICIÓN SEMI-ESCÉPTICA, estamos obligados a negar la PROPOSICIÓN DE SENTIDO COMÚN. Tomadas en conjunto, la PROPOSICIÓN CONECTIVA y la PROPOSICIÓN SEMI-ESCÉPTICA constituyen un *obstáculo conceptual* para que podamos tener conocimientos sobre el mundo externo. Por ello, las paradojas escépticas permiten plantear 'preguntas de posibilidad' (*how-possible questions*): ¿Cómo es posible saber que *p*, dado que parece haber obstáculos conceptuales para tener dicho conocimiento?

7. Dos tipos de soluciones

Es posible alcanzar la conclusión escéptica mediante consideraciones aparentemente plausibles. Luego, el escepticismo es un problema interesante incluso si no hay un sujeto (real o hipotético) que adopta una posición escéptica. Podemos pensar en el PRINCIPIO CONECTIVO como un intento por formular una norma implícita en nuestras prácticas epistémicas ordinarias (Stroud, 1991, 36). Si dicha norma articula una supuesta condición para el conocimiento, no es posible resolver la paradoja escéptica si no se adopta una postura acerca de cómo funciona el conocimiento en nuestras prác-

ticas epistémicas ordinarias. Es por eso por lo que muchas teorías epistemológicas contemporáneas pueden entenderse, al menos en parte, como soluciones a paradojas escépticas. Además, si una teoría epistemológica tiene consecuencias escépticas, tendremos una buena razón para rechazarla. En la epistemología contemporánea, formular una paradoja escéptica es análogo a encontrar un 'bug' en un programa informático. La existencia de la paradoja revela fallos en nuestra comprensión del conocimiento.

Dado que las tres proposiciones llevan a una contradicción, no basta con insistir en que la tesis escéptica es absurda. La paradoja escéptica ya nos ha mostrado tal absurdidad al revelar una contradicción en nuestra comprensión de ciertos conceptos epistémicos. Si queremos preservar la PROPOSICIÓN DE SENTIDO COMÚN, estamos obligados a rechazar la PROPOSICIÓN CONECTIVA o la PROPOSICIÓN SEMI-ESCÉPTICA. Idealmente, también deberíamos explicar la plausibilidad inicial de la proposición que decidamos rechazar (DeRose, 1995).

Muchos filósofos contemporáneos piensan que la solución escéptica debería ser nuestro último recurso; es una solución que sólo adoptaríamos después de haber examinado y rechazado las soluciones anti-escépticas disponibles. Por tal razón, examinaremos inicialmente dos tipos de soluciones anti-escépticas.

Un primer tipo de solución consiste en rechazar la PROPOSICIÓN CONECTIVA. Como la PROPOSICIÓN CONECTIVA deriva su plausibilidad del PRINCIPIO CONECTIVO, no parece posible rechazar la PROPOSICIÓN CONECTIVA sin ofrecer razones para pensar que el conocimiento no tiene que cumplir la condición que impone el PRINCIPIO CONECTIVO. Luego, el objetivo será mostrar que la posesión de conocimientos puede coexistir pacíficamente con nuestra incapacidad para descartar las hipótesis escépticas. Una solución satisfactoria debería explicarnos por qué el PRINCIPIO CONECTIVO no refleja el funcionamiento del conocimiento en nuestras prácticas ordinarias. Idealmente, también deberíamos formular un principio alternativo que no tenga consecuencias escépticas.

Un segundo tipo de solución rechaza la PROPOSICIÓN SEMI-ESCÉPTICA. Sus partidarios aceptan que la posesión de conocimientos sobre el mundo externo exige que podamos descartar las hipótesis escépticas, pero argumentan que es posible satisfacer la condición impuesta por el PRINCIPIO CONECTIVO[12].

Los dos tipos de solución responden a motivaciones opuestas. Quienes rechazan el PRINCIPIO CONECTIVO suelen pensar que éste impone condiciones demasiado exigentes. Por su parte, quienes rechazan la PROPOSICIÓN SEMI-ESCÉPTICA suelen pensar que la condición que impone el PRINCIPIO CONECTIVO no es tan exigente como podría parecer. Así, incluso aquellos que rechazan la conclusión escéptica están divididos respecto a qué tan exigentes deben ser las condiciones para el conocimiento. Este punto se aclarará en las próximas dos secciones.

8. La paradoja de cierre

Distintas paradojas escépticas emplean principios conectivos diferentes. Muchos principios conectivos pueden entenderse como maneras diferentes de desarrollar PEA (Sección 6). Recordemos que dicho principio exige que, para saber que *p*, es necesario poder descartar todas aquellas proposiciones que uno sabe que son incompatibles con p. Muchos filósofos han entendido esta exigencia en términos de conocimiento: 'descartar *he*' no es otra cosa que *saber* que *no-he* (o saber que *he* es falsa). Podemos capturar esta idea mediante el siguiente principio de CIERRE:

12 Descartes (1641) intentó rechazar la PROPOSICIÓN SEMI-ESCÉPTICA mediante una prueba de la existencia de un Dios garante de verdad. Muchos filósofos consideran que la estrategia teísta de Descartes está condenada al fracaso, pues comete una petición de principio. Es por eso por lo que el segundo tipo de solución suele asociarse con G. E. Moore, el más destacado defensor del sentido común en la tradición analítica. En sus textos sobre este tema, Moore (1939, 1953) acepta el PRINCIPIO CONECTIVO y la PROPOSICIÓN DE SENTIDO COMÚN.

> CIERRE. Si un sujeto, *S*, sabe que *p*, y además *S* sabe que *p* implica *q*, entonces *S* sabe que *q*[13].

Un principio de cierre identifica ciertas operaciones que, de aplicarse a ciertas entidades, arrojan esas mismas entidades. Por ejemplo, los números naturales están 'cerrados' bajo la operación de la adición, pues, al sumar un número natural a otro número natural, siempre obtendremos otro número natural. Algo diferente ocurre con la resta. El resultado de 2–5 es –3, un número negativo (Kvanvig, 2006). De manera análoga, CIERRE nos dice que, si aplicamos la operación de la implicación conocida al conocimiento, nunca nos saldremos del 'área cerrada' del conocimiento; siempre obtendremos más conocimientos.

Existe un debate acerca de cómo debe formularse CIERRE. La mayoría de las propuestas buscan excluir contraejemplos que trivializarían dicho principio (Alspector-Kelly, 2019; David y Warfield, 2008; Hawthorne, 2014; Pritchard, 2016). Sin embargo, no necesitamos entrar en tales discusiones para entender por qué muchos teóricos aceptan uno u otro principio de CIERRE. Supongamos que yo sé que Sara tiene dos hermanos y que también sé que la proposición <Sara tiene dos hermanos> implica la proposición <Sara tiene más de un hermano>. Luego, parece seguirse que sé que Sara tiene más de un hermano. Supongamos que yo sé que Canadá es parte de Norteamérica y que también sé que, si un territorio es parte de Norteamérica, dicho territorio es parte del planeta Tierra. Luego, parece seguirse que sé que Canadá es parte del planeta Tierra. Ejemplos como éstos parecen indicar que CIERRE (o un principio análogo) es válido.

Si aplicamos CIERRE a una hipótesis escéptica *he*, tenemos una nueva PROPOSICIÓN CONECTIVA:

13 Para formulaciones análogas, véase Cohen (1999), DeRose (1995), Dretske (1970, 2014), Nozick (1981), Sosa (1999) y Vogel (1990).

> PROPOSICIÓN CONECTIVA. Si *S* sabe que *p*, y además *S* sabe que *p* implica *no-he*, entonces *S* sabe que *no-he*.

Recordemos que las hipótesis escépticas están diseñadas para que el sujeto no pueda detectar que se encuentra en los escenarios escépticos descritos por tales hipótesis. Ante la imposibilidad de detectar que uno está en un escenario escéptico descrito por *he*, ¿cómo podría llegar a saber que *no-he*? Dado que no parece posible saber algo semejante, esto nos conduce a una nueva PROPOSICIÓN SEMI-ESCÉPTICA:

> PROPOSICIÓN SEMI-ESCÉPTICA. Aunque *S* sabe que *p* implica *no-he*, *S* no sabe que *no-he*.

Lamentablemente, si aceptamos la conjunción de la PROPOSICIÓN CONECTIVA y la PROPOSICIÓN SEMI-ESCÉPTICA, estamos obligados a rechazar la PROPOSICIÓN DE SENTIDO COMÚN y adoptar la tesis escéptica:

> TESIS ESCÉPTICA. No es el caso que: *S* sabe que *p*.

A continuación, presentaremos algunas respuestas influyentes a la paradoja de cierre.

8.1 Rechazo de cierre

Aunque algunos casos ordinarios parecen validar CIERRE, algunos filósofos han argumentado que dicho principio no es válido de manera irrestricta. En su opinión, existen casos ordinarios en los cuales un sujeto, *S*, sabe que *p*, además *S* sabe que *p* implica *q*, pero *S* no sabe que *q*. Fred Dretske (1970, pp.1015–1016) ofrece un ejemplo clásico. Supongamos que Ricardo está en el jardín zoológico. Al llegar al corral marcado 'Cebra', Ricardo ve un espécimen equino con rayas blancas y negras y forma la creencia en la proposición <Ese animal es una cebra>. Como Ricardo sabe reconocer una cebra a simple vista y no tiene razones para dudar, parece que se cumplen todas las condiciones para que Ricardo sepa que

ese animal es una cebra. Supongamos ahora que Ricardo sabe que la proposición <Ese animal es una cebra> implica la proposición <Ese animal no es una mula hábilmente disfrazada para que parezca una cebra> (pues los dos ejemplares de 'ese' refieren al mismo animal). En este caso, parece menos plausible asumir que Ricardo también sabe que ese animal no es una mula hábilmente disfrazada para que parezca una cebra. Después de todo—señala Dretske—la evidencia visual que Ricardo tenía para pensar que era una cebra "ha sido efectivamente neutralizada"; dicha evidencia visual es compatible con que Ricardo esté ante una mula hábilmente disfrazada para que parezca una cebra. Luego, es posible que uno sepa que ese animal es una cebra, que además sepa que la proposición <Ese animal es una cebra> implica la proposición <Ese animal no es una mula hábilmente disfrazada para que parezca una cebra>, pero que no sepa que ese animal no es una mula hábilmente disfrazada para que parezca una cebra.

El contraejemplo funciona únicamente si suponemos que la evidencia que le permite a Ricardo saber que ese animal es una cebra tiene que ser la misma evidencia que le permitiría saber que ese animal no es una mula hábilmente disfrazada para que parezca una cebra. Sin embargo, muchos autores han cuestionado tal asunción. Tales autores comienzan señalando que CIERRE no especifica que la evidencia que subyace al conocimiento de la proposición inicial tenga que ser la misma evidencia que le permitiría a Ricardo conocer la verdad de la proposición implicada. Lo único que dice CIERRE es que, cuando uno sabe que *p* implica *q*, uno no puede saber que *p* a menos que también sepa que *q*. Y es posible cumplir esa exigencia si la evidencia para saber que *p* es diferente de la evidencia que permitiría saber que *q*. A continuación, tales autores sostienen que los sujetos humanos adultos sí cuentan con evidencia independiente para saber que la hipótesis de la mula disfrazada es falsa. Por inducción, sabemos que es muy poco probable que un engaño semejante no fuese descubierto oportunamente; también sabemos que un engaño semejante daría lugar a una multa y que dicha posibilidad disuadiría a los propietarios del jardín zoológi-

co de efectuar tal engaño (Pritchard, 2012; Vogel, 1990; Williams, 2001). En suma, al tomar en cuenta la existencia de evidencia inductiva, el contraejemplo de Dretske pierde gran parte de su poder persuasivo[14].

Las objeciones anteriores muestran que no es posible rechazar CIERRE sin adoptar compromisos teóricos sustantivos acerca de la naturaleza de la evidencia disponible para un sujeto. Por tal razón, una estrategia más popular apela directamente a ciertos análisis del conocimiento. Tales análisis introducen algunas condiciones necesarias para el conocimiento que son fáciles de satisfacer por nuestras creencias ordinarias en el mundo externo, pero que nuestras creencias en las negaciones de las hipótesis escépticas no pueden satisfacer. De tener éxito, tales análisis mostrarían que es posible conocer la verdad de muchas proposiciones ordinarias sobre el mundo externo sin conocer la falsedad de las hipótesis escépticas.

Una primera estrategia se conoce como la 'teoría de las alternativas relevantes'. Inspirados por ciertas observaciones de Austin (1946), varios autores han señalado que, para saber que *p*, no es necesario conocer la negación de todas las alternativas a *p*. Basta con poder conocer la negación de un subconjunto de tales alternativas, también conocidas como las 'alternativas relevantes'. En el caso de la cebra, Austin diría que podemos saber que ese animal es una cebra con tal de que podamos reconocer sus rasgos característicos en circunstancias favorables. Esto ocurriría si no hay otras especies animales que tengan la misma apariencia visual de las cebras en las circunstancias en las que se encuentra Ricardo. Luego, en circunstancias favorables, la alternativa <Ese animal no es una mula hábilmente disfrazada para que parezca una cebra> no es relevante. Si, por el contrario, el sujeto se encontrase en circunstancias desfavorables donde es frecuente que las mulas sean hábilmente disfrazadas para que parezcan cebras, reconocer que un animal

14 Vogel (1990) plantea otros contraejemplos que evitan esta objeción. Para una discusión influyente de tales casos, véase Hawthorne (2004). Echeverri (2020) señala que la apelación a evidencia inductiva no es exitosa si se formula CIERRE como un principio diacrónico.

tiene rayas blancas y negras no será suficiente para saber que ese animal es una cebra. En tales circunstancias, la alternativa <Ese animal no es una mula hábilmente disfrazada para que parezca una cebra> sí es relevante. Después de todo, allí sí tendríamos buenas razones para pensar que el animal que Ricardo está viendo podría ser una mula hábilmente disfrazada como una cebra.

El mismo razonamiento vale para la hipótesis del CC. Supongamos que Ricardo está en un entorno en el cual las personas son frecuentemente capturadas por científicos malvados, quienes a su vez ponen sus cerebros en cubetas, los conectan a computadoras, y así sucesivamente. En dicho entorno, Ricardo no puede saber que ese animal es una cebra a menos que pueda saber que (él mismo) no es un CC. Sin embargo, en un entorno donde tales episodios no suelen ocurrir, Ricardo puede saber que ese animal es una cebra, aunque no pueda saber que (él mismo) no es un CC. En resumen, estar en circunstancias favorables permite adquirir conocimientos ordinarios sobre el mundo externo pese a que uno no sepa que ciertas hipótesis escépticas son falsas.

Consideremos con mayor atención el razonamiento que respalda la propuesta de Austin. Para Austin (1946), si hemos de identificar las normas que gobiernan implícitamente el concepto de conocimiento, conviene examinar las distintas maneras en que evaluamos el conocimiento en el lenguaje ordinario. Resulta evidente que, al evaluar cotidianamente si alguien sabe o no sabe, nunca le exigimos que conozca las negaciones de hipótesis escépticas. Incluso en las ciencias, donde las normas epistémicas son más estrictas, no exigimos que los científicos expliquen cómo determinaron que no son cerebros en cubetas o que no estaban siendo engañados por un genio maligno. De hecho, si alguien plantease exigencias semejantes, las desecharíamos como poco razonables. Dado que CIERRE parece autorizarnos a plantear exigencias que son poco razonables, CIERRE distorsiona el concepto ordinario de conocimiento[15].

15 Pritchard (2010, 2012) ha intentado reconciliar la teoría de las alternativas relevantes con un principio diacrónico de CIERRE. Para Maddy (2017), Austin defiende la tesis más radical según la cual no hay algo así como *el* concepto CONOCIMIENTO. Lo

Stroud (1991, cap. 2) ha objetado que el argumento anterior es inválido. Aunque en la vida diaria nos parecería poco razonable objetarle a alguien que no sabe que *p* porque no conoce la negación de ciertas hipótesis escépticas, de allí no se sigue que el concepto ordinario de conocimiento no plantee la exigencia de tener conocimientos anti-escépticos. En otras palabras, la evidencia sobre la conducta lingüística de la gente es compatible con la exigencia que plantea CIERRE. Para que el argumento austiniano funcione, es necesario suponer que la conducta lingüística *refleja directamente* nuestra comprensión del concepto ordinario de conocimiento. Sin embargo, existe una visión alternativa: la conducta lingüística bien podría reflejar ciertos aspectos prácticos que no afectan a la verdad de CIERRE, aunque sí determinan qué tipos de exigencias nos parecen razonables desde un punto de vista práctico en la vida diaria. Más específicamente, el *acto de formular una exigencia* puede ser poco razonable, aunque la *exigencia misma* emane de un principio verdadero. A modo de analogía, pensemos en casos en los que sería de mala educación hacer un comentario, aunque ese comentario dijera algo cierto. Por paridad de razonamiento, CIERRE aún podría gobernar tácitamente nuestras prácticas epistémicas, aunque los requisitos de la acción, la cooperación y la comunicación nos lleven a ignorar la exigencia de poder saber que las hipótesis escépticas son falsas:

> Lo que es o no es apropiado o razonable hacer está determinado por la situación que se afronta, por los propósitos o intereses que se tienen en el momento, por la apreciación que hacemos de la situación y, como destaca Descartes, por el tiempo de que se dispone. Sería absurdo quedarse de pie durante un buen rato en un camión que se está llenando rápidamente tratando de decidir cuál es precisamente el mejor lugar para sentarse. Ya que sentarse en algún sitio es mucho mejor que permanecer de pie, aunque ciertamente no sea tan bueno

único que tenemos es una multiplicidad de usos específicos de palabras como 'saber' y 'conocimiento'. Esta lectura la compromete con cierto escepticismo sobre los conceptos (Echeverri 2018).

> como sentarse en el mejor de todos los lugares, lo mejor que puede hacerse es sentarse rápidamente. En general, nuestras acciones pueden tener más éxito y producir una mayor satisfacción en la medida en que podemos reflexionar con más tiempo, de manera más rigurosa y con más información, pero forma parte de la naturaleza misma de la vida práctica el hecho de que a menudo no podamos llevar este proceso lo suficientemente lejos para conseguir la clase de certeza que de otro modo nos gustaría tener. Hacemos lo mejor que podemos de acuerdo con las circunstancias. (Stroud, 1991, p. 61)

Estas observaciones le permiten a Stroud explicar por qué surge la paradoja escéptica. La paradoja surge cuando hacemos explícito un principio que gobierna implícitamente nuestras prácticas epistémicas ordinarias y evaluamos dicho principio haciendo abstracción de la situación específica que se afronta y de los propósitos o intereses que se tienen en el momento. Para Stroud, el argumento escéptico no distorsiona el concepto ordinario CONOCIMIENTO, sino que examina dicho concepto desde un punto de vista objetivo que abstrae de los imperativos que gobiernan "la práctica social y las exigencias de la acción, la cooperación y la comunicación" (Stroud, 1991, p. 65) para determinar si *realmente* sabemos lo que creemos saber. Esto nos permite entender la fuerza que ejerce la paradoja escéptica cuando la consideramos por primera vez. Al examinar el conocimiento de manera objetiva, nos parece que éste debe cumplir una condición que no se puede cumplir. Esta perspectiva objetiva ya está implícita en la obra cartesiana, pues la investigación filosófica que emprende Descartes le exige dejar de lado todas sus preocupaciones prácticas (Stroud, 1991, p. 62)[16].

Pese a sus méritos indudables, el argumento de Stroud enfrenta varias dificultades. En primer lugar, Stroud pasa sutilmente de la tesis plausible según la cual puede haber exigencias verdaderas que parezcan poco razonables en la vida diaria a la tesis menos plausible según la cual una práctica puede estar gobernada por una exigen-

16 Williams (1978) se refiere a ese proyecto cartesiano como un 'proyecto de investigación pura'.

cia verdadera que nunca puede cumplirse. Segundo, Austin podría objetarle a Stroud que todos los conceptos humanos responden a necesidades prácticas específicas. Al separar el concepto de conocimiento de las exigencias de la vida práctica, resulta enigmático por qué empleamos dicho concepto y por qué una práctica humana podría estar estructurada en torno a un concepto que plantea exigencias que no se pueden cumplir. Tercero, en los últimos años, muchos autores han defendido la tesis de la intrusión pragmática (*pragmatic encroachment*). Dicha tesis señala que ciertas diferencias relativas a lo que está en juego desde un punto de vista práctico son relevantes para determinar si un sujeto tiene o no conocimientos (Fantl y McGrath, 2009; Stanley. 2005). Si la tesis de la intrusión pragmática es correcta, no es viable entender el concepto de conocimiento como algo completamente separado de nuestros intereses prácticos.

Muchos filósofos han intentado precisar el concepto de alternativa relevante y las condiciones para saber que una alternativa relevante es falsa (Dretske, 1971, 1981; Goldman, 1976; Pritchard, 2010; Stine, 1976; para una crítica influyente, véase Vogel, 1999). Sin embargo, es quizá Robert Nozick (1981) quien ha planteado el análisis más influyente (véase también Becker, 2007).

La idea básica de Nozick es que, para saber que *p*, es necesario 'seguir el rastro' (*track*) de la verdad de *p* a través de mundos posibles cercanos. Es así como podemos diferenciar una creencia que es verdadera por pura suerte de un caso genuino de conocimiento. Una condición necesaria para que una creencia siga el rastro de la verdad es que se cumpla la condición de SENSIBILIDAD:

> SENSIBILIDAD. La creencia de *S* de que *p* es sensible si y sólo si: Si *p* no fuera verdadera, *S* no creería que *p*[17].

17 La formulación oficial de Nozick (1981, 172–187) está relativizada a métodos. Tal relativización le permite evitar ciertos contraejemplos.

Evaluamos los condicionales contrafácticos de este tipo en relación con un conjunto de mundos posibles que están ordenados por su similitud con el mundo actual (Lewis, 1973; Stalnaker, 1968). De manera más específica, identificamos los mundos más cercanos en los cuales el antecedente es verdadero y nos preguntamos si, en esos mundos, el consecuente también es verdadero. Según la interpretación de Nozick, SENSIBILIDAD exige que, en los mundos más cercanos en los cuales *p* no es verdadera, el sujeto no crea que *p*[18]. Esta teoría predice que, en circunstancias favorables, la creencia de Ricardo de que ese animal es una cebra satisface SENSIBILIDAD. Después de todo, en un mundo cercano en el cual el animal fuese un cocodrilo, Ricardo—quien sabe reconocer las cebras a simple vista—no creería que ese animal es una cebra. Algo diferente ocurre con las negaciones de las hipótesis escépticas. En los mundos más cercanos donde tales hipótesis escépticas son verdaderas, Ricardo todavía creería que esas hipótesis no son verdaderas. Así, si Ricardo fuese un CC, mantendría la creencia de que (él mismo) no es un CC. Después de todo, ésta y otras hipótesis escépticas están diseñadas para que Ricardo no pueda detectar el engaño. Luego, la creencia de Ricardo en la proposición <No soy un CC> no satisface SENSIBILIDAD.

SENSIBILIDAD ofrece una manera elegante de precisar el concepto de alternativa relevante. En el mundo de Ricardo hay cebras y cocodrilos. Así, cuando Ricardo está ante una cebra, un mundo en el cual está ante un cocodrilo es un mundo cercano. Asumiendo que sólo hay un animal en el corral, la proposición <Hay un cocodrilo en el corral> es una alternativa relevante cuya falsedad Ricardo debe conocer para poder saber que hay una cebra en el corral. Por el contrario, un mundo en el cual Ricardo es un CC es un mundo muy lejano, pues un mundo en el cual Ricardo es un CC es muy diferente del mundo actual. Luego, la proposición <Soy un CC> es una alternativa irrelevante. Aunque SENSIBILIDAD pre-

18 Esta interpretación le obliga a revisar ciertos aspectos de los marcos semánticos de Lewis y Stalnaker.

dice que las negaciones de las hipótesis escépticas no son cognoscibles, también nos permite ignorar tales hipótesis con impunidad. No saber que uno no es un CC no es un obstáculo para tener conocimientos ordinarios sobre el mundo externo.

La teoría de Nozick ha dado lugar a muchos debates. Algunos autores han intentado mostrar que ciertos conocimientos inductivos no satisfacen SENSIBILIDAD. Así, aunque SENSIBILIDAD permita evitar el escepticismo sobre el mundo externo, también tiene el efecto colateral de conducir al escepticismo sobre el conocimiento inductivo (Vogel, 1987). Otros teóricos han señalado que la negación de CIERRE nos obliga a aceptar ciertas 'conjunciones abominables' como: "Sé que ese animal es una cebra, pero no sé que no soy un CC que sólo está alucinando que ese animal es una cebra" (DeRose, 1995, pp. 27–28). Tales conjunciones suenan incoherentes. Para muchos autores, su incoherencia respalda la validez irrestricta de CIERRE[19].

8.2 Rechazo de la proposición semi-escéptica

La teoría de Nozick analiza el conocimiento mediante el concepto modal de posibilidad. Muchos filósofos aceptan que la CONDICIÓN ANTI-SUERTE del conocimiento debe explicarse en términos modales, pero insisten en que no estamos obligados a aceptar SENSIBILIDAD. Hay condiciones modales del conocimiento que son compatibles con conocer las negaciones de las hipótesis escépticas. Esta parece ser una buena noticia, pues muchos filósofos consideran que CIERRE es un principio no-negociable (Cohen, 1999; Hawthorne, 2014; Pritchard, 2016; Williamson, 2000; algunas voces disidentes: Alspector-Kelly, 2019; Avnur, 2012; Coliva, 2015; Sherman y Harman, 2011; Sosa 2021).

SEGURIDAD es una condición modal compatible con CIERRE (Sosa 1999, 2002):

19 Ha habido intentos por mostrar que tales combinaciones no son incoherentes si se contextualizan de la manera adecuada (Avnur, 2012; Coliva, 2015; Sherman y Harman, 2011).

> SEGURIDAD. La creencia de *S* de que *p* es segura si y sólo si: Si *S* creyera que *p*, *p* no sería falsa.

Hay un debate acerca de cómo formular SEGURIDAD[20]. Sin embargo, podemos trabajar con la idea intuitiva de base: *S* creería que *p* sólo si *p* fuera verdadera. Una lectura influyente sostiene que esta condición se satisface cuando el consecuente es verdadero en la mayoría de los mundos cercanos en los cuales el antecedente también es verdadero (Pritchard, 2005a). SENSIBILIDAD y SEGURIDAD arrojan el mismo veredicto en relación con las proposiciones ordinarias sobre el mundo externo. En circunstancias favorables, Ricardo creería que ese animal es una cebra sólo si ese animal fuera una cebra. De estar ante un cocodrilo, Ricardo creería que ese animal es un cocodrilo. Sin embargo, SEGURIDAD arroja un veredicto diferente en relación con las negaciones de las hipótesis escépticas. Intuitivamente, un mundo en el cual uno es un CC es muy distinto del mundo actual. Luego, tal mundo no hace parte de la mayoría de los mundos cercanos en los cuales *S* cree que (él mismo) no es un CC. Luego, no hay mundos cercanos en los cuales uno cree que uno no es un CC y además no es verdad que uno no es un CC. En consecuencia, las creencias en las negaciones de las hipótesis escépticas son seguras.

Aunque SEGURIDAD es sólo una condición necesaria para el conocimiento, nos promete superar un obstáculo aparente para rechazar la PROPOSICIÓN SEMI-ESCÉPTICA. Después de todo, nos permite tener conocimientos anti-escépticos a pesar de que las hipótesis escépticas estén diseñadas para que el sujeto no pueda detectar que se encuentra en los escenarios escépticos descritos por tales hipótesis. Además, nos permite explicar por qué la PROPOSICIÓN SEMI-ESCÉPTICA puede parecer convincente. Para Sosa

20 Al igual que ocurre con SENSIBILIDAD, se suele relativizar SEGURIDAD a una base o método. Las interpretaciones habituales de esta condición también llevan a revisar ciertos aspectos de los marcos semánticos de Lewis y Stalnaker.

(1999, 2002), la diferencia entre SEGURIDAD y SENSIBILIDAD es tan sutil que es muy fácil confundirlas. Dicha diferencia sutil explica por qué pensamos equivocadamente que no podemos conocer las negaciones de las hipótesis escépticas.

Se han formulado varios contraejemplos a SEGURIDAD (para un contraejemplo influyente, véase Comesaña, 2005). Por razones de espacio, nos limitaremos a mencionar un problema relacionado con el escepticismo. Recordemos que SEGURIDAD sólo ofrece una condición necesaria para el conocimiento. Luego, la apelación a SEGURIDAD sólo será convincente si las creencias en las negaciones de las hipótesis escépticas pueden cumplir las demás condiciones del conocimiento. Supongamos que un sujeto nunca ha considerado la hipótesis del CC. Para saber que esa hipótesis es falsa, nuestro sujeto deberá formar la creencia anti-escéptica correspondiente de manera apropiada. Sin embargo, no está claro cómo podría hacer algo semejante. Una opción sería apelar a la percepción. Sin embargo, esta vía no parece estar disponible. Cuando formamos creencias con base en la percepción, detectamos los rasgos distintivos de ciertas entidades. Así, Ricardo forma la creencia de que ese animal es una cebra porque detecta visualmente sus rayas blancas y negras. Lamentablemente, las hipótesis escépticas están diseñadas para que el sujeto no pueda detectar que se encuentra en los escenarios escépticos. Incluso si aceptamos que las creencias en las negaciones de las hipótesis escépticas pueden ser seguras, ello no nos explica cómo podríamos formar tales creencias de manera apropiada. Pero, si no hay una manera apropiada de formar tales creencias, no parece posible tener conocimientos anti-escépticos[21].

8.3 Contextualismo

En la epistemología, el contextualismo es la tesis según la cual el valor semántico de expresiones epistémicas como 'saber', 'tener evidencia' y 'tener justificación' depende del contexto, de manera

21 Algunas observaciones de Wittgenstein (1969) apuntan a este problema. Para una discusión reciente, véase Greco (2020).

análoga a lo que ocurre con palabras como 'yo', 'alto' o 'plano'. Si Carlos dice 'Yo soy Napoleón', lo que dice es falso porque ese ejemplar de 'yo' refiere a Carlos. Si Napoleón dice 'Yo soy Napoleón', lo que dice es verdadero porque el segundo ejemplar de 'yo' refiere a Napoleón. Del mismo modo, distintas oraciones de la forma '*S* sabe que *p*' o '*S* no sabe que *p*' pueden tener distintas condiciones de verdad en distintos contextos de atribución[22]. Ello se debe a que distintos contextos de atribución imponen estándares diferentes y a que tales estándares determinan si las oraciones de conocimiento son verdaderas o no. En un contexto en el cual los estándares son altos, la oración 'Carlos sabe que hay una cebra en el corral' puede ser falsa. En otro contexto en el cual los estándares son bajos, esa misma oración puede ser verdadera. Existen tantas maneras de desarrollar esta idea como hay maneras de entender el concepto de estándar. Distintas teorías atribuyen estándares variables a distintos parámetros relacionados con las palabras 'saber' y 'conocimiento'. Así, algunos se enfocan en los umbrales para la atribución de la justificación (Cohen 1988, 1999), la evidencia (Neta 2002, 2003), o la fuerza de la posición epistémica del sujeto (DeRose 1995, 2002, 2009, 2017). Las teorías disponibles también difieren en relación con los factores que determinan el nivel de los estándares. Entre los factores más citados se encuentran los intereses, propósitos y temas de la conversación, la prominencia de ciertas posibilidades de error y lo que está en juego desde un punto de vista práctico (*practical stakes*)[23].

Algunos contextualistas han visto en la teoría de las alternativas relevantes un precursor del contextualismo (Cohen 1988; Lewis 1996). Desde su perspectiva, el contexto de atribución determina cuándo una alternativa es relevante[24]. Dado que Austin parte del

22 Nos referiremos a tales oraciones como 'oraciones de conocimiento'.

23 Algunas concepciones contextualistas no se enfocan en las atribuciones de conocimiento. Una de ellas es el contextualismo epistemológico de Michael Williams (2001), el cual se opone a la concepción del conocimiento como una especie natural.

24 Cohen (1999) abandona esta postura y desarrolla su propuesta contextualista independientemente de la teoría de las alternativas relevantes.

análisis del uso del lenguaje ordinario, las afinidades son innegables. Sin embargo, la teoría de las alternativas relevantes no implica el contextualismo. Para el teórico de las alternativas relevantes, son las circunstancias del sujeto las que determinan qué alternativas son relevantes y, por tanto, si el sujeto tiene o no conocimiento. Para el contextualista, el factor determinante de las condiciones de verdad de una oración de conocimiento es el contexto de atribución. Dado que las circunstancias del sujeto pueden concebirse de tal forma que no incluyan aspectos conversacionales y dado que quien realiza la atribución puede ser diferente del sujeto de conocimiento, ambas teorías son diferentes (Pryor, 2001; DeRose, 2009).

Pese a que el contextualismo es una tesis lingüística y no una tesis sobre el conocimiento, sus defensores piensan que el contextualismo nos permite ofrecer una nueva solución a la paradoja de cierre. En contextos ordinarios, los estándares son bajos. Luego, en tales contextos, las atribuciones de conocimientos ordinarios sobre el mundo externo y de conocimientos de las negaciones de las hipótesis escépticas pueden ser verdaderas. En contextos escépticos, los estándares son altos. Luego, en tales contextos, las atribuciones de conocimientos ordinarios sobre el mundo externo y de conocimientos de las negaciones de las hipótesis escépticas son falsas. Dado su carácter lingüístico, el contextualismo no se pronuncia sobre CIERRE, sino sobre una versión metalingüística de dicho principio. Sin embargo, el análisis anterior permite argumentar que cierta versión metalingüística de CIERRE es verdadera a través de los contextos. Después de todo, un condicional material puede ser verdadero cuando su antecedente y su consecuente son falsos, como ocurre en los contextos escépticos[25].

Una ventaja aparente de esta solución es que permite explicar parte de la fuerza de la paradoja de cierre. Nos parece que la

25 Algunos contextualistas niegan CIERRE o versiones metalingüísticas de éste (Heller, 1999). DeRose (2009, cap. 4) ha argumentado que los participantes en una conversación pueden resistirse a que suban los estándares para la atribución de conocimiento. Por tanto, aunque es frecuente que suban los estándares cuando examinamos la paradoja escéptica, tal efecto no es inevitable.

conclusión escéptica es correcta porque, al examinar las distintas oraciones que expresan la paradoja, no nos percatamos de que se produce un cambio sutil en los estándares de atribución de conocimiento. El contextualismo también permite explicar un aspecto del escepticismo que interesó a Hume (1739-1740): las dudas escépticas suelen estar circunscritas a contextos en los cuales hacemos filosofía y no tienen ningún impacto en la vida diaria (Greco 2007, 630). Ello se debe a que los estándares ordinarios son mucho más bajos que los estándares que operan cuando consideramos las hipótesis escépticas. Pese a que el contextualismo concede que las atribuciones de conocimiento son falsas en contextos escépticos, al menos preserva la verdad de gran parte de nuestras atribuciones de conocimiento en contextos ordinarios.

Quizá el aspecto más desconcertante del contextualismo es que parece obligarnos a aceptar cierta forma de 'ceguera semántica'. Cualquier locutor competente puede reconocer fácilmente que el valor semántico de palabras como 'yo' o 'alto' cambia según el contexto. Por el contrario, el contextualismo sobre las palabras epistémicas ha pasado inadvertido durante siglos. Este problema se ve exacerbado por la fuerte impresión que tenemos de que la conclusión escéptica contradice al sentido común; si estuviésemos ante un caso genuino de variabilidad contextual, no debería haber contradicción alguna entre las oraciones de conocimiento en contextos escépticos y las oraciones de conocimiento en contextos ordinarios. Y, de no haber contradicción, resulta difícil ver por qué nos parece estar ante un problema escéptico. Si los locutores competentes no son conscientes del carácter contextual de las palabras epistémicas, tampoco está claro cómo pueden éstos ajustar automáticamente los estándares a distintos contextos de atribución (Conee, 2014; Feldman, 1999; Pritchard, 2016; Schiffer, 1996; para algunas respuestas a este problema, véase Cohen, 1999, 2014; DeRose 2009, cap. 5)[26].

26 Si los locutores competentes pueden estar sistemáticamente engañados acerca del funcionamiento contextual de las palabras epistémicas, tampoco queda claro cómo éstos podrían rehusarse a que los estándares suban en contextos escépticos. Véase la nota 25.

El contextualismo también enfrenta un problema análogo al que enfrentan los defensores de SEGURIDAD. Dado que se trata de una tesis lingüística y no de una teoría epistemológica, el contextualismo no explica cómo, mientras permanezcamos en contextos ordinarios, podemos tener conocimientos ordinarios sobre el mundo externo y conocer las negaciones de las hipótesis escépticas. Así, en el mejor de los casos, el contextualismo es una teoría anti-escéptica incompleta. En el peor de los casos, el contextualismo distorsiona la paradoja de cierre, la cual señala que, incluso en los contextos ordinarios, no podemos conocer las negaciones de las hipótesis escépticas y, por tanto, no tenemos conocimientos ordinarios sobre el mundo externo (Conee, 2014; Feldman, 1999; Greco, 2007; Kornblith, 2000; Pritchard, 2016).

9. La paradoja de subdeterminación

En un comienzo, la paradoja de cierre constituyó la manera canónica de presentar el escepticismo sobre el mundo externo. Sin embargo, pronto se vio que había otras maneras de entender la exigencia de poder 'descartar' hipótesis escépticas. Tal es el caso de la paradoja de subdeterminación, donde tal exigencia se traslada a la evidencia que respalda nuestras creencias en proposiciones ordinarias sobre el mundo externo. En términos generales, para tener conocimientos de proposiciones ordinarias sobre el mundo externo, es necesario que la evidencia del sujeto 'favorezca' a tales proposiciones sobre hipótesis escépticas conocidas. Sin embargo, es difícil ver cómo podría cumplirse tal exigencia[27].

La paradoja de subdeterminación apela a un principio de subdeterminación como PRINCIPIO CONECTIVO. Dicho principio se aplica a conceptos que son necesarios, pero no suficientes, para ciertas formas de conocimiento, como el concepto EVIDENCIA.

27 Algunos autores sostienen que la paradoja de subdeterminación es 'más fundamental' que la paradoja de cierre (Cohen, 1999; Conee y Feldman, 2004; Greco, 2000, 2007). Más recientemente, Pritchard (2016) ha argumentado que las dos paradojas son lógicamente diferentes y que obedecen a motivaciones diferentes. Por ello, no es posible resolverlas desplegando una única estrategia anti-escéptica.

Luego, es posible formular el PRINCIPIO CONECTIVO como una condición necesaria, pero no suficiente, para ciertas formas de conocimiento. De manera alternativa, es posible formular dicho principio en relación con otro estatus epistémico menos exigente que el conocimiento, como la justificación. En esta sección exploraremos la segunda opción:

> SUBDETERMINACIÓN. Si *S* sabe que *p* y *q* son incompatibles, y además la evidencia de *S* no favorece a *p* sobre *q*, entonces *S* carece de justificación para creer que *p*[28].

Supongamos que Ricardo sabe que <Hay una cebra en el corral> es incompatible con <Hay un cocodrilo en el corral>. Supongamos además que la evidencia de Ricardo no favorece a la primera proposición sobre la segunda proposición. Quizá las condiciones de iluminación no son muy buenas y Ricardo no logra ver la forma equina y las rayas blancas y negras del animal. En este caso, SUBDETERMINACIÓN nos dice que Ricardo no tiene justificación para creer que hay una cebra en el corral. Por fortuna, esta conclusión no nos lleva al escepticismo, pues hay muchos otros casos en los cuales Ricardo podría tener mejor evidencia, como cuando logra ver a una cebra a una distancia razonable y a plena luz del día.

Lamentablemente, basta con aplicar SUBDETERMINACIÓN a las hipótesis escépticas para alcanzar la conclusión escéptica. Después de todo, SUBDETERMINACIÓN exige que, para tener justificación para creer proposiciones ordinarias sobre el mundo externo, es preciso contar con evidencia que favorezca a tales proposiciones sobre las hipótesis escépticas. Sin embargo, es difícil ver cómo podríamos tener tal evidencia.

Una opción es que nuestra evidencia empírica favorece a las proposiciones ordinarias sobre el mundo externo sobre las hipóte-

28 Algunos principios análogos han sido propuestos por Brueckner (1994), Cohen (1998), Pritchard (2016), Vogel (2004) y Yalçin (1992). Aunque Stroud no menciona SUBDETERMINACIÓN, apela implícitamente a él en su discusión del problema escéptico (Stroud, 1991, p. 32).

sis escépticas. Sin embargo, PARIDAD parece excluir esta posibilidad (Sección 5). Si aceptamos PARIDAD, es posible que haya escenarios escépticos en los cuales nuestras creencias ordinarias sobre el mundo externo son masivamente falsas mientras tenemos la misma evidencia empírica que tenemos en los casos buenos. Si las hipótesis escépticas pueden replicar exactamente la evidencia empírica que tenemos en los casos buenos, nuestra evidencia empírica es compatible con estar masivamente equivocados. Pero, si nuestra evidencia empírica es compatible con estar masivamente equivocados, nuestra evidencia empírica no nos da mejores razones a favor de las proposiciones ordinarias sobre el mundo externo que de las hipótesis escépticas. Luego, no parece que dicha evidencia empírica favorezca a las proposiciones ordinarias sobre el mundo externo sobre las hipótesis escépticas.

Dado que carecemos de evidencia empírica que satisfaga SUBDETERMINACIÓN, quizá tengamos razones *a priori* que favorezcan a las proposiciones ordinarias sobre el mundo externo frente a las hipótesis escépticas[29]. Desafortunadamente, esta opción no parece ser prometedora. Las proposiciones ordinarias sobre el mundo externo, al igual que las hipótesis escépticas, versan sobre rasgos a la vez *contingentes* y *específicos* de entidades que existen *en* el espacio: <Hay una cebra en el corral>, <Soy un cerebro *en* una cubeta cuyas experiencias *están siendo* producidas por una computadora>, etc. Si tuviésemos consideraciones *a priori* que favorecieran a las proposiciones ordinarias sobre el mundo externo frente a las hipótesis escépticas, entonces sería posible tener justificación acerca de la (no) instanciación de rasgos *contingentes* y *específicos* de entidades que existen en el espacio sin apelar a la percepción. Eso parece comprometernos con la existencia de una fuente mágica de justificación (Avnur, 2012; Coliva, 2015; McDowell, 1996; Neta, 2011, 2019; Pritchard, 2012, 2016; para una visión alternativa, véase Hawthorne, 2002).

29 Por ejemplo, Cohen (2010) y Wedgwood (2013) derivan la justificación de las negaciones de las hipótesis escépticas de la racionalidad de ciertas reglas inferenciales.

Si la distinción entre fuentes empíricas y *a priori* es exhaustiva, no parece posible cumplir con la exigencia que plantea SUBDETERMINACIÓN. Luego, la conclusión escéptica parece inevitable. Un sujeto cualquiera carece de justificación para creer múltiples proposiciones ordinarias sobre el mundo externo. Este punto refuerza la impresión general según la cual apelar a SEGURIDAD o al carácter contextual del vocabulario epistémico es insuficiente como estrategia anti-escéptica. Es difícil entender cómo un sujeto podría adquirir conocimientos sobre el mundo externo si PARIDAD excluye la evidencia empírica y si, además, no hay razones *a priori* que respalden proposiciones que versan sobre rasgos a la vez contingentes y específicos de entidades que existen en el espacio.

9.1 Rechazo de subdeterminación

Una manera de bloquear la paradoja es rechazar SUBDETERMINACIÓN. Para tal efecto, podría defenderse una nueva versión de la teoría de las alternativas relevantes, esta vez restringida a la justificación. Es posible que *S* tenga justificación para creer que *p*, aunque la evidencia que tiene *S* no favorezca a *p* sobre las hipótesis escépticas. Una solución semejante tendría que lidiar, empero, con el problema de las conjunciones abominables: "Tengo evidencia para creer que ese animal es una cebra, pero no tengo evidencia de que no soy un CC que está alucinando de manera no-verídica que ese animal es una cebra".

En *Sobre la Certeza*, Ludwig Wittgenstein (1969) plantea algunas observaciones que parecen mitigar el carácter contraintuitivo de esta postura. Wittgenstein sostiene que todas nuestras prácticas epistémicas tienen lugar dentro de un marco que las hace posibles. En un pasaje famoso, compara los componentes de ese marco con las bisagras de una puerta: "Si queremos que la puerta gire", escribe Wittgenstein, "las bisagras deben permanecer colocadas" (Wittgenstein, 1969, párr. 343). De manera análoga, las preguntas y las dudas que planteamos dependen de que algunas proposiciones estén exentas de duda (Wittgenstein, 1969, párr. 341). Entre tales proposiciones se encuentran las negaciones de hipótesis escépticas como <No soy un

CC> o, de manera más general, la proposición <No soy víctima de un engaño perceptual y cognitivo generalizado>.

Muchos seguidores de Wittgenstein han visto en estas observaciones una manera de negar SUBDETERMINACIÓN sin sucumbir a la conclusión escéptica. En su opinión, no podemos tener evidencia empírica que hable a favor de las bisagras (Coliva, 2015, 2022; Wright, 1985, 2004, 2014). Por tanto, las bisagras no pueden estar justificadas por evidencia empírica. Sin embargo, un análisis adecuado del papel de las bisagras podría ayudarnos a preservar la justificación de las proposiciones ordinarias sobre el mundo externo.

Para Wright (2004), las bisagras son los contenidos de actitudes de confianza (*trust*). Si Ricardo confía en que (él mismo) no es un CC, Ricardo se encuentra en un estado doxástico que es incompatible con que suspenda el juicio respecto de dicha hipótesis. Para Ricardo, es como si él creyera que (él mismo) no es un CC, pero sin contar con evidencia alguna que hable en contra de tal hipótesis escéptica. Pese a su falta de respaldo evidencial, las actitudes de confianza en las bisagras permiten regular todas las actividades cognitivas. Por ejemplo, las bisagras determinan si un sujeto cuenta con evidencia a favor de una proposición, si puede fiarse de una fuente epistémica para responder a una pregunta, o si cierta fuente epistémica debería prevalecer sobre otra fuente epistémica (Wright, 2014, pp. 242–243). Dado que las bisagras permiten alcanzar ciertos objetivos epistémicos, como maximizar la formación de creencias verdaderas, Wright (1985, 2004) piensa que es epistémicamente racional confiar en las bisagras, aunque tal confianza no esté basada en evidencia. Wright introduce el término 'autorización' (*entitlement*) para referirse a este tipo de racionalidad epistémica no-evidencial.

Desafortunadamente, la postura de Wright no es del todo convincente. Algunos autores han señalado que no se puede mostrar que las actitudes de confianza promueven la obtención de nuestros objetivos epistémicos sin hacer suposiciones sustantivas sobre el mundo externo. Por lo tanto, los argumentos que propone Wright a favor de su teoría de la autorización cometen una petición de

principio (Echeverri, 2024; Williams, 2012). Por otra parte, muchos filósofos han rechazado el intento de Wright por abrirle espacio a un tipo de racionalidad epistémica no-evidencial (Echeverri, 2024; Jenkins, 2007; Pritchard, 2005b). Una analogía quizá sea suficiente para ilustrar el problema. Consideremos la proposición sustantiva <El número de planetas en el universo es par>. A falta de evidencia empírica relevante, no parece ser epistémicamente racional confiar en que el número de planetas en el universo es par, incluso si hacerlo podría permitirnos formar muchas creencias verdaderas sobre el universo. Algunas concepciones epistemológicas más tradicionales nos ofrecen una explicación de ese veredicto. Uno no puede tener una actitud hacia una proposición que sea incompatible con la suspensión del juicio sobre esa proposición a menos que tenga evidencia relevante para la verdad o la falsedad de tal proposición. El problema es que la teoría de la autorización nos obliga a violar esta máxima en relación con proposiciones no menos sustantivas como <No soy un CC>. En resumen, es un misterio por qué uno *debería* y *podría* confiar en dicha proposición sin tener evidencia que la respalde[30].

30 En respuesta a estas dificultades, Coliva (2015, 2022) ha intentado mostrar que las bisagras pueden ser epistémicamente racionales si se las considera como constitutivas de la racionalidad humana, del mismo modo en que ciertas reglas son constitutivas del juego de ajedrez. Para exámenes críticos de dicha propuesta, véase Avnur (2017) y Echeverri (2023a).

Otros teóricos han desarrollado concepciones alternativas de las bisagras. En un extremo están las teorías epistémicas que intentan mostrar que (ciertas) bisagras pueden figurar como contenidos de conocimientos (Engel, 2016; Greco 2020; Neta, 2021; Pritchard, 2011). Tales propuestas permiten preservar SUBDETERMINACIÓN. En el otro extremo están aquellas teorías que construyen las bisagras como entidades que no son aptas para ser conocidas (Moyal-Sharrock, 2004; Pritchard, 2016; Schönbaumsfeld, 2016). Algunos autores de este último grupo han intentado preservar SUBDETERMINACIÓN complementando el marco de las bisagras con una u otra forma de disyuntivismo epistemológico. Discutiremos esta última postura a continuación.

9.2 Rechazo de proposición semi-escéptica

Varios filósofos han señalado que la PROPOSICIÓN SEMI-ESCÉPTICA descansa sobre una concepción demasiado restringida de la evidencia disponible en los casos buenos. Veamos por qué.

Si uno está en un caso bueno, parece apropiado utilizar construcciones fácticas de la forma '*S* ve que *p*' para respaldar sus creencias perceptuales. Si Pedro me llama por teléfono para preguntarme si Lara vino a trabajar, yo podría utilizar una construcción fáctica para responder a su pregunta. Al visitar el cubículo de Lara y verla trabajando, podría responderle: 'Sí, veo que Lara vino a trabajar'. Si además carezco de razones para dudar y estoy en un caso bueno en el cual puedo emplear mis capacidades de manera fiable, lo que le dije a Pedro era verdad; vi que Lara vino a trabajar (Echeverri, 2023b; Kern, 2017; McDowell, 2002a, 2002b; Neta, 2009, 2011; Pritchard, 2012, 2016).

Las observaciones anteriores motivan un tipo de posición que se conoce como 'disyuntivismo epistemológico' (DE). La idea básica es que la evidencia que uno tiene en un caso bueno es mejor que la evidencia que uno tiene en un caso malo. En un caso bueno, puedo tener evidencia fáctica que implica la verdad de su contenido. Si veo que Lara vino a trabajar, es verdad que Lara vino a trabajar. Como esta evidencia fáctica no está disponible en el caso malo, tal evidencia favorece a las proposiciones ordinarias acerca del mundo externo sobre las hipótesis escépticas. Dado que los casos buenos y los casos malos son indiscriminables introspectivamente, los partidarios del DE describen su posición mediante una disyunción excluyente: o bien el sujeto ve que *p* (en los casos buenos), o bien al sujeto sólo le parece que ve que *p* (en los casos malos).

De ser exitosa, esta propuesta ofrece diagnósticos interesantes del problema escéptico. Para Pritchard (2016), pensamos que nuestras razones están subdeterminadas porque tácitamente adoptamos la tesis de la insularidad de las razones[31]. Esta tesis dice que las razones que tenemos en los casos buenos son compatibles con

31 Pritchard emplea 'razones' para referirse a evidencia a la cual tenemos acceso reflexivo.

la falsedad generalizada de las proposiciones sobre el mundo externo. El ejemplo de la conversación telefónica con Pedro sugiere que la tesis de la insularidad de las razones no está respaldada por nuestras prácticas ordinarias. Para McDowell (2011) y Kern (2017), la PROPOSICIÓN SEMI-ESCÉPTICA puede parecer inevitable sólo si adoptamos una imagen cuestionable de la falibilidad humana. Pensamos que, si tenemos capacidades de conocimiento falibles, es porque todos nuestros logros epistémicos están basados en un factor compartido por los casos buenos y los casos malos (Burge, 2005; 2011). Sin embargo, esta concepción no es obligatoria. Una visión alternativa concibe la falibilidad humana, no como un factor compartido entre los casos buenos y los casos malos, sino como una propiedad de nuestras capacidades de conocimiento. Es propio de una capacidad falible que pueda tener dos tipos de manifestaciones: empleos exitosos y empleos no-exitosos. Cuando un empleo es exitoso, dicho empleo nos proporciona razones fácticas que excluyen cualquier posibilidad de error. Cuando un empleo es no-exitoso, dicho empleo se queda corto desde un punto de vista epistémico. Por ejemplo, en lugar de proporcionarnos conocimientos perceptuales, nos proporciona creencias falsas fundadas en experiencias no-verídicas (Kern, 2017; McDowell, 2011, 2013; Millar, 2019 y Schellenberg, 2018 ofrecen maneras diferentes de desarrollar este tipo de respuesta)[32].

32 Moore (1944), Austin (1962) y Dretske (1969) desarrollan ideas que prefiguran muchos de estos temas. La teoría del 'conocimiento primero' de Williamson (2000) también ofrece una respuesta análoga. Williamson (2000) argumenta que la postura escéptica presupone que la evidencia que uno tiene está determinada por sus poderes de discriminación introspectiva. Sin embargo, la evidencia que uno tiene puede ir más allá de lo que uno puede discriminar introspectivamente. Supongamos que un sujeto está en un caso bueno y que tiene conocimientos sobre el mundo externo. Para Williamson, la evidencia (E) de un sujeto consiste en todos los conocimientos del sujeto (C) (E=C). Luego, pese a sus limitaciones de discriminación introspectiva, el caso bueno y el caso malo son asimétricos desde el punto de vista de la evidencia. Si un sujeto sabe que *p*, es verdad que *p*. Luego, un sujeto que está en un caso bueno tiene evidencia que favorece a *p* sobre todas aquellas hipótesis escépticas que implican *no-p*. Si, por el contrario, el sujeto se encuentra en un caso malo, sólo le parece que sabe que *p*. Como tal parecer no implica la verdad de *p*, su evidencia es menos buena que la evidencia disponible en el caso malo.

El DE es una de las estrategias anti-escépticas más discutidas en los últimos años. Por desgracia, el DE no está exento de problemas. Quizá la dificultad más notable la representa el 'nuevo problema del genio maligno'. Este problema emerge cuando consideramos a un sujeto que se encuentra en un escenario escéptico, pero utilizamos dicho escenario para evaluar la justificación de las creencias de ese sujeto. Pensemos en el escenario del CC. Muchos críticos piensan que, si a Ricardo le parece que ese animal es una cebra, Ricardo tiene justificación suficiente para creer que ese animal es una cebra. También piensan que la justificación de Ricardo no difiere de manera relevante de la justificación que Ricardo tendría en un caso bueno. Un argumento a favor de esta tesis es que, al formar su creencia, Ricardo no se ha comportado de manera irresponsable al formar su creencia. Después de todo, Ricardo no puede detectar que se encuentra en ese escenario escéptico. Podemos encapsular este razonamiento en la siguiente tesis:

> IDENTIDAD EPISTÉMICA. La justificación que tenemos en los casos buenos es la misma que la justificación que tenemos en los casos malos.

Como el DE postula una asimetría de evidencia o razones entre los casos buenos y los casos malos y tal asimetría basta para tener mejor justificación en los casos buenos, el DE viola IDENTIDAD EPISTÉMICA[33].

Los defensores del DE suelen 'morder la bala' y ofrecer caracterizaciones alternativas de nuestra posición epistémica en los escenarios escépticos. Una solución popular sostiene que el sujeto en el caso malo no tiene justificación alguna. En el caso malo, formar creencias empíricas no es apropiado, ni está permitido desde un punto de vista epistémico. Sin embargo, dado que el caso malo es indiscriminable introspectivamente de un caso bueno, podemos

33 Para una discusión de este problema, véase los ensayos en Dorsch y Dutant (de próxima publicación).

excusar al sujeto que está en el caso malo por creer lo que cree. Como las excusas son evaluaciones epistémicas meramente negativas, el sujeto en cuestión no posee justificación (Littlejohn de próxima publicación; Pritchard, 2012; Williamson de próxima publicación; para una crítica de esta postura, véase Madison, 2018).

Algunos autores han insistido en que esta respuesta no es suficiente. En su opinión, las creencias del sujeto que está en el caso malo tienen ciertas características epistémicas positivas que el concepto EXCUSA no captura (Ranalli, 2021). Una respuesta posible es que su insatisfacción se desprende de una equivocación conceptual. Una evaluación epistémica puede referirse a una creencia (prospectiva) o al sujeto que forma dicha creencia (prospectiva). Una cosa es decir que es apropiado o que está permitido *creer* que *p* desde un punto de vista epistémico. Y otra cosa muy distinta es sostener que el sujeto se ha comportado correctamente desde un punto de vista epistémico. Si el sujeto ha hecho todo lo que está a su alcance para formar una creencia verdadera, podemos evaluar al sujeto como un agente epistémicamente responsable. De allí no se sigue, empero, que la creencia (prospectiva) del sujeto goce de un estatus epistémico positivo, es decir, que dicha creencia (prospectiva) sea apropiada o que esté permitido formarla.

Una solución alternativa consiste en introducir distintos tipos o niveles de justificación. Así, podría afirmarse que el sujeto en el caso bueno y el sujeto en el caso malo comparten un tipo de evidencia fenoménica, pero que sólo el sujeto que está en el caso bueno goza, además, de evidencia fáctica (Schellenberg, 2018). Luego, pese a que ambos sujetos están justificados, no gozan del mismo tipo o grado de justificación. Se ha objetado que esta respuesta no es suficiente, pues no explica por qué la evidencia o las razones fácticas son necesariamente mejores que la evidencia o las razones no-fácticas (Ranalli, 2021). Quizá podría replicarse que las razones fácticas están subdeterminadas de una forma en que las razones no-fácticas no lo están.

9.3 Limitar el daño

Ninguna de las soluciones anti-escépticas que hemos examinado está exenta de problemas. Por tanto, quizá podríamos aceptar la conclusión escéptica, pero intentar 'limitar el daño'. Pese a que no tengamos evidencia que favorezca a las proposiciones ordinarias acerca del mundo externo sobre las hipótesis escépticas, todavía es posible identificar aspectos epistémicos positivos que están presentes únicamente en los casos buenos. Por ejemplo, podría afirmarse que sólo un sujeto que se encuentra en un caso bueno puede formar creencias sobre el mundo externo que no son verdaderas por pura suerte. Quizá esto sea suficiente para satisfacer un concepto mínimo de conocimiento. Esta estrategia exige explicar, empero, por qué muchos filósofos no se han dado por satisfechos con una asimetría epistémica centrada exclusivamente en el conocimiento.

Algunas formas recientes de pragmatismo también pueden verse como propuestas escépticas que intentan limitar el daño. Tales teorías sostienen que, aunque no tengamos justificación epistémica para creer proposiciones ordinarias sobre el mundo externo, contamos con justificación práctica para creer tales proposiciones y dicha justificación práctica es suficiente para proporcionarnos un tipo de justificación global para creer tales proposiciones (*all things considered justification*) (Hirsch, 2015; Rinard, 2021). Queda por ver si podríamos reconciliar tales posturas pragmatistas con la manera como formamos y evaluamos nuestras creencias en la vida diaria.

10. Conclusiones

Es frecuente entender el escepticismo como una posición incoherente que podemos ignorar como teóricamente irrelevante. Esta concepción pasa por alto otra manera más fecunda de entender el escepticismo. Según esta concepción, el escepticismo es una serie de paradojas que emergen cuando reflexionamos filosóficamente sobre el funcionamiento de nuestros conceptos epistémicos ordinarios. En este capítulo, hemos 1) mostrado cómo esta visión permite otorgarle un lugar central al escepticismo en la epistemología contemporánea, 2) explorado los rasgos centrales de algunas paradojas

escépticas y 3) presentado algunas soluciones influyentes. Dadas las limitaciones de espacio, hemos dejado de lado otras paradojas, posturas y argumentos de gran importancia. No obstante, esperamos que este recorrido parcial sirva como punto de partida para continuar explorando este fascinante tema[34].

Referencias

Alspector-Kelly, M. (2019). *Against Knowledge Closure*. Cambridge University Press.

Alston, W. P. (1991). *Perceiving God. The Epistemology of Religious Experience*. Cornell University Press.

Austin, J. L. (1946). Other Minds. En J. L. Austin, *Philosophical Papers* (pp. 76–116). Oxford University Press.

Austin, J. L. (1962). *Sense and Sensibilia*. Oxford University Press.

Avnur, Y. (2012). Closure Reconsidered. *Philosophers' Imprint, 12*, 1–16.

Avnur, Y. (2017). On What Does Rationality Hinge? *International Journal for the Study of Skepticism, 7*, 246–257.

Becker, K. (2007). *Epistemology Modalized*. Routledge.

Bergmann, M. (2021). *Radical Skepticism and Epistemic Intuition*. Oxford University Press.

Brueckner, A. (1994). The Structure of the Skeptical Argument. *Philosophy and Phenomenological Research, 54*(4), 827–835.

Burge, T. (2005). Disjunctivism and Perceptual Psychology. *Philosophical Topics, 33*(1), 1–78.

Burge, T. (2011). Disjunctivism Again. *Philosophical Explorations, 14*(1), 43–80.

Burnyeat, M. (1982). Idealism and Greek Philosophy: What Descartes Saw and Berkeley Missed. *Philosophical Review, 91*(1), 3–40.

Byrne, A. (2004). How Hard Are Sceptical Paradoxes? *Noûs, 38*(2), 299–325.

34 La elaboración de este texto ha contado con el apoyo del Programa UNAM-PAPIIT (IA 400124: "Escepticismo y racionalidad"). También agradezco a Octavio Cervantes y a un evaluador anónimo por sus comentarios sobre este texto.

Byrne, A. (2014). McDowell and Wright on Anti-Scepticism, etc. En D. Dodd y E. Zardini (eds.), *Scepticism and Perceptual Justification* (pp. 275–296). Oxford University Press.

Carnap, R. (1928). *Pseudoproblemas en la filosofía. La psique ajena y la controversia sobre el realismo* (Trad. L. Mues de Schrenk). IIFs-UNAM.

Cassam, Q. (2007). *The Possibility of Knowledge*. Oxford University Press.

Cavell, S. (1979). *The Claim of Reason*. Oxford.

Chalmers, D. (2005). The Matrix as Metaphysics. En C. Grau (ed.), *Philosophers Explore the Matrix*. Oxford University Press.

Clarke, T. (1972). The Legacy of Skepticism. *Journal of Philosophy, 69*(20), 754–769.

Cohen, S. (1988). How to Be a Fallibilist. *Philosophical Perspectives, 2*, 91–123.

Cohen, S. (1998). Two Kinds of Skeptical Argument. *Philosophy and Phenomenological Research, 58*(1), 143–159.

Cohen, S. (1999). Contextualism, Skepticism, and the Structure of Reasons. *Philosophical Perspectives, 13*, 57–89.

Cohen, S. (2010). Bootstrapping, Defeasible Reasoning, and *A Priori* Justification. *Philosophical Perspectives, 24*, 141–159.

Cohen, S. (2014). Contextualism Defended y Contextualism Defended Some More. En M. Steup, J. Turri y E. Sosa (eds.), *Contemporary Debates in Epistemology*. (2ª ed., Vol. 14). (pp. 69–75; 79–83). Wiley Blackwell.

Coliva, A. (2015). *Extended Rationality: A Hinge Epistemology*. Palgrave.

Coliva, A. (2022). *Wittgenstein Rehinged*. Anthem.

Comesaña, J. (2005). Unsafe Knowledge, *Synthese, 146*(3), 395–404.

Conant, J. (2004). Varieties of Scepticism. En Denis McManus (ed.). *Wittgenstein and Scepticism* (pp. 97–136). Routledge.

Conee, E. (2014). Contextualism Contested y Contextualism Contested Some More. En M. Steup, J. Turri y E. Sosa (eds.). *Contemporary Debates in Epistemology*. (2ª, Vol. 14). (pp. 60–69; 75–79). Wiley Blackwell.

Conee, E. y R. Feldman (2004). *Evidentialism: Essays in Epistemology*. Oxford University Press.

David, M. y T. A. Warfield (2008). Knowledge-Closure and Skepticism. En Q. Smith (Ed.), *Epistemology: New Essays* (pp. 137–187). Oxford University Press.

DeRose, K. (1995). Solving the Skeptical Problem. *Philosophical Review, 104*(1), 1–52.

DeRose, K. (2002). Assertion, Knowledge, and Context. *Philosophical Review, 111*(2), 167–203.

DeRose, K. (2009). *The Case for Contextualism: Knowledge, Skepticism, and Context*, (Vol. 1). Oxford University Press.

DeRose, K. (2017). *The Appearance of Ignorance: Knowledge, Skepticism, and Context*. (Vol. 2). Oxford University Press.

Descartes, R. (1641). *Meditaciones Metafísicas y Otros Textos*. (Trad. E. López y M. Graña). Gredos.

Dorsch, F. y J. Dutant (Eds.) (de próxima publicación). *The New Evil Demon Problem*. Oxford University Press.

Dretske, F. (1969). *Seeing and Knowing*. Chicago University Press.

Dretske, F. (1970). Epistemic Operators. *Journal of Philosophy, 67*(24), 1007–1023.

Dretske, F. (1971). Conclusive Reasons. *Australasian Journal of Philosophy, 49*(1), 1–22.

Dretske, F. (1981). *Knowledge and the Flow of Information*. MIT Press.

Dretske, F. (2014). The Case Against Closure. En M. Steup, J. Turri y E. Sosa (Eds.), *Contemporary Debates in Epistemology*. (2ª ed., vol. 14). (pp. 13–25). Wiley Blackwell.

Echeverri, S. (2017). How to Undercut Radical Skepticism. *Philosophical Studies, 174*(5), 1299–1321.

Echeverri, S. (2018). Epistemology Without Concepts? *Metascience, 27*(1), 117–121.

Echeverri, S. (2020). Perceptual Knowledge, Discrimination, and Closure. *Erkenntnis, 85*(6), 1361–1378.

Echeverri, S. (2023a). Moderatism and Truth. *Canadian Journal of Philosophy, 53*(3), 271–287.

Echeverri, S. (2023b). A-Rational Epistemological Disjunctivism. *Philosophy and Phenomenological Research, 106*(3), 692–719.

Echeverri, S. (2024). Humean Skepticism and Entitlement. En S. Stapleford y V. Wagner (Eds.), *Hume and Contemporary Epistemology* (pp. 183–205). Routledge.

Engel, P. (2016). Epistemic Norms and the Limits of Epistemology. *International Journal for the Study of Skepticism, 6*(2–3), 228–247.

Fantl, J. y M. McGrath (2009). *Knowledge in an Uncertain World*. Oxford University Press.

Feldman, R. (1999). Contextualism and Skepticism. *Philosophical Perspectives, 1*, 91–114.

Firth, R. (1978). Are Epistemic Concepts Reducible to Ethical Concepts? En A. Goldman y J. Kim (Eds.), *Values and Morals* (pp. 215–229). D. Reidel.

Goldman, A. (1976). Discrimination and Perceptual Knowledge. *Journal of Philosophy, 73*(11), 771–791.

Goldman, A. (1979). What is Justified Belief? En G. Pappas (ed.), *Justification and Knowledge* (pp. 1–25). D. Reidel.

Greco, J. (2000). *Putting Skeptics in Their Place: The Nature of Skeptical Arguments and Their Role in Philosophical Inquiry*. Cambridge University Press.

Greco, J. (2007). External World Skepticism. *Philosophy Compass, 2*(4), 625–649.

Greco, J. (2020). *The Transmission of Knowledge*. Cambridge University Press.

Hawthorne, J. (2002). Deeply Contingent A Priori Knowledge. *Philosophy and Phenomenological Research, 65*(2), 247–269.

Hawthorne, J. (2004). *Knowledge and Lotteries*. Oxford University Press.

Hawthorne, J. (2014). The Case for Closure. En M. Steup, J. Turri y E. Sosa (Eds.), *Contemporary Debates in Epistemology* (pp. 40–56). (2ª ed., Vol. 14). Wiley Blackwell.

Heller, M. (1999). Relevant Alternatives and Closure. *Australasian Journal of Philosophy, 77*(2), 196–208.

Hirsch, E. (2015). *Radical Skepticism and the Shadow of Doubt*. Bloomsbury.

Hume, D. (1739-1740). *A Treatise of Human Nature*. Oxford University Press, 2000.

Jenkins, C. (2007). Entitlement and Rationality. *Synthese, 157*(1), 25–45.

Kant, I. (1781/1787). *Crítica de la razón pura* (Trad. Mario Caimi). Fondo de Cultura Económica, UAM, UNAM, 2018.

Kern, A. (2017). *Sources of Knowledge: On the Concept of a Rational Capacity for Knowledge*. Harvard University Press.

Kornblith, H. (2000). The Contextualist Evasion of Epistemology. *Philosophical Issues, 10*, 24–32.

Kvanvig, J. (2006). Closure Principles. *Philosophy Compass, 1*(3), 256–267.

Levin, M. (2000). Demons, Possibility and Evidence. *Noûs, 34*(3), 422–440.

Lewis, D. (1973). *Counterfactuals*. Blackwell.

Lewis, D. (1996). Elusive Knowledge. *Australasian Journal of Philosophy, 74*(4), 549–567.

Littlejohn, C. (de próxima publicación). A Plea for Epistemic Excuses. En F. Dorsch y J. Dutant (eds.). *The New Evil Demon Problem*. Oxford University Press.

Maddy, P. (2017). *What Do Philosophers Do? Skepticism and the Practice of Philosophy*. Oxford University Press.

Madison, B. J. C. (2018). On Justifications and Excuses. *Synthese, 195*(10), 4551–4562.

McDowell, J. (1996). *Mind and World*. (2a ed.). Harvard University Press.

McDowell, J. (2002a). Knowledge and the Internal Revisited. *Philosophy and Phenomenological Research, 64*(1), 97–105. Reimpreso en J. McDowell (2009), *The Engaged Intellect* (pp. 278–287). Harvard University Press.

McDowell, J. (2002b). Responses. En N. H. Smith (ed.). *Reading McDowell: On Mind and World* (pp. 269–305). Routledge.

McDowell, J. (2011). *Perception as a Capacity for Knowledge*. Marquette University Press.

McDowell, J. (2013). Tyler Burge on Disjunctivism (II). *Philosophical Explorations, 16*(3), 259–279.

Millar, A. (2019). *Knowing by Perceiving*. Oxford University Press.

Moore, G. E. (1925). A Defence of Common Sense. En J. H. Muirhead (Ed.), *Contemporary British Philosophy* (pp. 192–233). George Allen & Unwinn. Reimpreso en T. Baldwin (Ed.) (1993). *G. E. Moore: Selected Writings* (pp. 106–133). Routledge.

Moore, G. E. (1939). Proof of an External World. *Proceedings of the British Academy, 25*(5), 273–300. Reimpreso en T. Baldwin (Ed.) (1993), *G. E. Moore: Selected Writings* (pp. 147–170). Routledge.

Moore, G. E. (1944). Four Forms of Scepticism. En G. E. Moore. *Philosophical Papers* (pp. 196–226). Allen & Unwin.

Moore, G. E. (1953). *Some Main Problems of Philosophy*. Routledge.

Moyal-Sharrock, D. (2004). *Understanding Wittgenstein's On Certainty*. Palgrave.

Neta, R. (2002). S Knows that P. *Noûs, 36*(4), 663–689.

Neta, R. (2003). Contextualism and the Problem of the External World. *Philosophy and Phenomenological Research, 66*(1), 1–31.

Neta, R. (2009). Mature Human Knowledge as a Standing in the Space of Reasons. *Philosophical Topics, 37*(1), 115–132.

Neta, R. (2011). A Refutation of Cartesian Fallibilism. *Noûs, 45*(4), 658–695.

Neta, R. (2021). An Evidentialist Account of Hinges. *Synthese, 198*(Suppl. 15), 3577–3591.

Nozick. R. (1981). *Philosophical Explanations*. Harvard University Press.

Popkin, R. (2003). *The History of Scepticism from Savonarola to Bayle*. (3a ed. aumentada). Oxford University Press.

Pritchard, D. (2005a). *Epistemic Luck*. Oxford University Press.

Pritchard, D. (2005b). Wittgenstein's *On Certainty* and Contemporary Anti-Scepticism. En D. Moyal-Sharrock y W. H. Brenner (eds.), *Readings of Wittgenstein's On Certainty* (pp. 189–224). Palgrave.

Pritchard, D. (2010). Relevant Alternatives, Perceptual Knowledge and Discrimination. *Noûs, 44*(2), 245–268.

Pritchard, D. (2011). Wittgenstein on Skepticism. En M. McGinn y O. Kuusela (eds.). *The Oxford Handbook of Wittgenstein* (pp. 521–547). Oxford University Press.

Pritchard, D. (2012). *Epistemological Disjunctivism*. Oxford University Press.

Pritchard, D. (2016). *Epistemic Angst: Radical Scepticism and the Groundlessness of Our Believing*. Princeton University Press.

Pryor, J. (2000). The Skeptic and the Dogmatist. *Noûs, 34*(4), 517–549.

Pryor, J. (2001). Highlights of Recent Epistemology, *British Journal for the Philosophy of Science, 52*(1), 95–124.

Pryor, J. (2014). There Is Immediate Justification. En M. Steup, J. Turri y E. Sosa (eds.), *Contemporary Debates in Epistemology* (pp. 202–221). (2ªed., Vol. 14). Wiley Blackwell.

Putnam, H. (1981). *Reason, Truth, and History*. Cambridge University Press.

Ranalli, C. (2021). Are There Heavyweight Perceptual Reasons? *International Journal for the Study of Skepticism, 11*(2), 93–118.

Rinard, S. (2021). Pragmatic Skepticism. *Philosophy and Phenomenological Research, 104*(2), 434–453.

Schellenberg, S. (2018). *The Unity of Perception: Content, Consciousness, Evidence*. Oxford University Press.

Schiffer, S. (1996). Contextualist Solutions to Scepticism. *Proceedings of the Aristotelian Society, 96*(1), 317–333.

Schönbaumsfeld, G. (2016). *The Illusion of Doubt*. Oxford University Press.

Sexto Empírico. *Esbozos Pirrónicos* (Trad. A. Gallego Cao y T. Muñoz Diego). Gredos.

Sherman, B. y G. Harman (2011). Knowledge and Assumptions. *Philosophical Studies, 156*(1), 131–140.

Sosa, E. (1999). How to Defeat Opposition to Moore. *Philosophical Perspectives, 13*, 141–153.

Sosa, E. (2002) Tracking, Competence, and Knowledge (pp. 264–287). En P. K. Moser (ed.), *The Oxford Handbook of Epistemology*. Oxford University Press.

Sosa, E. (2021). *Epistemic Explanations: A Theory of Telic Normativity, and What It Explains*. Oxford University Press.

Stalnaker, R. (1968). A Theory of Conditionals. En N. Rescher (Ed.), *Studies in Logical Theory* (pp. 98–112). Blackwell.

Stanley, J. (2005). *Knowledge and Practical Interests*. Oxford University Press.

Stine, G. C. (1976). Skepticism, Relevant Alternatives, and Deductive Closure. *Philosophical Studies, 29*(4), 249–261.

Strawson, P. F. (1985). *Scepticism and Naturalism: Some Varieties*. Columbia University Press.

Stroud, B. (1968). Transcendental Arguments. *Journal of Philosophy, 65*(9), 241–256.

Stroud, B. (1984). Scepticism and the Possibility of Knowledge. *Journal of Philosophy, 81*(10), 545–551. Reimpreso en B. Stroud (2000), *Understanding Human Knowledge* (pp. 1–8). Oxford University Press.

Stroud, B. (1991). *El escepticismo filosófico y su significación* (Trad. L. García Urriza). Fondo de Cultura Económica.

Vogel, J. (1987). Tracking, Closure, and Inductive Knowledge. En S. Luper-Foy (Ed.), *The Possibility of Knowledge: Nozick and His Critics* (pp. 197–215). Rowman & Littlefield.

Vogel, J. (1990). Are There Counterexamples to the Closure Principle? En M. Roth y G. Ross (Eds.), *Doubting: Contemporary Perspectives on Skepticism* (pp. 13–27). Kluwer.

Vogel, J. (1999). The New Relevant Alternatives Theory. *Philosophical Perspectives, 13*, 155–180.

Vogel, J. (2004). Skeptical Arguments. *Philosophical Issues, 14*, 426–455.

Wedgwood, R. (2013). A Priori Bootstrapping. En A. Casullo y J. C. Thurow (Eds.). *The A Priori in Philosophy*. Oxford University Press.

Williams, B. (1978). *Descartes: el proyecto de una investigación pura* (Trad. L. Benítez). IIFs-UNAM.

Williams, M. (2001). *Problems of Knowledge: A Critical Introduction to Epistemology*. Oxford University Press.

Williams, M. (2008). Hume's Skepticism. En J. Greco (ed.), *The Oxford Handbook of Skepticism* (pp. 80–107). Oxford University Press.

Williams, M. (2010). Descartes' Transformation of the Sceptical Tradition. En R. A. H. Bett (Ed.), *The Cambridge Companion to Ancient Scepticism* (pp. 288–313). Cambridge University Press.

Williams, M. (2012). Wright Against the Sceptics. En A. Coliva (Ed.), *Mind, Meaning, and Knowledge: Themes from the Philosophy of Crispin Wright* (pp. 352–375). Oxford University Press.

Williamson, T. (2000). *Knowledge and Its Limits*. Oxford University Press.

Williamson, T. (de próxima publicación). Justifications, Excuses, and Skeptical Scenarios. En J. Dutant y F. Dorsch (eds.), *The New Evil Demon Problem*. Oxford University Press.

Wittgenstein, L. (1969). *On Certainty* (Ed. G. E. M. Anscombe y G. H. von Wright). Blackwell.

Wright, C. (1985). Facts and Certainty. *Proceedings of the British Academy, 71*, 429–472.

Wright, C. (1991). Scepticism and Dreaming: Imploding the Demon. *Mind, 100*(1), 87–116.

Wright, C. (2004). Warrant for Nothing (and Foundations for Free)? *Proceedings of the Aristotelian Society Supplementary volumes, 78*(1), 167–212.

Wright, C. (2014). On Epistemic Entitlement (II): Welfare State Epistemology. En D. Dodd y E. Zardini (eds.), *Scepticism and Perceptual Justification* (pp. 213–248). Oxford University Press.

Yalçin, Ü (2004). Skeptical Arguments from Underdetermination. *Philosophical Studies, 68*(1), 1–34.

TERCERA SECCIÓN

ÉTICA ANALÍTICA

CAPÍTULO VIII

TRES TEORÍAS DE LA AGENCIA RESPONSABLE

Fernando Rudy Hiller[1]

RESUMEN

Los estudios sobre responsabilidad moral ocupan un lugar central dentro de la filosofía moral contemporánea anglófona o "analítica". Todos ellos parten, de una u otra manera, del artículo fundacional de Strawson "Libertad y resentimiento", según el cual la pregunta central para entender la responsabilidad moral de un agente no es si éste poseía libre albedrío al realizar una acción, sino si es merecedor de alguna de las "actitudes reactivas" (indignación, reproche, encomio, gratitud) a causa de ésta. Este es el punto de encuentro de la enorme mayoría de las teorías contemporáneas. Sin embargo, existe gran discrepancia entre ellas respecto de dos asuntos: 1)

1 Investigador del Instituto de Investigaciones Filosóficas de la UNAM. Ha publicado diversos artículos sobre filosofía de la acción y responsabilidad moral, entre otros temas.

cómo entender las actitudes reactivas; 2) qué capacidades agenciales son necesarias y suficientes para hacerse acreedor de ellas. En este artículo propondré una taxonomía tripartita de las teorías contemporáneas con base en las respuestas que ofrecen a cada una de estas cuestiones, a saber: teorías "de respuesta a razones", "conversacionales" y "de la calibración interpersonal". Concluiré el artículo bosquejando un argumento de por qué la mejor interpretación de Strawson la ofrecen las teorías de la calibración.

Palabras clave: Filosofía moral, teorías de la agencia responsable, teorías conversacionales, teorías de calibración interpersonal, Strawson.

ABSTRACT

Studies on moral responsibility occupy a central place within contemporary Anglophone or "analytic" moral philosophy. All of them, in one way or another, take as their point of departure P.F. Strawson's foundational article *Freedom and Resentment*, according to which the central question for understanding an agent's moral responsibility is not whether the agent possessed free will when performing an action, but whether they are deserving of any of the so-called "reactive attitudes" (indignation, blame, praise, gratitude) in response to it. This constitutes the common ground shared by the vast majority of contemporary theories. However, there is significant disagreement among them on two main issues: (1) how reactive attitudes should be understood; and (2) which agential capacities are necessary and sufficient for being a fitting target of such attitudes. In this article, I propose a tripartite taxonomy of contemporary theories based on the answers they offer to these questions—namely, "reasons-responsive" theories, "conversational" theories, and "interpersonal calibration" theories. I will conclude by outlining an argument as to why the latter provide the most compelling interpretation of Strawson's view.

Keywords: Moral philosophy, theories of responsible agency, conversational theories, interpersonal calibration theories, Strawson.

Los estudios sobre responsabilidad moral ocupan un lugar privilegiado dentro de la ética contemporánea, sobre todo en la tradición anglófona o "analítica". Si bien dicho campo de estudio engloba una gran variedad de temas, el más importante entre ellos es el de determinar cuáles son las capacidades psicológicas que debe poseer un agente para poder ser clasificado como moralmente responsable por sus actos. A pesar de la diversidad de teorías que sobre esta cuestión han surgido, la enorme mayoría de ellas toma como punto de partida el artículo fundacional de Peter Strawson "Libertad y resentimiento" (1962/1992)[2]. En dicho texto, Strawson argumenta que el punto crítico para entender la responsabilidad moral de un agente no es si éste poseía libre albedrío al realizar su acción, sino si, de acuerdo con nuestras prácticas ordinarias, es merecedor de alguna de las "actitudes reactivas" (indignación, resentimiento, admiración, gratitud, etc.) a causa de aquella.

En su momento, el artículo de Strawson provocó una auténtica revolución en los estudios sobre responsabilidad, que hasta ese momento tenían como preocupación principal la de zanjar la cuestión de si la responsabilidad moral era o no compatible con el determinismo[3]. La aportación de Strawson consistió en mostrar que lo verdaderamente relevante para los estudios sobre responsabilidad no radica en la discusión de una tesis metafísica, sino en entender las prácticas de atribución de responsabilidad tal como se ejercen en el día a día. Si enfocamos en ellas nuestra atención, sostenía Strawson, veremos que la verdad o falsedad del determinismo es por completo irrelevante, pues dichas prácticas se enfocan únicamente en evaluar la buena o mala voluntad que los agentes

2 Otros dos textos clásicos para el análisis de la responsabilidad en la tradición analítica son: "A plea for excuses" (Austin 1961) y "Responsibility and retribution" (Hart 1968). A pesar de su indudable importancia, ninguno de ellos sigue ejerciendo sobre la discusión contemporánea una influencia comparable al texto de Strawson. Agradezco a un revisor anónimo haberme recordado la necesidad de mencionar ambos artículos.

3 El artículo de Harry Frankfurt, "Alternate possibilities and moral responsibility" (1969/1988), es la fuente clásica en la filosofía analítica contemporánea para la tesis de que la responsabilidad moral no requiere posibilidades alternativas y, por lo tanto, que ésta es plenamente compatible con el determinismo.

demuestran en sus actos y las posibles explicaciones que pueden esgrimir para excusar su mal comportamiento. Algunas de esas explicaciones –como haber sido coaccionado o ser ignorante de las consecuencias de una acción– muestran que un agente que *en general* es responsable no lo es respecto de una acción en particular; sin embargo, otras –como padecer demencia o una compulsión suficientemente grave– indican que el agente en cuestión no es responsable *en absoluto*. Si esto es así, entonces, a partir de un análisis de las prácticas de responsabilidad mismas, podemos dilucidar cuáles son las capacidades psicológicas que debe poseer un agente para poder participar en ellas.

Este es el punto de encuentro de casi todas las teorías contemporáneas sobre la agencia responsable. Sus discrepancias se deben al desacuerdo respecto de dos cuestiones centrales e íntimamente relacionadas entre sí: 1) cuál es la función específica de las actitudes reactivas dentro de las prácticas de responsabilidad moral; 2) qué capacidades agenciales son necesarias para hacerse acreedor a ellas. En el presente texto propondré una taxonomía tripartita de las teorías contemporáneas con base en las respuestas que ofrecen a ambas cuestiones y, además, argumentaré a favor de la superioridad de una de ellas. Así, las teorías que llamaré "punitivistas" sostienen que las actitudes reactivas negativas, como resentimiento e indignación, son una especie de sanciones informales con las que se busca castigar al ofensor o, al menos, señalar que es merecedor de castigo por una acción moralmente incorrecta, y, en consonancia con esto, afirman que las capacidades agenciales que debe poseer un agente para ser responsable son aquellas que le habrían permitido comportarse correctamente –y así evitar el castigo– siguiendo las razones morales pertinentes. En segundo lugar, las teorías que llamaré "conversacionales" o "comunicativas" mantienen que las actitudes reactivas son herramientas con las cuales se busca establecer una conversación moral con el ofensor a fin de que éste reconozca su falta y se disculpe, y, en consecuencia, afirman que las capacidades agenciales requeridas son aquellas que le permiten comprender, tras realizar una acción incorrecta y haber sido

censurado por otros, las razones morales en contra de su acción y los cursos correctivos que debe seguir para restaurar las relaciones que han sido dañadas. Finalmente, las teorías que llamaré "de la calibración interpersonal" aseveran que las actitudes reactivas son los medios con los que contamos para calibrar o ajustar nuestras relaciones interpersonales, en el sentido de dar a las personas el trato que merecen en virtud de la buena o mala voluntad que muestran hacia nosotros o hacia aquellos con quienes simpatizamos. En esta tercera familia de teorías, las capacidades constitutivas de un agente responsable son aquellas que le permiten expresar "significado moral" a través de sus acciones.

En lo que sigue, y a la par que expongo los rudimentos de estas tres teorías, argumentaré que la mejor interpretación de las prácticas ordinarias de atribución de responsabilidad la ofrecen las teorías de la calibración interpersonal. De esta manera, la contribución del presente texto reside, por un lado, en clarificar el debate contemporáneo sobre la agencia responsable y, por otro, en ofrecer una serie de argumentos encaminados a resolverlo.

1. ¿Qué es la responsabilidad moral?

Antes de dar paso a la descripción del panorama contemporáneo acerca de la psicología moral de la agencia responsable, conviene dejar sentado un concepto mínimo de responsabilidad moral, pues, de lo contrario, podría suponerse que quizá el desacuerdo entre las distintas teorías pudiera deberse a que en el fondo están discutiendo sobre temas distintos. Propongo, entonces, la siguiente definición:

> *Responsabilidad moral.* La responsabilidad moral engloba las evaluaciones y reacciones que son apropiadas en respuesta a qué tan bien o mal la conducta de las personas se ajusta a ciertas normas y expectativas interpersonales, en particular las relacionadas con la moralidad.

Como fue señalado antes, a partir de Strawson hay virtual unanimidad en que las evaluaciones y reacciones características de la responsabilidad moral –es decir, de las atribuciones de responsabi-

lidad– se manifiestan a través de lo que el mismo Strawson bautizó como "actitudes reactivas". Son éstas un conglomerado de actitudes con componentes cognitivos (juicios y creencias), afectivos (sentimientos y emociones) y conductuales, las cuales merecen el calificativo de "reactivas" en virtud de que se despliegan en reacción a la percepción de buena o mala voluntad manifestada por las personas en sus acciones. La buena o mala voluntad es, a su vez, una función de si, y en qué medida, la conducta de aquellas se ajusta a las normas y expectativas propias de la moralidad.

El campo de las actitudes reactivas admite dos biparticiones: por un lado, entre las actitudes reactivas positivas y negativas y, por otro, entre las actitudes reactivas personales e impersonales. Examinémoslas por turnos. La primera bipartición tiene que ver con la valencia que caracteriza a una actitud reactiva, la cual es positiva cuando se busca encomiar (*praise*) a un agente o negativa cuando se busca censurarlo (*blame*). Como puede suponerse, una actitud reactiva positiva –tal como gratitud o admiración– ocurre en respuesta a la buena voluntad que manifiesta un agente al llevar a cabo una acción moralmente correcta o, más precisamente, supererogatoria. Esto es, para que sintamos gratitud o admiración hacia alguien, normalmente no basta con que esa persona se limite a cumplir con su deber, sino que es necesario que exceda aquello que el deber estrictamente le exige. Por ejemplo, sería extraño experimentar gratitud o admiración hacia un automovilista que detuviera su vehículo cuando la luz del semáforo cambia a rojo, mientras que no sería raro hacerlo si esa misma persona se desvía considerablemente de su camino para atender una situación de emergencia como llevar a un herido al hospital. Por otro lado, una actitud reactiva negativa –cuyos ejemplos paradigmáticos son el resentimiento y la indignación– se da ante la percepción de indiferencia o, todavía peor, de mala voluntad de parte de un agente cuando su conducta cae por debajo de lo moralmente esperable. Por ejemplo, supongamos que usted, quien va cargado con demasiados libros, deja caer al piso varios de ellos y yo, que voy con las manos desocupadas, paso de largo sin detenerme a ayudarlo.

En esta situación, su resentimiento hacia mí estaría más que justificado, ya que, si bien al seguir con mi camino no manifiesto mala voluntad hacia usted, por supuesto que manifiesto indiferencia respecto de sus intereses, y esta indiferencia bien puede dar lugar –y usualmente lo hace– a una actitud reactiva negativa de su parte. Si esto es así, entonces expresiones ya no de indiferencia sino de declarada mala voluntad (por ejemplo, si yo tirara a propósito sus libros de un manotazo), con más razón dan lugar a manifestaciones todavía más acaloradas de resentimiento e indignación.

Antes de pasar a la segunda bipartición, será provechoso hacer explícitos los tres componentes de las actitudes reactivas señalados anteriormente. En primer lugar, estas actitudes poseen un componente cognitivo, ya que involucran creencias acerca del comportamiento de las personas (quién hizo qué) y juicios acerca de cómo se compadece aquél con las normas morales aplicables en cada caso. En segundo lugar, poseen un componente afectivo, pues usualmente los juicios y creencias mencionados hace un momento carecen de la sobriedad típica de, digamos, los juicios matemáticos, sino que van acompañados de emociones que varían, tanto en contenido como en intensidad, en razón de qué tan grave es la falta que se imputa al agente o qué tan meritorio es el favor que se recibió de él. En tercer lugar, las actitudes reactivas están conectadas con diversas disposiciones conductuales, las cuales, de nueva cuenta, varían dependiendo de si la acción que dio lugar a ellas fue moralmente buena o moralmente mala y de qué tan grave o meritorio fue la falta o el favor recibidos. Así, por ejemplo, cuando experimentamos gratitud hacia alguien que nos ha beneficiado, nos vemos motivados a expresar nuestro agradecimiento y, de ser posible, a retribuir el favor, mientras que, por el contrario, cuando experimentamos resentimiento dirigido a quien nos ha dañado, tenemos la disposición a expresar nuestro malestar y a tomar medidas encaminadas a modificar (o incluso terminar) nuestras relaciones con el ofensor.

Pasemos a la segunda bipartición en la clasificación de las actitudes reactivas. En “Libertad y resentimiento”, Strawson hizo

notar que éstas se dividen, además de en positivas y negativas, en personales e impersonales. Una actitud reactiva personal surge en respuesta a una ofensa o a un beneficio cuyo recipiendario es la persona misma que experimenta la actitud en cuestión; en cambio, una actitud reactiva impersonal tiene lugar cuando el blanco de la ofensa o el beneficio es una tercera persona con quien simpatiza aquel que experimenta la actitud. Así, tomando como ejemplo las cuatro actitudes reactivas paradigmáticas –resentimiento, indignación, gratitud y admiración–, diremos que tanto el resentimiento y la gratitud son actitudes personales, mientras que la indignación y la admiración son impersonales. Ahora bien, a pesar de que Strawson sostuvo que, al ser desinteresadas, solo las actitudes reactivas impersonales son propiamente actitudes morales y, en esa medida, solo ellas expresan atribuciones de responsabilidad, los teóricos contemporáneos asumen de modo unánime que cualquier actitud reactiva, sea personal o impersonal, es un vehículo apropiado para responsabilizar a los agentes por su conducta.

Lo expuesto hasta aquí sobre las actitudes reactivas constituye el terreno compartido por los teóricos de la responsabilidad moral. Como veremos más adelante, las discrepancias comienzan una vez que aquéllos se abocan a contestar la pregunta: ¿qué función cumplen las actitudes reactivas más allá de ser los vehículos para transmitir censura y encomio? Unos sostienen que estas actitudes (por el lado negativo) tienen por objetivo castigar o, cuando menos, señalizar el merecimiento de castigo de quienes han violado una norma moral; otros afirman que el propósito de las mismas es iniciar una conversación moral con el ofensor, y aun otros aseveran que su función es regular las relaciones interpersonales con quienes han demostrado falta de compromiso con la moralidad. Sin embargo, antes de proceder a analizar cada una de estas posturas, conviene decir unas palabras preliminares acerca del otro ángulo del desacuerdo, a saber, el que concierne a la psicología moral de la agencia responsable.

2. Agencia y responsabilidad

Como expliqué al inicio, las divergencias entre los teóricos de la responsabilidad moral recaen tanto en su concepción de la función específica que tienen las actitudes reactivas –más allá de ser los vehículos de las atribuciones de censura o encomio– dentro de las prácticas de responsabilización como en su descripción de las capacidades psicológicas que debe poseer un agente para hacerse acreedor a ellas[4]. Hemos visto ya algo sobre el primer punto, así que conviene ahora abordar brevemente el segundo.

No todo agente es un agente responsable. Los animales y los niños pequeños son agentes, mas no son moralmente responsables por sus actos. Por definición, un agente es un organismo capaz de actuar, es decir, de efectuar acciones, o sea movimientos controlados con los cuales se desplaza a través de su medio ambiente, realiza modificaciones en él e interactúa con los objetos y con otros organismos. Para ello, un agente precisa de diversas capacidades motrices y, por el lado psicológico, de capacidades perceptivas e intelectivas. Por supuesto, "intelectivo" no debe entenderse aquí como la facultad de razonamiento abstracto a través de conceptos (de la que carecen tanto los animales como los niños muy pequeños), sino como las habilidades de adaptar inteligentemente el propio comportamiento en función de los cambios en el ambiente y de aprender nuevas formas de actuar en respuesta a ciertos estímulos.

¿Qué se requiere para transformar un mero agente en un agente responsable? Dado que el consenso es que solo un agente humano adulto es genuinamente responsable de sus actos, la respuesta inmediata que viene a la mente es "razón". Esta respuesta, sin ser equivocada, resulta de poca ayuda, ya que el término "razón", referido a la facultad distintiva de los seres humanos adultos, es demasiado general como para ayudarnos a dilucidar las cualidades que

4 Recuérdese que, en mi interpretación del debate contemporáneo, ambas cuestiones están íntimamente relacionadas, pues qué capacidades psicológicas se postulen como necesarias para la agencia responsable dependerá de qué función se asigne a las actitudes reactivas dentro de las prácticas de responsabilidad.

debe poseer un agente para ser responsable. Es necesario, pues, ser más precisos. Para que un agente sea moralmente responsable de sus actos, debe, en primer lugar, contar con la capacidad de comprender los requisitos morales básicos, por ejemplo, los que llaman a evitar el daño gratuito a la persona de otros, a prestar ayuda y a reciprocar la ayuda recibida, a respetar la propiedad ajena, a ser veraz y honesto en el trato común, etc. Desde luego, estos requisitos básicos adoptan multitud de formulaciones más precisas en distintos contextos históricos y sociales, y un agente responsable debe ser mínimamente capaz de comprender qué es lo que la moralidad imperante en su propio contexto le exige en diferentes circunstancias. Pero la mera comprensión no basta, así que, en segundo lugar, un agente responsable debe poder regular su comportamiento en atención a lo que la moralidad exige de él. Esta capacidad es de tipo volitivo, ya que involucra autocontrol, es decir, la habilidad de modificar los impulsos a actuar con el fin de adaptarlos a las exigencias de determinadas normas, en este caso, normas morales.

En suma, pues, un agente moralmente responsable es aquel que es capaz de entender lo que la moralidad exige de él en una diversidad de circunstancias y de actuar en consecuencia (no siempre, desde luego, pero sí en una mayoría de los casos). Alguien podría preguntar en este punto: ¿y qué hay del libre albedrío? En efecto, hasta antes de la publicación de "Libertad y resentimiento", había consenso en que la capacidad distintiva de los agentes responsables era precisamente el libre albedrío, entendido como la facultad de "haber obrado de otro modo" a como el agente lo hizo en realidad. Como puede advertirse, en esa comprensión de la agencia responsable, el famoso problema de la libertad y el determinismo resultaba central, pues, para averiguar si los seres humanos poseían o no libre albedrío, era necesario establecer, primero, si las acciones de los agentes humanos están regidas por el determinismo causal que gobierna todos los fenómenos y, segundo, en caso de una respuesta afirmativa, si el libre albedrío es o no compatible con el determinismo. No obstante, en palabras de David Shoemaker (2020, p. 212), el ensayo de Strawson trajo consigo el salutífero efecto de

"librarnos de la búsqueda de la libertad", es decir, de reorientar la atención de los filósofos para que, en vez de ocuparse de una cuestión metafísica como el problema de la libertad y el determinismo, se enfocara en una serie de preguntas éticas concernientes a las prácticas ordinarias de atribución de responsabilidad[5]. Como quedó dicho al inicio, el presente ensayo se inscribe dentro de la tradición strawsoniana y, en consecuencia, el problema metafísico será dejado completamente de lado.

Ahora bien, lo dicho hasta el momento acerca de la agencia responsable es, al igual que lo concerniente a las características generales de las actitudes reactivas, terreno común entre los distintos teóricos contemporáneos. ¿Dónde reside, pues, el desacuerdo acerca de la psicología moral de la agencia responsable que constituye el principal interés de este ensayo? Para localizar dicho desacuerdo, es preciso, tal como hicimos con las actitudes reactivas, pasar de una descripción general de las capacidades propias de la agencia responsable a una descripción específica. La pregunta que divide a los diferentes teóricos es, entonces, la siguiente: más allá de las capacidades de comprender en general los requisitos morales y de ajustar el propio comportamiento a ellos, ¿de qué debe ser capaz un agente *al momento de la acción* para ser responsabilizado por esa acción en particular? En el debate contemporáneo se registran tres clases de réplicas a esta pregunta, las cuales corresponden a las tres posiciones descritas al inicio.

Los teóricos que defienden una postura de tipo punitivista afirman que, *antes* de que la acción (cuando ésta es incorrecta) ocurra, el agente debe haber sido capaz de reconocer y reaccionar al menos a una razón moral en contra de aquella, pues solo en este caso es posible sostener que dicho agente pudo haber actuado correctamente, lo cual, según dichas teorías, es necesario para que

5 Claro está que ello no quiere decir que ya nadie se dedique a estudiar el problema de la libertad y el determinismo. Bien al contrario, éste es un campo tan activo como el de los estudios sobre responsabilidad; la diferencia radica en que, tras la publicación de "Libertad y determinismo", lo que antes constituía un solo campo de estudio se ha dividido en dos. Véase mi texto Rudy Hiller (2021) para más detalles al respecto.

sea justo responsabilizarlo (y quizá castigarlo) por haber llevado a cabo esa acción. Por otro lado, los teóricos que se adhieren a una postura conversacional niegan que la susceptibilidad a razones *ex ante* (antes de efectuar la acción) sea necesaria para ser llamado a cuentas, pero afirman que la susceptibilidad *ex post* sí lo es, con lo cual quieren decir que, para ser responsabilizado por un acto en particular, y tras haber realizado éste, el agente debe ser capaz de reconocer –quizá a consecuencia de ser objeto de reproches– que *había* al menos una razón moral para no haber actuado como lo hizo y, además, deber ser capaz de responder adecuadamente a esta toma de conciencia mediante el establecimiento de una "conversación moral" encaminada a restaurar las relaciones que han sido dañadas entre él y la parte ofendida[6]. Finalmente, los teóricos de la calibración interpersonal se caracterizan por disentir tanto de unos como de otros, pues sostienen que la capacidad *local* (esto es, presente al momento de la acción o inmediatamente después de ocurrida ésta) de susceptibilidad a razones, ya sea *ex ante* o *ex post*, no es necesaria para ser llamado a cuentas por una acción específica, lo que significa que, de acuerdo con ellos, no es necesario que el agente haya sido capaz de actuar correctamente ni tampoco que pueda entablar conversación moral alguna para que sea legítimo responsabilizarlo por su conducta. En su postura, por el contrario, basta con que el agente exprese "significado moral" a través de sus acciones para que las actitudes reactivas le sean apropiadamente dirigidas (más adelante se buscará dilucidar en qué consiste esta clase de significado).

3. Teorías punitivistas

Contamos ahora con los elementos necesarios para comprender el debate contemporáneo sobre la agencia responsable. Comenzaremos por discutir las teorías punitivistas.

6 Presenté inicialmente esta distinción entre sensibilidad a razones *ex ante* y *ex post* en Rudy Hiller (2022).

La idea de que existe una conexión íntima entre la responsabilidad moral y el merecimiento de castigo es muy antigua e intuitiva. Tanto en el pensamiento religioso como en el pensamiento jurídico, ser responsable por una acción incorrecta implica estar sujeto a ciertas sanciones de parte de la autoridad competente (ya sea la divinidad o el Estado) y, en la medida en que la responsabilidad moral se considera como el tipo más básico de responsabilidad, es natural pensar que ésta tiene que ver fundamentalmente con las consecuencias (buenas o malas) a las que un agente se hace acreedor en razón de su comportamiento (Hieronymi, 2019). La conexión entre responsabilidad moral y castigo explica también por qué, hasta antes de "Libertad y resentimiento", se daba por sentado que el libre albedrío constituía una condición necesaria de la agencia responsable, pues es parte del sentido moral ordinario que es justo o apropiado castigar a alguien por sus actos solo si la persona pudo haber evitado la acción por la cual es castigada, es decir, solo si pudo haber actuado de un modo distinto a como de hecho lo hizo.

Sin embargo, como expliqué más arriba, el artículo de Strawson mostró que era posible teorizar sobre la responsabilidad moral haciendo a un lado la problemática metafísica sobre el libre albedrío y concentrándose exclusivamente en los aspectos éticos de las prácticas ordinarias, en particular en las condiciones bajo las cuales las actitudes reactivas se consideran apropiadas. ¿Significó ello que Strawson mismo desligó también la responsabilidad moral del merecimiento de castigo? Por extraño que parezca, la respuesta a esta pregunta es negativa, pues Strawson sostuvo que las actitudes reactivas forman parte de un continuo de prácticas morales basadas en "la imposición de sufrimiento al ofensor que es una parte esencial del castigo" (1962/1992, p. 33, traducción modificada). Hacia el final de esta sección argumentaré que Strawson se equivoca en este punto, pues, como se verá, las actitudes reactivas no son asimilables a la imposición de sanciones y castigos.

No obstante, es preciso reconocer de inmediato que este es un punto altamente controvertido, ya que muchos teóricos contem-

poráneos concuerdan con Strawson en que las actitudes reactivas no solo forman parte del mismo conjunto de prácticas basadas en las sanciones y los castigos, sino que ellas mismas son una especie (bastante atenuada) de éstos. Por ejemplo, Jay Wallace (1994), al describir las actitudes reactivas, afirma que "la conducta punitiva pertenece al síndrome de respuestas al cual predisponen las actitudes reactivas. … Aprendemos los conceptos de indignación, resentimiento y culpa en parte al aprender cómo se conectan con la conducta punitiva" (p. 68). Por su parte, Gary Watson (1996/2004, p. 275) sostiene que ser llamado a cuentas en el ámbito moral, no menos que en el legal, involucra ser objeto de sanciones, y añade que "las sanciones involucradas en ser responsabilizado moralmente consisten en ser el blanco de estas actitudes [reactivas]" (p. 278). Victoria McGeer (2013) concuerda. Así mismo, Gideon Rosen (2015: 83) postula que las actitudes reactivas negativas contienen el deseo de que el ofensor sufra algún perjuicio en pago por su acción y la creencia de que dicho perjuicio es merecido. Igualmente, Dana Nelkin (2013; 2016) argumenta que existe una íntima conexión entre ser el blanco apropiado de las actitudes reactivas negativas y merecer sanciones. Y Michael McKenna (2019), buscando especificar en qué consiste el aspecto punitivo de las actitudes reactivas mismas, propone que éstas, al ser expresadas, infligen tres clases de "daños": impiden que el ofensor interactúe con otros de modo normal; fuerzan a éste a dar explicaciones por su conducta y afectan su estabilidad emocional.

Así pues, como puede observarse en esta pequeña pero representativa muestra de filósofos contemporáneos que concuerdan con Strawson en la supuesta conexión entre las actitudes reactivas y el castigo, dichas actitudes son interpretadas o bien como sanciones en sí mismas, o bien como disposiciones a sancionar a los ofensores, o bien como actitudes que señalizan el merecimiento de sanciones. La extendida creencia en esta conexión justifica la afirmación de Watson (1996/2004, p. 280) de que es común interpretar la responsabilidad moral como una "práctica de tipo legal" (*legal-like practice*) que sirve los fines de la justicia retributiva. Si

esto es así, continúa Watson (p. 276), entonces para determinar si un agente es moralmente responsable por sus actos incorrectos, no menos que para establecer si es legal o criminalmente responsable por ellos, resulta esencial saber si, al momento de la acción, contaba con una oportunidad razonable de haberse comportado correctamente y haber evitado así la sanción correspondiente (que, en el caso de la responsabilidad moral, puede no ir más allá de ser el blanco de resentimiento e indignación).

Esta condición –que el mismo Watson denomina "evitabilidad"– sobre la censura moral es, bajo diversas formulaciones, ampliamente aceptada por los teóricos punitivistas. La formulación más explícita la encontramos en un artículo escrito conjuntamente por David Brink y Dana Nelkin (2013), quienes afirman que "la censura [*blame*] y el castigo presuponen que el agente tuvo una oportunidad propicia [*fair*] para haber evitado actuar incorrectamente" (285). ¿Qué es lo que hace que una oportunidad sea propicia en este sentido? De acuerdo con Brink y Nelkin, ello depende de dos elementos denominados "competencia normativa" y "control situacional". El primero involucra las capacidades para detectar y responder, al momento de la acción, a las razones morales pertinentes en contra de ella (si es incorrecta), en tanto que el segundo mienta la ausencia de factores en las circunstancias del agente –como por ejemplo coacción– que impidan el ejercicio de aquellas capacidades[7].

A pesar de parecer autoevidente, el requisito de competencia normativa (o, en mis términos, de sensibilidad a razones *ex ante*) ha sido negado por los teóricos de la calibración interpersonal, quienes, como veremos más adelante, sostienen que para responsabilizar a un agente por una acción incorrecta no es necesario que

7 Es importante aclarar que los teóricos punitivistas que aceptan, de uno u otro modo, el requisito de evitabilidad sobre la censura moral son en su mayoría compatibilistas, es decir, piensan que no es necesario poseer libre albedrío "libertario" –el tipo de libertad que es incompatible con el determinismo– para ser responsable. Véase por ejemplo Wolf (1990), Wallace (1994), Fischer y Ravizza (1998), Pettit (2002), M. Smith (2004), Nelkin (2011), Brink y Nelkin (2013), Vargas (2013), McKenna (2013) y McGeer y Pettit (2015).

aquél haya sido capaz, al momento de la acción, de detectar y responder adecuadamente a las razones morales pertinentes, pues, incluso si este no es el caso, las actitudes reactivas siguen siendo aplicables, siempre y cuando no exista una excusa –como coerción o demencia– suficientemente convincente. Dado este desacuerdo, los teóricos punitivistas precisan de un argumento a favor del requisito de competencia normativa. Dicho argumento es el "argumento de la justicia" (*fairness argument*), el cual ha sido articulado por Watson (1996/2004), Wallace (1994) y Brink y Nelkin (2013). El argumento puede resumirse así:

Responsabilizar a las personas por acciones incorrectas implica experimentar actitudes reactivas negativas, las cuales son una especie de sanciones informales que nos disponen, además, hacia otras actividades más abiertamente punitivas. Pero, dado que adoptar dichas actitudes expone a los blancos de éstas a sufrir cierta clase de daños, su adopción está sujeta a evaluación a la luz de normas morales y, en particular, normas de justicia (*fairness*) que determinan cuándo está permitido someter a las personas a daños de este tipo (Wallace, 1994, p. 94; Watson, 1996/2004, p. 273). Una norma de justicia básica es que un agente sujeto a determinadas obligaciones y a la amenaza de sanciones en caso de que incumpla con ellas debe poseer las capacidades necesarias para atenerse a las primeras y así evitar las segundas. Sin embargo, puesto que las obligaciones morales están respaldadas por razones que las justifican y que ofrecen al mismo tiempo motivos para que los agentes cumplan con ellas, se sigue que quien está sujeto a esas obligaciones debe ser capaz de comprender y adoptar las razones morales pertinentes. Y, a la inversa, es injusto exigir el cumplimiento de una obligación moral específica a un agente si éste es incapaz de entender las razones que la respaldan y actuar en consonancia con ellas. Por lo tanto, para que sea justo llamar a cuentas a un agente por la violación de obligaciones morales, dicho agente debe poseer la doble capacidad de comprender razones morales y de gobernar su conducta a la luz de ellas (Wallace, 1994, p. 162).

Este argumento ofrece una justificación moral para el requisito de competencia normativa que forma parte de la condición de evitabilidad: es moralmente inadecuado (al ser injusto) dirigir actitudes reactivas y otras formas más enérgicas de sanción hacia un agente incapaz de comprender las razones morales particulares en contra de su acción y de actuar en consecuencia. Ahora bien, dicho requisito, en apariencia irreprochable, es en realidad más exigente de lo que parece a primera vista. Para entender por qué, es necesario volver a la distinción entre capacidades generales y capacidades locales mencionada de paso anteriormente. El término "capacidad general" refiere a la habilidad que tiene una persona para llevar a cabo cierta actividad sin hacer mención de si puede o no ejercerla en el momento presente. Por ejemplo, aun en este momento, mientras estoy sentado escribiendo, poseo la capacidad general de andar en bicicleta, puesto que, de tener una a la mano, podría, si así lo quisiera, subirme en ella y pedalear. Por otro lado, el término "capacidad local" alude a la facultad de ejercer en el momento presente una capacidad general que ya se posee.

¿Qué tipo de capacidad es la capacidad de detectar y responder a razones morales que, de acuerdo con los teóricos punitivistas, es necesaria para la responsabilidad moral? Los teóricos "de la vieja guardia" como Wolf (1990), Wallace (1994) y Fischer y Ravizza (1998) suponían que bastaba con poseer estas capacidades de manera general para ser responsable por una acción en particular; sin embargo, Manuel Vargas (2013), un miembro más joven de esta corriente teórica, apoyado en evidencia empírica proveniente de la psicología social situacionista, argumentó convincentemente que, dado que las capacidades de detección y respuesta a razones morales son sumamente variables según las circunstancias y el tipo de razones en juego, no se justifica la confianza en que la posesión general de aquéllas basta para hacer de una persona un agente responsable en un contexto determinado. En contraste, Vargas propuso un modelo "circunstancialista", según el cual un agente es responsable por sus acciones en ciertas circunstancias y no en otras dependiendo de si posee o no en un grado suficiente las capacida-

des locales de detección y respuesta a razones morales específicas. Para ayudar a determinar qué cuenta como "suficiente" a este respecto, Vargas formuló el siguiente test: en relación con una acción incorrecta en particular y en un contexto determinado, un agente exhibe la sensibilidad a razones suficientes para ser llamado a cuentas si, *en una proporción adecuada de contextos deliberativamente similares*, detecta y responde a las razones morales pertinentes (Vargas, 2013, pp. 217-22).

Como ilustración del test, considérese el caso de un automovilista que no respeta la luz roja del semáforo. En ausencia de una excusa convincente (trasladaba a un herido al hospital, digamos), experimentaremos cierto grado de indignación ante su acto, lo cual, como hemos visto, es una forma de responsabilizarlo por él. De acuerdo con el test de Vargas, nuestra indignación estará justificada solo si el conductor, en una proporción adecuada de contextos deliberativamente similares, en efecto detiene su auto en respuesta –quizá implícita y automática– a las razones morales pertinentes (como por ejemplo que debe respetar la seguridad de los peatones y de otros conductores), pues ello proporcionaría evidencia de que en este caso específico en que ignoró el semáforo sí era capaz de haber actuado correctamente. Desde luego, en la vida real no es usual que contemos con esta evidencia, y por ello Vargas acepta que su test no da respuestas contundentes para responsabilizar a este o a aquel agente en particular; no obstante, sostiene que su utilidad radica en establecer cómo debe entenderse la capacidad distintiva de los agentes responsables (p. 222).

Sin embargo, el problema más importante para la teoría circunstancialista de Vargas no radica en el hecho de que no resulta fácil determinar si un agente en específico satisface o no aquel test, sino en que nuestras prácticas ordinarias de responsabilidad basadas en las actitudes reactivas no parecen presuponer la capacidad local *ex ante* de sensibilidad a razones. El siguiente caso es, a mi parecer, decisivo. Harvey Weinstein, el ex productor de Hollywood encarcelado por violación, es sin duda uno de los paradigmas mundiales de un hombre poderoso que se aprovecha de su poder para

ultrajar a mujeres que trabajan para él y, en cuanto tal, es unánimemente repudiado. Si alguien en el mundo fue objeto de actitudes reactivas negativas por las fechas en que su caso salió a la luz, ese alguien fue Weinstein. Ahora bien, ¿poseía Weinstein, al momento de cometer la serie de crímenes por los que fue condenado, las capacidades de reconocer y adoptar las razones morales en contra de sus acciones? Sin duda poseía la capacidad *general* de entender y seguir requerimientos morales, pues no se comportaba de modo similar en todos los ámbitos de su vida; pero, como Vargas señalaría correctamente, la pregunta importante para los teóricos punitivistas no es si Weinstein era, en general, un agente moralmente competente, sino, más bien, si lo era en relación con las consideraciones morales pertinentes que condenaban su comportamiento hacia las mujeres.

No obstante, puesto que su conducta en este ámbito a lo largo de su vida revela más allá de toda duda que, en una proporción adecuada de contextos deliberativamente similares, Weinstein no respondía adecuadamente a las razones morales pertinentes en contra de acciones de esta clase, entonces parece que la conclusión que se nos impone, al menos según el test de Vargas, es que no poseía las capacidades locales de detección y respuesta a dichas razones y, consecuentemente, no era moralmente responsable por sus crímenes (lo cual no implica necesariamente que tampoco lo fuera en el ámbito jurídico). Ello a su vez significaría que las actitudes reactivas negativas dirigidas contra Weinstein, tanto por sus víctimas como por los millones de personas que han simpatizado con ellas, no están justificadas.

Es importante tener en claro que el caso de Weinstein, con ser muy aparatoso en su despliegue de maldad, no es en realidad una rareza. En la vida cotidiana, y en relación con faltas mucho menores, todos nosotros manifestamos comportamientos moralmente inadecuados que nos hacen objeto de actitudes reactivas, pero respecto de los cuales es plausible pensar que nos vemos afligidos por una especie de "ceguera moral local" comparable, si no en grado, al menos sí en tipo a la de Weinstein. Así pues, respecto de estas fal-

tas relativamente menores, es posible plantear la misma pregunta que surgió en el caso de aquél: ¿es esa clase de ceguera moral un impedimento para dirigir al ofensor actitudes reactivas plenamente justificadas? Confío en que el sentido moral ordinario del lector lo inclinará a dar una respuesta negativa.

Antes de concluir esta sección, quisiera enunciar claramente cuál es la conclusión que deseo extraer de las consideraciones precedentes. El punto no es que la norma moral a la que se apela en el argumento de la justicia esbozado arriba –según la cual es injusto sancionar a un agente cuando carece de las capacidades indispensables para evitar la sanción– sea incorrecta; más bien, lo que enseña el ejemplo de Weinstein es que las actitudes reactivas no son sanciones, puesto que, a diferencia de éstas, siguen siendo plenamente apropiadas aun cuando el agente a quien van dirigidas carece de la capacidad local de sensibilidad a razones que le habría permitido evitar la acción incorrecta que dio lugar a ellas. Y en la medida en que, siguiendo a Strawson, mantenemos que dichas actitudes son la característica distintiva de las atribuciones de responsabilidad, se sigue también que esa sensibilidad (entendida en sentido *ex ante*) no es aquello que en realidad distingue a un agente responsable de uno que no lo es.

4. Teorías conversacionales

Desde mi perspectiva, el ejemplo de Weinstein basta para mostrar concluyentemente que, mientras nos adhiramos al marco teórico strawsoniano, no debemos suponer que la sensibilidad a razones *ex ante* –aquella que satisfaría el requisito de evitabilidad– es una condición necesaria de la agencia responsable. Sin embargo, ello no basta para descartar sin más la sensibilidad a razones, pues, como se recordará, es posible conceptualizar ésta de manera *ex post*, esto es, como la capacidad de entender, tras haber realizado una acción incorrecta y haber recibido la crítica de otros a través de las actitudes reactivas, por qué dicha acción era inapropiada y también qué acciones correctivas es necesario emprender para restablecer las relaciones con aquellos a quienes se ha ofendido (lo cual incluye, por

supuesto, la capacidad de efectivamente llevar a cabo esas acciones correctivas). Esta manera de conceptualizar la sensibilidad a razones es la base de las teorías llamadas "conversacionales" o "comunicativas", según las cuales las actitudes reactivas son, en palabras del creador de esta línea teórica, "formas incipientes de comunicación" (Watson, 1987/2004, p. 230)[8].

Watson llega a esta conclusión tras investigar bajo qué condiciones, de acuerdo con la teoría de Strawson, una persona no es apta para ser responsabilizada. Según Strawson, condiciones como psicopatía son incompatibles con las actitudes reactivas negativas debido a que vuelven inaplicable la exigencia moral básica de mostrar cierto grado de consideración o buena voluntad hacia los demás. Watson sugiere que esto es así porque dichas actitudes, que incorporan la expresión de la exigencia básica, presuponen que aquel a quien van dirigidas es capaz de comprender el mensaje que buscan transmitir, el cual incluye tanto aquella exigencia como el resentimiento o indignación que suscitó el haberla pasado por alto. En razón de lo anterior, Watson concluye que la condición esencial para poder llamar a alguien a cuentas es que posea la capacidad de entablar una conversación moral[9].

Para ser capaz de comprender el mensaje transmitido por las actitudes reactivas no basta con percibir *ex post* las razones morales que respaldan la queja de la parte ofendida (Darwall, 2006,p. 75; Watson, 2011), sino que se requiere además de una variedad de disposiciones afectivas –tales como la susceptibilidad a experimentar culpa y arrepentimiento (Macnamara, 2015, p. 216) y empatía con

8 Como puede verse, Gary Watson es entonces una figura central tanto para las teorías punitivistas como para las comunicativas, puesto que, en distintos artículos, sentó las bases de ambas.

9 En realidad, Watson es ambivalente respecto a esta conclusión debido a la posibilidad de agentes con los que sea imposible dialogar moralmente, no porque sean incapaces, sino porque rechazan de tajo nuestras categorías morales (Watson, 1987/2004, p. 239). Es difícil saber cómo integrar esta ambivalencia en una interpretación global del artículo en cuestión, pero, en cualquier caso, las tesis gemelas de que las actitudes reactivas poseen una dimensión comunicativa y de que la capacidad de comprender su mensaje es central para la agencia responsable han sido adoptadas por diversos teóricos, en especial por Darwall (2006), Shoemaker (2007), McKenna (2012) y Macnamara (2015).

los sentimientos de la parte ofendida (Shoemaker, 2007)–, así como disposiciones conductuales, por ejemplo a pedir disculpas y a tomar medidas correctivas y restaurativas (McKenna, 2012, p. 91).

Ahora bien, una objeción que puede presentarse contra la idea de que las actitudes reactivas son en sí mismas comunicativas es que uno puede experimentarlas incluso en casos en los que el agente a quien van dirigidas está ausente o, aun estando presente, se rehúsa a darse por aludido y se niega así a entablar una conversación moral (Talbert, 2012). En respuesta, el defensor de la teoría conversacional sostiene que las actitudes reactivas son esencialmente comunicativas no porque siempre consigan "hacer contacto" con el ofensor, sino en virtud de que poseen dos características distintivas de cualquier mensaje: tienen un contenido representacional y tienen por función la de suscitar en otros la comprensión de dicho contenido. En cuanto al contenido, las actitudes reactivas representan al agente como alguien que exhibió mala voluntad al violar una norma moral; en cuanto a su función, estas actitudes producen comprensión de aquel contenido a través del sentimiento de culpa que despiertan en el ofensor, ya que es a través de éste que las personas entienden que han quebrantado una norma (Macnamara, 2015). Es importante tener en claro que la función asignada en esta teoría a las actitudes reactivas –generar comprensión en el ofensor a través de sentimientos de culpa– no depende en absoluto de las intenciones de la persona que las alberga, sino que es una "función etiológica" (Macnamara, 2015, p. 225), es decir, es el resultado de la historia evolutiva de cómo esas actitudes llegaron a existir. Así pues, según estos teóricos, las actitudes reactivas siguen siendo comunicativas aun cuando quien las experimenta no quiere o no puede emplearlas para transmitir mensaje alguno y aun cuando el blanco de aquéllas se rehúsa a darse por enterado.

En mi opinión, las teorías comunicativas o conversacionales representan una significativa mejora respecto de las teorías punitivistas, ya que es más plausible concebir a las actitudes reactivas como entidades comunicativas en el sentido explicado hace un momento que como instrumentos para implementar sanciones; en consecuen-

cia, es más plausible sostener que la sensibilidad a razones *ex post*, a diferencia de la sensibilidad *ex ante*, es constitutiva de la agencia responsable. Sin embargo, en mi estimación, hay buenas razones para negar a fin de cuentas ambos postulados de las teorías conversacionales. Las razones en cuestión se basan en consideraciones similares a las que me llevaron a descartar las teorías punitivistas, a saber, la existencia de "puntos ciegos" en la comprensión moral que aquejan a cualquier agente ordinario. Esos puntos ciegos constituyen contraejemplos convincentes contra la tesis de que la capacidad *local* de entender el mensaje que transmiten las actitudes reactivas en un contexto específico y en relación con una acción incorrecta en particular es necesaria para la justificación de esas actitudes.

Piénsese por ejemplo en una persona razonablemente comprometida con el respeto a la diversidad de opciones de vida, pero que es incapaz de aceptar que el transexualismo sea una de ellas (esto es, una opción legítima que debe ser respetada). Esta persona, sin ser abiertamente hostil o agresiva hacia las personas transexuales, sí se permite, cuando el tema sale a discusión, expresar sus opiniones en términos que resultan ofensivos para aquéllas. En una ocasión, alguien le reprocha su insensibilidad, a lo que la persona responde encogiéndose de hombros y negándose a admitir que su comportamiento es inaceptable, así como a pedir disculpas y emprender acción correctiva alguna. Ahora bien, la misma pregunta que planteé arriba en relación con las teorías punitivistas vuelve a ser relevante aquí: ¿es ese tipo de ceguera moral local un impedimento para dirigir al ofensor actitudes reactivas plenamente justificadas? Como allá, la respuesta que ofrece aquí el sentido moral ordinario es, me parece, contundentemente negativa: la persona tránsfoba no queda exenta de atribuciones de responsabilidad a través de las actitudes reactivas en virtud de su incapacidad local para comprender *ex post* el tipo de consideraciones morales que están en juego en el ejemplo.

En respuesta a esta clase de contraejemplo a la teoría conversacional, los filósofos que se adhieren a ella responden que una cosa es ser auténticamente incapaz de comprender ciertas consideracio-

nes morales y otra muy diferente poseer la capacidad pero negarse a ejercitarla (Watson, 2011; Macnamara, 2015, p. 232, n. 37). Estos filósofos parecen asumir que las capacidades morales relevantes para la agencia responsable son necesariamente capacidades "globales", esto es, son tales que, si uno las posee en determinado contexto y respecto de ciertas razones, entonces uno las posee en todo contexto y respecto de todas las razones. De esta manera, su diagnóstico del ejemplo sería que la persona tránsfoba, al ser capaz en varios contextos de respetar la diversidad de opciones de vida, por supuesto que puede comprender las razones en juego en el caso de la transexualidad. Esta respuesta adolece, sin embargo, de dos serios problemas. Por un lado, pasa por alto el hecho de que las capacidades morales son sumamente sensibles al contexto de ejecución (véase más arriba la discusión del "circunstancialismo" de Vargas) y, por otro, pierde de vista que las capacidades morales no están dadas en un agente de una vez y para siempre, sino que son susceptibles de mejoramiento a lo largo del tiempo (McGeer, 2018). Si este es el caso, aunque la persona tránsfoba podría sin duda llegar a deshacerse de sus prejuicios y aprender así a respetar la transexualidad, ello no quiere decir que, tal como esa persona es en el momento presente, pueda comprender las razones pertinentes y pedir disculpas por su comportamiento.

En suma, pues, este ejemplo muestra que las actitudes reactivas no dejan de ser apropiadas en virtud de que el agente a quien van dirigidas sea incapaz de comprender el mensaje que buscan transmitir (que violó una norma moral y debe disculparse), lo cual significa que tampoco la sensibilidad a razones *ex post* es una condición necesaria de la agencia responsable.

5. Teorías de la calibración interpersonal

Llegamos así a la tercera clase de teorías sobre la agencia responsable mencionadas al inicio, las teorías de la calibración interpersonal. Como lo adelanté, éstas se caracterizan por negar que la sensibilidad a razones –tanto *ex ante* como *ex post*– constituya una condición necesaria para ser legítimamente llamado a cuentas a través

de las actitudes reactivas. Por lo tanto, dichas teorías les niegan a estas actitudes el estatus tanto de sanciones como de portadoras de mensajes. En cambio, en su opinión las actitudes reactivas son más bien herramientas con las cuales calibramos o ajustamos nuestras relaciones interpersonales en respuesta al "significado moral" que transmiten las acciones de los demás. En lo que sigue, ahondaré en ambos aspectos de la propuesta.

Thomas Scanlon (2008, 2015), quien formuló originalmente la teoría (que yo llamo) de la calibración interpersonal, sostiene que la función de las actitudes reactivas negativas no debe buscarse en el efecto que producen en el ofensor (ya sea sancionarlo o enmendarlo), sino en las razones por las cuales quien experimenta esas actitudes las encuentra valiosas[10]. Las razones en cuestión, prosigue Scanlon (2015, p. 92), "derivan de las razones que tenemos para estar concernidos con las actitudes normativas de los demás", esto es, derivan del interés, compartido por todos, de entablar relaciones interpersonales que estén en sintonía con las actitudes que los demás nos prodigan. No queremos confiar en quienes no son de confianza, ni ser amigables con quienes no se preocupan por nosotros, ni promover los proyectos de alguien que ve a los demás exclusivamente en términos instrumentales, ni tener erróneamente en alta estima a quien no lo merece. Para que esta calibración interpersonal se desarrolle apropiadamente, nuestra conducta debe ser receptiva al significado moral detrás de las acciones de los demás (Scanlon, 2015, p. 93). Las actitudes reactivas –tanto las negativas como las positivas– son las herramientas con las cuales llevamos a cabo la calibración.

Así, por ejemplo, si alguien que creíamos nuestro amigo evidencia con sus acciones que no tiene mayor interés en nosotros (no se preocupa por nuestro bienestar, busca nuestra compañía solo cuando requiere ayuda, no muestra interés en nuestros proyectos, etc.), la reacción natural es de resentimiento acompañado de una

10 Otros filósofos que adoptan el marco teórico de Scanlon son Talbert (2012), A. Smith (2019) y Hieronymi (2019).

serie de manifestaciones conductuales encaminadas a modificar o quizá a concluir con la relación: dejaremos de compartir nuestras confidencias con esa persona, no la ayudaremos más en sus predicamentos, ya no la invitaremos a departir con nosotros, etc. Si bien es cierto que la persona que sufre estas consecuencias puede llegar a percibirlas como castigos que se le imponen por su conducta, en realidad no son tales, pues, como explica Scanlon (2015, p. 93), "no son implementadas *porque* sean dañinas desde la perspectiva de quien las padece, sino por razones que tienen que ver con nuestro propio interés en el tipo de relación que mantenemos con él".

Como vimos anteriormente, McKenna (2019) insistiría en que, puesto que con independencia de las intenciones de quien adopta las actitudes reactivas, éstas, al ser expresadas, generan de hecho ciertos perjuicios a quien van dirigidas, entonces su expresión debe estar sujeta a determinadas normas morales, como por ejemplo la que impone el requisito de evitabilidad discutido antes. Pero el teórico de la calibración interpersonal rechaza esta conclusión, pues, de acuerdo con él, el único requisito válido que gobierna a las actitudes reactivas es el que Scanlon, (2008,p. 180) denomina "precisión psicológica": estas actitudes son apropiadas si y sólo si aquellos a quien van dirigidas mostraron en efecto indiferencia o mala voluntad en su conducta (cuando ésta es incorrecta) o, en otras palabras, si y sólo si sus acciones tienen el significado moral que aparentemente expresan. Si los destinatarios de estas actitudes son o no capaces de haberse comportado de mejor modo o de entablar una conversación moral, ello es irrelevante para evaluar la idoneidad de las reacciones dirigidas a ellos.

Antes de proseguir, conviene aclarar la noción de significado moral[11]. El significado moral de una acción depende de las razones que impulsaron al agente a llevarla a cabo y de lo que dichas razones muestran acerca de la relevancia otorgada por él a los intereses de los demás. Por ejemplo, la acción de ayudar a alguien en apuros tiene dos significados morales muy diferentes si, en un caso,

11 Scanlon (2008) y McKenna (2012) discuten a fondo esta noción.

el agente actuó por una preocupación genuina por el bienestar de aquella persona mientras que en otro actuó por interés propio (en previsión de una recompensa, digamos). Pero, además, el significado moral de una acción depende no solo de las razones por las cuales el agente la realizó, sino también de qué otros hechos con relevancia moral es capaz de apreciar, Tomemos como ejemplo a dos agentes que realizan la misma acción incorrecta –como burlarse de la tartamudez de alguien– por la misma razón –para hacer reír a sus amigos–. Si uno de ellos es un niño pequeño y el otro un adulto moralmente competente, es plausible pensar que sus acciones, a pesar de ser exteriormente idénticas y de haber sido motivadas por razones similares, transmiten significados morales diferentes. Mientras que el significado de la acción del niño puede ser simplemente "Me parece divertida la forma en que habla esta persona", la acción del adulto también significa algo como "y mi diversión es más importante que la humillación que esta persona probablemente sentirá". Esta capa adicional de significado depende de qué otros hechos con relevancia moral el adulto registra, como, por ejemplo, el hecho de que las personas son emocionalmente vulnerables y el hecho de que mofarse de ellas por sus características inusuales suele generarles una profunda aflicción[12].

Ahora bien, de lo que *no* depende el significado moral de una acción incorrecta es de la capacidad local del agente para responder adecuadamente –ya sea de manera *ex ante* o *ex post*– a las razones morales en contra de ella. Para ver claramente que esto es así, podemos echar mano de lo que denomino "el test de la evaporación". Lo que el test nos pide es enfocar la atención en un malhechor que de manera espontánea despierta nuestras actitudes reactivas y, a continuación, estipular que el agente en cuestión no era capaz, ni antes ni después de la acción, de comprender cabalmente las razones morales en contra de ésta y de actuar en consecuencia. Finalmente, debemos preguntarnos: ¿el objetable significado moral que transmiten las acciones del malhechor se evapora o desapare-

12 Utilicé este mismo ejemplo en mi artículo Rudy-Hiller (2024).

ce a causa de la incapacidad local que hemos estipulado? En otras palabras, ¿inhibe esa incapacidad nuestra tendencia a experimentar actitudes reactivas negativas hacia él?

Para responder estas preguntas, sugiero volver a los dos ejemplos introducidos antes. En el caso de Weinstein, es plausible suponer que su incapacidad local para comprender las razones morales en contra de su trato violento hacia las mujeres estaba causada, al menos en parte, por lo mucho que gozaba sometiéndolas a su voluntad; en el caso de la persona tránsfoba, la causa de su intolerancia seguramente radica en la aversión visceral que le produce la mera mención de la transexualidad. ¿Proveen esas incapacidades una razón suficiente para reinterpretar el significado moral de las acciones de ambos personajes o, puesto de otra manera, constituyen aquéllas un obstáculo para experimentar actitudes reactivas negativas hacia ellos? Por tercera vez en este trabajo, apelo al sentido moral ordinario con la confianza de que, también aquí, la respuesta a estas preguntas será negativa[13].

Por las razones anteriormente expuestas, el teórico de la calibración interpersonal le niega al principio de evitabilidad –el cual rige la aplicación de sanciones y castigos– relevancia alguna para evaluar a las actitudes reactivas. ¿Apela este teórico a algún otro principio moral para justificar el requisito de precisión psicológica que, según él, es el único que regula dichas actitudes? Sí lo hace, en efecto. El principio relevante es enunciado por Scanlon del modo siguiente:

13 Queda claro que, para el teórico de la calibración interpersonal, la sensibilidad a razones no es una condición necesaria de la agencia responsable. ¿Cuáles son, desde su perspectiva, las capacidades que sí lo son? Para Scanlon (1998), así como para Talbert (2008), basta con ser un agente racional, esto es, que gobierna su conducta con base en razones, aunque no necesariamente morales. Por mi parte, sostengo que, además de racionalidad práctica, para transmitir el tipo de significado relevante para las atribuciones de responsabilidad es necesario agregar un mínimo de competencia moral a fin de que el agente sea capaz de apreciar ciertos hechos éticamente relevantes, tales como –para volver a los mencionados antes– que las personas somos seres emocionalmente vulnerables y que hacer mofa de nuestras características inusuales constituye una falta de respeto. El ejemplo discutido más arriba del niño y del adulto que se burlan del tartamudeo de alguien más tiene por objeto resaltar la importancia del requisito de competencia moral mínima.

> El trato usual que prodigamos a las demás personas, y que se ve modificado a consecuencia de la formación de las actitudes reactivas, no se debe a ellas de modo incondicional, sino solo en caso de que manifiesten actitudes apropiadas hacia nosotros y hacia los demás (2015, p. 94).

Como puede apreciarse, el concepto moral básico empleado en este pasaje es el de *reciprocidad*: las otras personas tienen derecho a exigirnos que les dispensemos buen trato a condición de que ellas hagan lo mismo con nosotros. Cuando esta condición no se cumple –como ocurre, por ejemplo, en el caso del amigo ingrato discutido antes–, entonces estamos en pleno derecho de modificar nuestra relación con esa persona a fin de prodigarle el trato que se merece. La noción de "merecimiento" en juego aquí no es, sin embargo, la misma a la que apelan los teóricos punitivistas cuando hablan de merecimiento de castigo; en este caso, el merecimiento no tiene que ver con sanciones, sino, justamente, con la calibración interpersonal, es decir, con el ajuste de nuestro trato hacia otros de modo que esté en sintonía con el significado moral de sus acciones. En consecuencia, para que una actitud reactiva negativa cualquiera esté justificada, basta con que la interpretación de dicho significado sea correcta, esto es, basta con que los actos del ofensor de hecho hayan estado motivados por las razones objetables que se le atribuyen; por lo tanto, las consideraciones de evitabilidad o de disposición al diálogo moral son irrelevantes. Es así como el principio de reciprocidad esbozado por Scanlon justifica el requisito de precisión psicológica que, en su propuesta, es el único estándar al cual deben atenerse las actitudes reactivas.

6. Conclusiones

En este trabajo he presentado mi interpretación del debate contemporáneo acerca de la psicología moral de la agencia responsable y las razones por las cuales considero que la tercera clase de teorías discutidas –las de la calibración interpersonal–, al explicar de modo más fiel las prácticas ordinarias de atribución de responsabilidad, son las que mejor se ajustan al marco teórico delineado por Strawson en

"Libertad y resentimiento". En la vida cotidiana, experimentamos las actitudes reactivas negativas de manera espontánea en respuesta a la interpretación –casi siempre automática– que damos a la conducta de los demás, sin parar mientes en si la persona a quien las dirigimos pudo haber actuado de mejor manera siguiendo las razones morales pertinentes o si podrá, tras haber comprendido los motivos de nuestro resentimiento e indignación, entablar una conversación moral al respecto. Más aún, incluso cuando contamos con evidencia suficiente de que un ofensor no era capaz de realizar ni lo uno ni lo otro en relación con una acción incorrecta en particular, ello no inhibe nuestras actitudes reactivas ni tampoco nos da razones para juzgar que, aunque no ocurra así, debería hacerlo. Como expliqué apelando al principio de reciprocidad que propone Scanlon (2015), tenemos, más bien, excelentes razones para juzgar y reaccionar a la conducta de los demás con base en el significado moral que ésta de hecho exhibe, sin ocuparnos en especulaciones acerca de lo que el agente pudo haber hecho al momento de la acción o de lo que podrá hacer posteriormente gracias a nuestra intervención.

Referencias

Austin, J. L. (1961). A plea for excuses. En *Philosophical Papers* (pp. 175-204). Oxford University Press.

Brink, D. y Nelkin, D. (2013). Fairness and the Architecture of Responsibility. *Oxford Studies in Agency and Responsibility*, 1, 284–314.

Darwall, S. (2006). *The Second-Person Standpoint*. MA: Harvard University Press.

Fischer, J. y M. Ravizza (1998). *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge University Press.

Frankfurt, H. (1969/1988). Alternate possibilities and moral responsibility. En *The importance of what we care about* (pp. 1-10). Cambridge University Press.

Hart, H. L. A. (1968). Postscript: Responsibility and retribution. En *Punishment and Responsibility* (pp. 210-237). Oxford University Press.

Hieronymi, P. (2019). I'll Bet You Think This Blame is About You. *Oxford Studies in Agency and Responsibility*, 5, 60–87.

Macnamara, C. (2015). Blame, communication and morally responsible agency. En R. Clarke, M. McKenna, and A. Smith (Eds) *The Nature of Moral Responsibility* (pp. 211-236). Oxford University Press

McGeer, V. (2013). Civilizing Blame. En Coates J. and N. Tognazzini (Eds.). *Blame: Its Nature and Norms* (pp. 162-188). Oxford University Press.

McGeer, V. (2019). Scaffolding Agency: A Proleptic Account of the Reactive Attitudes. *European Journal of Philosophy*, 27, (2), 301-323.

McGeer, V. y P. Pettit (2015). The Hard Problem of Responsibility. *Oxford Studies in Agency and Responsibility*, 3, 160-188.

McKenna, M. (2012). *Conversation and Responsibility*. Oxford University Press.

McKenna, M. (2013). Reasons-Responsiveness, Agents, and Mechanisms. *Oxford Studies in Agency and Responsibility*, 1, 151-84.

McKenna, M. (2019). Basically Deserved Blame and its Value. *Journal of Ethics and Social Philosophy*, 15 (3), 1-28.

Nelkin, D. (2011). *Making Sense of Freedom and Responsibility*. Oxford University Press.

Nelkin, D. (2013). Desert, Fairness, and Resentment. *Philosophical Explorations*, 16 (2), 117-132.

Nelkin, D. (2016). Accountability and Desert. *The Journal of Ethics*, 20, 173-189.

Pettit, P. (2002). The Capacity to Have Done Otherwise En *Rules, Reasons, and Norms* (pp. 257-273). Oxford University Press.

Rosen, G. (2015). The alethic conception of moral responsibility. En R. Clarke, M. McKenna, and A. Smith (Eds.). *The Nature of Moral Responsibility* (pp. 65-88). Oxford University Press.

Rudy Hiller, F. (2021). It's (almost) all about desert: on the source of disagreements in responsibility studies. *The Southern Journal of Philosophy*, 59 (3), 386-404.

Rudy Hiller, F. (2022). The moral psychology of moral responsibility. En J. Doris y M.

Vargas (Eds.). *The Oxford Handbook of Moral Psychology* (pp. 509-542). Oxford University Press.

Rudy Hiller, F. (2024). Accountability, reasons-responsiveness, and narcos moral responsibility. *Asian Journal of Philosophy*. Disponible en línea: 10.1007/s44204-024-00188-1

Scanlon, T. M. (1998). *What We Owe to Each Other*. MA: Harvard University Press.

Scanlon, T. M. (2008). *Moral Dimensions*. MA: Harvard University Press.

Scanlon, T .M. (2015). Forms and Conditions of Responsibility. En R. Clarke, M.

McKenna, and A. Smith (Eds.) *The Nature of Moral Responsibility* (pp. 89-111) Oxford University Press..

Shoemaker, D. (2007). Moral address, moral responsibility, and the boundaries of the moral community. *Ethics*, 118 (1), 70-108.

Shoemaker, D. (2020). Responsibility: the State of the Question. Fault Lines in the Foundations. The Southern Journal of Philosophy, 58 (2), 205-237.

Smith, M. (2004).Rational Capacities. En Ethics and the A Priori. (pp. 114-135) Cambridge University Press..

Smith, A. (2019). Who's Afraid of a Little Resentment? *Oxford Studies in Agency and Responsibility*, 6, 85-111.

Strawson, P. (1962/1992). Libertad y resentimiento. En *Cuadernos de Crítica*, Vol. 47. UNAM.

Talbert, M. (2008). Blame and Responsiveness to Moral Reasons: Are Psychopaths Blameworthy? *Pacific Philosophical Quarterly*, 89, 516-535.

Talbert M. (2012). Moral Competence, Moral Blame, and Protest. *Journal of Ethics*, 16 (1),89-109.

Vargas, M. (2013). *Building Better Beings*. Oxford University Press.

Wallace, J. (1994). *Responsibility and the Moral Sentiments*. MA: Harvard University Press.

Watson, G. (1987/2004). Responsibility and the Limits of Evil. En *Agency and Answerability* (pp. 219-259). Oxford University Press.

Watson, G. (1996/2004). Two Faces of Responsibility. En *Agency and Answerability* (pp. 260-288). Oxford University Press.

Watson, G. (2011).The trouble with psychopaths. En , R. J. Wallace, R. Kumar, and S. Freeman (Eds.). *Reasons and Recognition*. Oxford University Press.

Wolf, S. (1990). *Freedom within Reason*. Oxford University Press.

CAPÍTULO IX

LA MÁQUINA DE EXPERIENCIAS DE NOZICK: ALCANCES DE LA CRÍTICA AL HEDONISMO ÉTICO

Luis Francisco Estrada Pérez[1]

RESUMEN

Algunos filósofos han juzgado que una vida buena debe evaluarse en función de la cantidad de felicidad que contiene (Nozick, 2013, p. 79). Epicuro (2005) afirmaba que la felicidad es la ausencia de dolor. Bentham (2000) que la felicidad debe evaluarse como maximización de placer y reducción del dolor. Conjuntamente han asumido que los términos felicidad y placer son sinónimos, de modo tal

1 Universidad Nacional Federico Villarreal (Perú). Magíster en Epistemología y candidato a Doctor en Filosofía por la Universidad Nacional Mayor de San Marcos (Perú). Miembro del Grupo de Investigación Episteme (UNMSM). Sus líneas de investigación son Filosofía del Lenguaje, Filosofía Analítica, Ética, Epistemología, Filosofía Política y Filosofía de la Mente.

que se estima el nivel de felicidad de una persona por la cantidad de placer que experimenta. Nozick (2017) presenta el experimento mental de la máquina de experiencias que grafica el escenario hipotético de un artilugio que brinda innumerables cantidades de experiencias placenteras: ¿se conectaría para siempre a esta máquina? Tradicionalmente se ha asumido este *Gedankenexperiment* como una crítica al hedonismo ético (Feldman, 2011), posición que sostiene que las acciones morales deben evaluarse según la cantidad de placer (felicidad, bienestar) que proporcionan (Sumner, 2003), pues, a juicio de Nozick, la cantidad de felicidad no es lo único relevante en la vida. El presente trabajo analizará el mencionado experimento mental y ponderará la crítica de Nozick al hedonismo ético en dos de sus variantes: hedonismo sensorial, posición que privilegia los datos sensoriales como objeto del placer; y hedonismo actitudinal, doctrina que prepondera la disposición mental hacia las sensaciones (Feldman, 2004). A diferencia de Feldman (2011), quien sostiene que el experimento mental de la máquina de experiencias no es concluyente respecto a su crítica al hedonismo ético, se afirmará que la crítica de Nozick encajaría en el llamado hedonismo actitudinal. Finalmente, siguiendo la interpretación de Sher (2004) y Shea - Kintz (2022), se postulará que los planteamientos de Nozick se vinculan con el perfeccionismo ético, doctrina que afirma que algunas formas de desarrollar la vida humana son objetivamente valiosas y tienen un valor intrínseco, lo cual se complementa con su visión de la unidad orgánica (Nozick, 1983).

Palabras clave: Hedonismo actitudinal, hedonismo ético, máquina de experiencias, Perfeccionismo ético. Nozick.

ABSTRACT

Some philosophers have judged that a good life should be evaluated based on the amount of happiness it contains (Nozick, 2013, p. 79). Epicurean (2005) stated that happiness is the absence of pain. Bentham (2000) that happiness should be evaluated as maximization of pleasure and reduction of pain. Also they have assumed that the terms happiness and pleasure are synonyms, such that a person's

level of happiness is estimated by the amount of pleasure they experience. Nozick (2017) presents the experience machine thought experiment that charts the hypothetical scenario of a gadget that delivers countless amounts of pleasurable experiences: would you plug into this machine forever? Traditionally, this *Gedankenexperiment* has been assumed as a criticism of ethical hedonism (Feldman, 2011), a position that maintains that moral actions should be evaluated according to the amount of pleasure (happiness, well-being) they provide (Sumner, 2003), since, in the opinion of Nozick, the amount of happiness is not the only relevant thing in life. The present work will analyze the aforementioned thought experiment and will ponder Nozick's criticism of ethical hedonism in two of its variants: sensory hedonism, a position that privileges sensory data as an object of pleasure; and attitudinal hedonism, a doctrine that predominates the mental disposition towards sensations (Feldman, 2004). Unlike Feldman (2011), who maintains that the thought experiment of the experience machine is not conclusive with respect to his criticism of ethical hedonism, it will be stated that Nozick's criticism would fit into the so-called attitudinal hedonism. Finally, following the interpretation of Sher (2004) and Shea - Kintz (2022), it will be postulated that Nozick's approaches are linked to ethical perfectionism, a doctrine that affirms that some ways of developing human life are objectively valuable and have a intrinsic value, which is complemented by his vision of organic unity (Nozick, 1983).

Keywords: Attitudinal hedonism, ethical hedonism, experience machine, ethical perfectionism, Nozick.

1. La máquina de experiencias

En su connotada obra *Anarquía, Estado y Utopía* (2017) Robert Nozick presenta el famoso experimento mental llamado la máquina de experiencias. Supongamos que existe un artefacto capaz de brindarnos ingentes cantidades de experiencias placenteras privadas a lo largo de nuestra vida. Al final de cuentas, en el mundo real, nuestro acceso al placer o dolor es una cuestión interna. Usted recibe datos sensoriales y los interpreta en la percepción. El debate clásico de

corrientes escépticas asume que nunca podremos tener un acceso que no esté basado precisamente en esa instancia interna: nuestras sensaciones, nuestros recuerdos, nuestras impresiones, etc., se fundamentan, en última instancia, en el fuero interno (Nagel, 2003, p.13-19). De modo tal que la opción que ofrece la máquina de experiencias no es del todo descabellada. Asumiendo que nuestro principal interés en la vida es maximizar las experiencias de placer y reducir las experiencias de dolor, sería racional que usted deseara entrar en este dispositivo, pues le ofrece todas aquellas experiencias placenteras que en algún momento le fueron esquivas: ser una estrella pop, un empresario multimillonario, un amante lascivo, etc. Incluso podrán ser experimentadas de forma tan vívida como en el mundo real. Si bien Nozick (2017, p. 53-54) señala que en el mundo tenemos acceso a máquinas de experiencias locales (como el alcohol, las drogas, el sexo, etc., que nos "desconectan" del mundo real) el gran reto no es permanecer por un tiempo estipulado dentro de la máquina de experiencias, sino conectarnos para siempre.

Si usted está interesado en maximizar sus experiencias placenteras y reducir las experiencias de dolor, la alternativa que le ofrece la máquina de experiencias es la mejor a la que podría acceder, pues le garantiza que tendrá las mismas sensaciones vívidas que percibe en el mundo real. No solo eso, sino que, si usted no tiene claro de acá a unos años que formas de placer puede experimentar, tendrá la asesoría de expertos para que oportunamente preprograme hasta el fin de sus días innumerables sensaciones pletóricas de placer. ¿Aceptaría este tipo de vida para siempre? ¿Consideraría que esta sería la mejor forma de vida a la que puede aspirar? Tradicionalmente, la corriente que ha enarbolado la necesidad de evaluar las acciones morales en función de la cantidad de placer que proporcionan es el hedonismo ético. Precisamente por ello, muchos investigadores han considerado que el objeto de la crítica de Nozick a través de su *puzzle* es la referida corriente. A continuación, se pondrá en debate dicha interpretación del experimento mental de Nozick partiendo de una definición clara del hedonismo ético.

2. Hedonismo ético: una doctrina de larga data

Feldman (2004, p. 21) afirma que el hedonismo sugiere entre el público no académico una corriente muy cercana a la frase "sexo, drogas y rocanrol". Ello revela una mala interpretación del hedonismo más serio (académico). Epicúreo (2004) mismo lamentó este tipo de tergiversación. Insistía en que cuando se habla de placer como el objetivo de nuestras acciones no se habla de placeres vulgares, como algunos pueden creer, sino que refiere, entre otras cosas, a la carencia de dolor corporal y ausencia de intranquilidad en el alma. Mill (2014, p. 61), también tuvo que hacer frente a los ataques de los críticos que sostenían que estaba defendiendo una doctrina valiosa sola para los cerdos[2]. Al igual que Epicúreo, tuvo que batallar para explicar que la doctrina hedonista no tenía estas desagradables implicancias. Incluso en nuestros días, muchos críticos del hedonismo se sienten incapaces de comulgar con una doctrina que entienden cercana a un sensualismo individualista[3].

Es menester, sostiene Feldman (ibídem), que se haga una lectura justa de la doctrina hedonista. Sin embargo, los problemas no tardan en llegar pues la literatura hedonista profesional está plagada de perspectivas defectuosas. De esta forma, el mayor problema no vendría por parte de la crítica sino, por el contrario, de la dificultad de enunciar de forma clara el hedonismo ético. Por ejemplo, Frankena (1988, p. 84) indica que para saber qué tipo de placer reclama el hedonista debemos tener en cuenta las siguientes premisas:

2 Se entiende que la referencia de Mill a los cerdos pretende enfatizar la visión estigmatizada del hedonismo como una teoría que defiende los placeres vulgares en lugar de una perspectiva que considere no solo la cantidad sino la calidad del placer.

3 Parfit (2004, p. 60) califica al hedonismo como una teoría indirectamente contraproducente.

(1) Felicidad= placer
(2) Todos los placeres son intrínsecamente buenos, o buenos en sí mismos[4]
(3) Solo los placeres son intrínsecamente buenos, o cualquier cosa que sea buena en sí misma es placentera en sí misma
(4) Lo placentero es el criterio de la bondad intrínseca. Es lo que hace que las cosas sean buenas en tanto que fines.

Sin embargo, aparte de las cuatro premisas antes señaladas, encontramos también diferencias respecto a la naturaleza del placer. Se puede medir cuantitativamente los niveles de dopamina que se producen al escuchar el género musical favorito, pero parece poco viable poder cuantificar la predilección por un género particular, pues ello involucra decisiones, actos volitivos, sesgos cognitivos, etc. Se sostiene que en el primer caso estamos frente a un enfoque cuantitativo del placer, esto es, que las sensaciones placenteras se pueden traducir a guarismos; mientras que, en el segundo caso, nos encontramos ante una perspectiva cualitativa del placer, dado que ciertas facetas placenteras no se pueden cuantificar. Epicuro (2005) y Bentham (2000) afirman que semejantes diferencias en cualidad no hacen diferencias a su bondad o valor. Ellos apuestan por la cantidad de placer más que por la discusión de la cualidad. Por el contrario, Mill (2014, p. 62) sostiene que una diferencia en cualidad, sí hacen diferencias en el valor, por ejemplo, los placeres intelectuales (amor al conocimiento, cumplimiento de un plan de vida, selección de valores, etc.) son superiores a los corporales (alimentación, concupiscencia, salud, etc.) aun cuando afirma que sin la satisfacción de los últimos es inviable alcanzar los primeros. Esta es la distinción clásica entre un hedonismo cuantitativo y cualitativo. Por ello, Frankena (1988) pide que se considere una quinta tesis:

4 Se entiende que la apelación a la bondad intrínseca del placer tiene por objetivo hacer justicia al programa hedonista original alejado a un individualismo sensualista y carente de objetividad.

(5) La bondad intrínseca de una actividad o experiencia es proporcional a la cantidad de placer que contiene (o más bien la cantidad balanceada de placer por encima de dolor contenido en ella o intrínseco a ella).

Frankena (1988, p. 88) asevera que, de las cinco tesis hedonistas presentadas, los defensores de una posición no- consecuencialista niegan (4) y (5). La mayoría de ellos niegan (2), pues insisten que algunos placeres son intrínsecamente malos, por ejemplo, placeres emparentados con el perjuicio de un tercero, o aquellos que involucran una mala disposición moral como la crueldad o malicia o el placer en la maldad. Algunos también rechazan (3), arguyendo que hay valores que son intrínsecamente buenos, pero que no contienen placer, aunque pueden ocasionarlo, por ejemplo, la belleza, la verdad, la virtud, etc.

A juicio de Sumner (2003, p. 92), el hedonismo clásico produce una teoría del estado mental acerca del bienestar porque considera el placer y el dolor como estados distintivos de la mente. La teoría resultante nos dice que nuestras vidas van bien cuando tenemos sentimientos placenteros y mal cuando tenemos sentimientos dolorosos, que se está mejor que otra persona cuando tenemos más de los primeros y menos de los últimos, que algo me beneficia cuando me causa lo primero y me daña cuando me provoca lo segundo y así sucesivamente.

3. Hedonismo ético: debates

Se ha mencionado la distinción entre el hedonismo cuantitativo que considera la bondad en función de la cantidad del placer y el hedonismo cualitativo que apunta a la cualidad del placer. Los hedonistas cuantitativos aceptan (5), pero los hedonistas cualitativos lo niegan. Mill (2014) evidentemente era partidario de un hedonismo cualitativo. Ello va en consonancia con su preferencia por los placeres intelectuales por encima de los corporales. Dos tipos de argumentos principales han sido esgrimidos en el debate entre los hedonistas y los no- hedonistas. En primer lugar, una línea de argumento psico-

lógica. Los hedonistas, sean cuantitativos o cualitativos, argumentan que el placer es un bien en sí mismo porque es, al final de cuentas, lo que todos buscamos obtener. De esta forma, Mill sostiene, a juicio de Frankena (1988, p. 85), que el placer y solo el placer es bueno como fin[5]. La premisa de este argumento es una doctrina psicológica, una teoría de la naturaleza humana, que es llamada hedonismo psicológico. La conclusión, sin embargo, es un juicio de valor. Como resultado de ello, muchos no- hedonistas, desde Moore (2005, p. 25)[6], han atacado a este argumento como ilógico[7]:

De

(a) El placer y solo el placer es deseado como un fin.

Uno no puede inferir:

(b) El placer y solo el placer es un bien como fin.

Lo cual resulta admisible, pues existen términos en (b) que no se pueden inferir de (a). Por ejemplo, el mero deseo resulta pobre para fundamentar ontológicamente al placer como un bien, v.g. desear una sustancia psicoactiva no la hace buena en sí misma. La posición de Mill no es que (b) se sigue lógicamente de (a), sino que (a) es verdadera como una teoría de la naturaleza humana; y, si la naturaleza humana está constituida para apuntar al placer, entonces

5 Frankena se apoya en la siguiente cita de Mill (2014, p.121): "Si la opinión que he manifestado ahora es psicológicamente verdadera- si la naturaleza humana está constituida de tal forma que no desea nada que no sea ya bien, una parte de la felicidad o un medio para la felicidad- no podemos contar con ninguna otra prueba, y no necesitamos otra, con relación a que éstas son las únicas cosas deseables"

6 To many persons it may seem clear that it would be our duty to prefer some pleasures to others, even if they did not entail a greater quantity of pleasure; and hence that though actions which produce a maximum of pleasure are perhaps, in fact, always right, they are not right because of this, but only because the producing of this result does in fact happen to coincide with the producing of other results.

7 Cf. Ross (1930, p. 138), Brentano (2009, p. 19), Kagan (1998, p.35) y Rawls (2010, p.501).

es absurdo o irrazonable negar que el placer sea lo bueno aun cuando es lógicamente posible.

Para Frankena (ibídem) un hedonista es capaz de poner este argumento en una forma lógicamente válida:

(a) El placer y solo el placer es deseado como un fin.
(b) Lo que es deseado como un fin y solo deseado como un fin es un bien como fin en sí mismo.
(c) El placer y solo el placer es un bien como fin en sí mismo.

Muchos, continúa Frankena (1988, p. 86), no- hedonistas aceptan (b) y rechazan (a). Aristóteles (1098b 10- 1098b 30), por ejemplo, señala que reclamar que el fin, en el sentido de realización, al que todas las cosas apuntan no es necesariamente el bien, es un sinsentido: si algo es el fin de una cosa, también tiene que ser un bien. En realidad, empieza argumentando que, el bien puede ser definido como aquello a lo que todas las cosas apuntan. Afirma que todas las cosas apuntan a la felicidad, pero niega que la felicidad sea el placer. La felicidad es una actividad de excelencia del alma, al igual que una actividad en concordancia con la moral y, en especial, con las virtudes intelectuales de excelencia que incluyen la ciencia, la sabiduría y otras formas de conocimiento. Es esta actividad de excelencia nuestro fin, no el placer; pues este último es un acompañante en nuestro ejercicio de excelencia. El argumento no prueba que el hedonismo está errado como una teoría del valor, aun cuando aceptemos (b), pero parecería demostrar que el hedonismo no ha sido probado, que probablemente no sea verdadero, y que no puede ser utilizado como parte de un argumento para establecer una teoría hedonística del valor.

Es preciso tener presente, sin embargo, que, en las bases de la presentación del hedonismo ético, dos tipos de cosas se asumen como buenas en tanto que fines: por un lado, algo semejante al placer, disfrute, satisfacción; por otro, alguna forma de excelencia o autoperfección. Muchos hedonistas y no hedonistas critican esta presentación hedonista, sosteniendo, por el contrario, que una

cosa no es buena solo por el hecho de ser deseada, v.g. alguien puede desear acosar a una persona, pero eso no convierte a su acto en algo bueno. Estos autores apelan a un tipo de revisión reflexiva del tipo de cosas que solemos tomar como fines o como buenas en sí mismas. Sidgwick (1962), por ejemplo, piensa que esta inspección es un proceso de intuición de juicios autoevidentes. El punto principal es que la revisión debe ser reflexiva y debe limitarse a sí misma rigurosamente a la cuestión de qué es lo bueno en sí mismo aparte de sus consecuencias e implicaciones morales. De esta manera la pregunta de forma cruda sería: "¿qué tipo de cosas son racionales de desear por su propio valor?" Luego en el corazón de la doctrina hedonista subyace una pregunta última sobre qué es lo más importante en la vida y cuál es la mejor manera de vivir. Precisamente, son estas las cuestiones últimas y medulares que el *puzzle* de Nozick desentraña y que se analizará a continuación.

3.1. Primer argumento contra el hedonismo

Feldman (2011, p. 67) señala que muchos especialistas no estiman que el hedonismo ético, entendiendo a esta doctrina como una teoría vinculada al bienestar, sea atacado directamente por el pasaje de la máquina de experiencias de Nozick. Según esta propuesta, el bienestar de una persona es directamente proporcional a la cantidad de mayor placer y mínimo dolor que experimente. Una de las consecuencias de esta forma de hedonismo es que una vida posible que tenga un gran balance de mayor placer y mínimo dolor debe tener mayor valor de bienestar que otra, *ceteris paribus*. Por ello, el reto del *puzzle* es indicar qué tipo de hedonista recusaría ingresar a la máquina de experiencias a pesar de que, en principio, ofrecería la mayor cantidad de placer y el mínimo dolor. Ello nos brinda una primera clave, señala Feldman, en la obra citada, para entender el punto principal que muchos especialistas encuentran en el pasaje de la máquina de experiencias.

Sumner (2003, p. 26) sostiene que la postulación de modelos de vida buena (para los seres humanos) son tan viejos como la filosofía. En el mundo occidental, que va desde los griegos hasta

nuestros días, los ideales que se han granjeado apoyo han invocado una gran variedad de conceptos: placer, felicidad, satisfacción de deseos o preferencias, el cumplimiento de las necesidades, el logro de objetivos o metas, el desarrollo de capacidades o potencialidades, el mantenimiento del funcionamiento normal, vivir una forma de vida acorde a la naturaleza propia de uno mismo, y por supuesto sin duda otros más. A su juicio (Sumner, 2003, p. 138) el bienestar es subjetivo, y en este sentido, el hedonismo carece de una referencia externa a los estados mentales del sujeto. La versión clásica del hedonismo suscribe tres tesis fundamentales: (1) la ecuación de bienestar y felicidad, (2) la reducción del bienestar al placer (y, por ende, a la ausencia de dolor), y (3) la consideración de los estados mentales de placer (y dolor). Desde que la mejor teoría acerca de la naturaleza del placer y el dolor los trata como sensaciones, (3) es inobjetablemente de estirpe subjetiva. Sin embargo, no parecería justo asumir que es simplemente la naturaleza mental del placer y el dolor la causante de hacer al hedonismo clásico vulnerable a objeciones de falsedad y error por su vinculación con el subjetivismo. Luego, algo no debe ser consistente en (1) o (2). La idea de que hay una interesante conexión entre el bienestar y la felicidad tiene al menos una plausibilidad inicial mucho mayor que cualquier otra consideración de la felicidad en términos de sentimientos placenteros. Por ello, si aspiramos a una nueva perspectiva del bienestar resulta valioso profundizar en una mejor (y menos hedonística) teoría acerca de la naturaleza de la felicidad.

Feldman (2011, p. 67) nos invita a imaginar un escenario en el que se empieza a vivir un tipo de vida, la vida real (VR). Las cosas ocurren razonablemente bien y si te contentas con tu (VR) podrías terminar con una vida placenteramente justa. Ahora, si usted estuviera en el escenario de Nozick y alguien le ofreciese la oportunidad de conectarse a una máquina de experiencias le diría que si lo hace tendría por el resto de su vida momentos intensos y duraderos de placer. Usted tendría su vida en la máquina de experiencias (VME), la cual le brindaría una significativa cantidad mayor de placer y mínima de dolor. Con estos presupuestos, podemos enunciar

una versión del argumento que para algunos se encuentra oculto en el pasaje de la máquina de experiencias:

Primer Argumento en contra del Hedonismo Ético

1. Si estuvieras en el escenario de Nozick no te conectarías para obtener una VME. Preferirías VR.
2. Si se da (1), entonces VME no tiene un nivel de bienestar mayor que VR.
3. Si VME no tiene un nivel de bienestar mayor que VR, entonces el hedonismo ético es falso.
4. Por consiguiente, el hedonismo ético es falso.

Al parecer, sostiene Feldman (2011, p. 68), (1) es plausible. Al menos, es admisible cuando se construye como una respuesta de lo que alguien haría si estuviese en el escenario de Nozick. Preferiríamos permanecer en la VR mucho antes que tener una VME. La respuesta sería igual de admisible si la VME ofreciera una simulación de la VR tan vívida que no se encuentre diferencia alguna entre ambas, pues para Nozick la VR es axiológicamente más significativa que la VME. Las dificultades vendrían en (2), pues a juicio de Feldman, del hecho de aceptar (1) no se infiere que se debería aceptar (2). Estimar la VR por encima de VME, no implica nada concerniente al valor de bienestar de ambas vidas. Hay distintas explicaciones posibles para ello. Algunas de las principales son epistémicas: ¿cómo puedes estar seguro de que la máquina trabajaría como advierte? Probablemente podría malograrse. ¿Cómo saber si en realidad la VME no ofrecerá mayor cantidad de bienestar que VR? Podría presentarse una situación en la que efectivamente el bienestar de VME sea mayor que el bienestar en VR. Rechazar conectarnos en estas circunstancias, continua Feldman, no tendría implicaciones respecto a los valores de bienestar, pero sí, en línea con Nozick, por el valor intrínseco que atribuimos a la vida en el mundo real y a nuestra conexión con este mundo[8].

8 Los argumentos de Nozick para no entrar en la máquina de experiencias se desarrollarán en el último apartado.

3.2. Segundo argumento contra el hedonismo

Para una situación como la planteada también se puede considerar que el sujeto no solo quiere estar bien informado, indica Feldman (2011, p. 69), sino que también es completamente egoísta y solo se interesa por sí mismo. Baier (2004, p. 282), afirma que el egoísmo común o también llamado psicológico, estipula que todos somos egoístas pues nos preocupamos por nuestro interés como por ejemplo alcanzar el mayor beneficio. Ahora bien, puede que esta definición resulte insuficiente para nuestro razonamiento. Por ello es mejor asumir que el sujeto es un "egoísta bienestarista", es decir, él se interesa por su propio bienestar y se interesa por su "bienestar moral", es decir, su interés no es un simple interés por un placer inmediato sensualista. ¿Puede darse el caso que tal persona aún se niegue a conectarse, dado que la VME le ofrecería la mayor cantidad de experiencias placenteras?, ¿mostraría ello que el hedonismo es falso? Este segundo argumento contra el hedonismo puede enunciarse formalmente:

Segundo Argumento en contra del Hedonismo Ético

1. Si estuviera en el escenario de Nozick, y supiera sin duda alguna que VME podría ser más placentera que VR, y fuese un "egoísta bienestarista", aun así, no se conectaría para obtener VME. Preferiría VR.
2. Si (1) es verdadero, VME no tiene un mayor valor de bienestar que VR.
3. Si VME no tiene un mayor valor de bienestar que VR, luego el hedonismo ético es falso.
4. Por consiguiente, el hedonismo ético es falso.

Feldman (2011, p. 70) considera que (2) está abierto a cuestionamientos por los mismos motivos que el primer argumento. Un problema adicional es que puedes rehusarte a conectarte porque encuentras la idea de una VME ignominiosa pues resulta extraña y antinatural. Sería una imagen semejante al paciente que rechaza un nuevo tratamiento médico nuevo porque le parece abominable,

pero sin mayor justificación. Aun cuando puede reconocer que llevar el tratamiento puede ser la mejor alternativa, prefiere permanecer enfermo "en condiciones naturales".

Por otra parte, Feldman (ibídem) sostiene que, si bien puedes saber con certeza que conectarte puede ser muy placentero, al ser un agente racional y un egoísta bienestarista, sin embargo, puedes rechazar conectarte. Esto sucedería si no crees en el hedonismo, por ejemplo. Supongamos, continua Feldman, que alguien es partidario de algún tipo de perfeccionismo, a saber, la doctrina que sostiene que hay ciertos fines teleológicos por encima de la mera satisfacción inmediata. Esta persona considerará que el bienestar está determinado por el abanico de posibilidades vinculadas a un ideal humano (libertad, verdad, honestidad, etc.). Tal persona estimará que los ideales humanos implican aspectos morales, intelectuales, estéticos, y otros logros. Al estar dicha persona ajustada a esta visión de la vida buena, no le resultará atractiva la VME. "¿Por qué desearía una vida tan miserable?", se preguntaría. "Sería mucho peor mi existencia si me conecto", podría responder, coincidiendo con las intuiciones de Nozick. Ahora bien, el hecho de que alguien no elegiría VME en dichas circunstancias muestra que dicha persona no cree que VME posea un valor de bienestar mayor que VR, pero no demuestra que dicha creencia sea verdadera. Por consiguiente, a juicio de Feldman (ibídem) la persona que se encuentra en el escenario de Nozick no está comprometida con alguna especie de teoría del bienestar antihedonista. Su elección de VR por encima de VME se debe más a su visión acerca del bienestar que al hecho de que la VR sea realmente la mejor vida. Esto concuerda con lo señalado por Tännsjö (1988, p. 65) que la presuposición del hedonismo de la posibilidad de comparar el bienestar de manera interpersonal es inviable dejando a la doctrina hedonista en una encrucijada.

Feldman (2011, p. 71) indica que lo anterior conlleva grandes dificultades. Si se afirma que el sujeto en el escenario de Nozick simplemente no tuviese una visión acerca de lo que constituye una buena vida, se volvería bastante dudoso que tenga una razón para elegir conectarse en una VME o permanecer con VR. Presumible-

mente no tendría bases para pensar que una vida fuese mejor que otra, salvo que se ampare en algún sesgo. Evidentemente, la mayoría de nosotros tenemos un presupuesto axiológico sobre el valor de VR y en esos términos se entiende el rechazo a VME por parte de Nozick. Por ello, las especulaciones de Feldman parecen implicar cierto tipo de nihilismo axiológico. Si afirmamos que el sujeto acepta al hedonismo como una teoría del bienestar, entonces seguramente el sujeto elegiría VME dada la situación que plantea el *puzzle*, a saber, que no existe diferencia cualitativa entre las experiencias placenteras en VR y VME. En este caso, la premisa (1) podría ser falsa. Si sostenemos que este sujeto aceptase alguna teoría no- hedonística del bienestar, entonces debería elegir VR, pero esto no mostraría nada acerca de la verdad del hedonismo, solo su preferencia particular. La elección del sujeto puede ser atribuible a su aceptación de una teoría del bienestar no – hedonística con pretensiones de universalidad y como bien señala Mackie (2000, p. 113) la posibilidad de universalizar no constituye una limitación para el subjetivismo hedonista. Por ejemplo, cualquiera puede ponerse en el lugar de otro sin que ello implique claudicar a sus propios deseos y preferencias.

3.3. Tercer argumento contra el hedonismo ético

Feldman (2011, p. 71) sostiene que una tercera interpretación puede conceder que el sujeto inmerso en el escenario de Nozick es "axiológicamente iluminado", es decir, un sujeto que prefiere un tipo de vida por encima de otro solo en el caso de que una sea realmente mejor para él que la otra. Se presupone que este sujeto tiene cierta capacidad para distinguir el valor intrínseco de ambas vidas. Ello evitaría la posibilidad de que el sujeto pueda rechazar conectarse porque se encuentra en las garras de alguna axiología defectuosa. Ahora podemos reformular el argumento:

Tercer Argumento en contra del Hedonismo Ético

1. Si estuvieses en el escenario de Nozick y supieses, sin duda alguna, que VME podría ser más placentera que VR, y fueses

completamente un egoísta bienestarista, racional, e iluminado axiológicamente, aun así, no te conectarías para obtener una VME. Permanecerías en VR.

2. Si se da (1), entonces VME no tiene un valor de bienestar mayor que VR.
3. Si VME no tiene un valor de bienestar mayor que VR, entonces el hedonismo ético es falso.
4. Por consiguiente, el hedonismo ético es falso.

Feldman (2011, p. 71- 72) considera que esta formulación del argumento contra el hedonismo es más plausible puesto que resultaría difícil imaginar cómo una persona con todos los rasgos descritos puede errar al elegir la mejor vida de bienestar. Su preocupación es obtener la vida que sea la mejor para sí mismo. Es una persona completamente racional y sus preferencias encajan con el valor real (intrínseco) de las vidas posibles que puede elegir. Harman (2000, p. 141) asevera que algo X tiene valor intrínseco en la medida en que X tiene un valor que se debe enteramente a lo que X es intrínsecamente sin considerar las relaciones de X con otras cosas. Lo que nos importa es saber qué haría una persona que es completamente racional, egoísta bienestarista y con conocimiento del valor intrínseco de las cosas en este caso. De esta forma, si el sujeto elige VR, se apoyará en la noción de que VR es la mejor vida, aun cuando pueda tener muchas relaciones y similitudes con la VME (v.g. el carácter vívido de las sensaciones placenteras), el valor de la VR radica en sus cualidades innatas. Por ello, en el escenario de Nozick se rechaza la VME pues intrínsecamente ofrece algo muy distinto a la VR y de menor estima. De esta manera se refutaría el hedonismo. Sin embargo, esta última versión del hedonismo ético resultaría demasiado sofisticada tanto que podría causar resquemor si es que realmente sigue en línea con el hedonismo tradicional, pues se busca una respuesta axiológica al problema del bienestar lo cual llevará a Nozick a plantear, en escritos posteriores, una revisión a nuestros concep-

tos de valor y significado de la vida, en suma, una posición cercana al perfeccionismo ético[9],[10] que se desarrollará más adelante.

Pero una vez que realizamos estos cambios en la descripción del escenario de elección, la plausibilidad de la premisa (1) comienza a desvanecerse, a juicio de Feldman. Aun cuando, en esta tercera versión, se ha reformulado el argumento con el objeto de presentarlo satisfactorio como refutación del hedonismo no resulta tal, pues un hedonista ético que privilegie el acceso a experiencias placenteras y estime a estas como lo intrínsecamente valioso optaría por conectarse. Luego, si Nozick pretendió refutar al hedonismo con su famoso *puzzle*, pues no fue tan certero. Es preciso encontrar una versión del hedonismo que permita asumir como efectiva la crítica de Nozick. Aún queda una tipología de hedonismo que recusaría conectarse a la máquina de experiencias: el hedonismo actitudinal.

4. Hedonismo actitudinal

4.1. Placer actitudinal

Hasta el momento se ha considerado, por razones metodológicas, que los portadores fundamentales del valor intrínseco positivo de placer son episodios en los que una persona siente una sensación placentera. En este sentido, se asume que el "placer" indica algún tipo de sentimiento o sensación. Pero el hedonismo actitudinal entiende el placer como algo diferente: una actitud (Feldman, 2004, p. 55). La distinción entre placer sensorial y placer actitudinal es antigua, aunque a menudo se pasa por alto o se ignora por incomprendida. Una persona que experimenta placer sensorial en un momento determinado no solo siente sensaciones placenteras. Por ejemplo, si le agradan (en el sentido que tiene una predisposición a sentir este tipo de placer sensorial) los sabores del pisco sour y

9 Shea, M. y Kintz, J (2022) sostienen que la posición de Nozick es cercana al perfeccionismo ético amparados en su visión orgánica de la vida y la unidad del valor. Este punto lo analizaremos hacia el final del presente trabajo.

10 Véase también Nozick (1983) en especial el capítulo cinco, Foundations of ethics.

el ceviche, puede experimentar placer sensorial mientras bebe un pisco sour y saborea un ceviche, pero también antes de experimentarlos, planificando, por ejemplo, qué día de la semana degustará estos alimentos sibaritas. El solo hecho de proyectar este tipo de eventos, pensar y vislumbrar la situación, también da pábulo a un tipo de placer.

La cuestión aquí, sin embargo, es que los placeres sensoriales se evidencian de manera sensitiva: cosas incuestionables como las sensaciones de calor y frío; la sensación de presión, las cosquillas y la picazón. Los placeres actitudinales son diferentes. Una persona siente placer actitudinal ante algún estado de cosas si lo disfruta, si está complacido con el mismo. Luego podría experimentar el placer sensorial. Por ejemplo, supongamos que usted es una persona amante de la paz. Asumamos que advierte el hecho de que no hay guerras en curso. El mundo está en paz, es decir sus expectativas pacifistas calzan con las condiciones del mundo. Probablemente esté satisfecho de ello. Se alegra de que el mundo esté en paz. Entonces ha sentido un placer actitudinal ante un hecho determinado: que el mundo está en paz. Los placeres actitudinales siempre están dirigidos a los objetos, del mismo modo que las creencias, las esperanzas y los miedos están dirigido a objetos. Éste es un aspecto en el que se diferencian de los placeres sensoriales. Los placeres sensoriales no solo están dirigidos a objetos de placer, como cuando en un momento *t* disfruta la lectura de una novela porque la lee atentamente en ese momento *t*, sino que también están determinados por estos objetos de placer: usted no puede disfrutar sensorialmente de un pisco sour si no tiene el estímulo de un pisco sour. Otra diferencia es que los placeres actitudinales no necesitan tener ninguna "sensación". Si se da el caso de que no reina la paz en el mundo, ello no impedirá que usted anhele la paz y esté enfocado en alcanzarla. Sabemos que tenemos placeres actitudinales no por una simple sensación, sino de la misma manera que sabemos cuándo creemos en algo, o lo esperamos, o tememos que pueda suceder.

Los placeres sensoriales usualmente están entrelazados con placeres actitudinales, de modo que cuando las personas están contentas por algo, a menudo también experimentan simultáneamente placer sensorial. Por ejemplo, mientras usted bebe un pisco sour, obtiene sabores placenteros (placeres sensoriales), y también puedes sentir placer por el hecho de estar bebiendo pisco sour (una actitud placentera). Independientemente de la frecuencia de este entrelazamiento, los fenómenos son distintos y pueden separarse. Un ejemplo que evidencia que se puede tener placer actitudinal sin placer sensorial sería el siguiente. Imagínese que está padeciendo una intensa sensación de dolor y no siente ningún placer sensorial en absoluto. Suponga que advierte que el dolor se vuelve menos intenso. Usted podría estar alegre de que el dolor empiece a decrecer. Describiría su situación diciendo que adopta actitudes placenteras en el hecho de que su dolor sensorial se está volviendo menos intenso. En esta situación es posible que todavía no sienta ninguna sensación sensorial placentera en absoluto. Incluso aunque esté experimentando un placer actitudinal ante un hecho determinado, lo único que "siente" estrictamente es un dolor disminuido.

Con ello se demuestra que el placer actitudinal es distinto del placer o displacer sensorial. Puedes disfrutar de algo actitudinalmente a la vez que no sientes ningún placer sensorial. Incluso puedes tener fruición por el decremento de un dolor sensorial. Los ejemplos antes descritos sugieren que es posible acceder al placer actitudinal ante algún hecho mientras no se siente ningún tipo de placer sensorial. Los fenómenos son, pues, distintos, aunque no enteramente separables. No es posible que alguien experimente placer sensorial en un momento sin experimentar también cambios actitudinales en términos de placer. Por ejemplo, resultaría controvertido que si a usted le agrada (actitudinalmente) el pisco sour no experimente placer actitudinal al momento de degustar (sensorialmente) una copa de pisco sour. Por ello, Feldman (2004, p. 56) sostiene que podemos definir los placeres sensoriales como sentimientos en los que quien los siente experimenta intrínsecamente placer actitudinal. En otras palabras, lo que hace que una

experiencia sea un placer sensorial es el hecho de que la persona que lo experimenta siente un placer actitudinal intrínseco por el hecho de experimentarlo.

La forma de hedonismo que Feldman pretende defender (2004, p. 57) implica la afirmación de que el placer actitudinal es el bien principal para el hombre incluso por encima del placer sensorial. Feldman señala algunos rasgos característicos del placer actitudinal intrínseco (p. 58):

a. *La conciencia del placer*. Una persona puede sentirse complacida por algo sin ser plenamente consciente del hecho de que está contento por ello. Por ejemplo, si alguien está completamente inmerso en un proyecto de redacción de una novela es posible que se concentre tan intensamente en el proyecto que tiene entre manos que se olvide de todo excepto de la trama y el desenlace. Es posible que esta persona no se percate del paso del tiempo. Es posible que este individuo no esté pensando en sí mismo o en su estado de ánimo. Sin embargo, si más tarde alguien le preguntase, inmediatamente se daría cuenta de que disfruta trabajando en ese proyecto. Por lo tanto, alguien podría sentir un placer actitudinal intrínseco durante un cierto tiempo por el hecho de que esta, por ejemplo, enfrascado en la redacción de una novela, aunque entonces no sea plenamente consciente del hecho de que está disfrutando de todo lo vinculado a la realización de una novela. Luego, si alguien disfruta de algo, señala Feldman (ibidem) inmediatamente esta persona reconocería que está disfrutando de esta actividad y de todo lo que ella implica.

b. *Placer y creencia*. Según Feldman (2004, p. 59), si disfrutas de un determinado estado de cosas, debes creer que dicho estado es verdadero. Por lo tanto, si me complace redactar una novela, entonces debo pensar que estoy redactando una novela. Podría parecer que existen casos en los que una persona se complace en algún estado de cosas en los que no creen. Por ejemplo, una persona ciertamente podría disfrutar de las cosas que observa en una película, o en una novela, o en una obra de teatro, sin creer que esas cosas realmente estén sucediendo. Este tipo de casos (place-

res en la ficción) no ponen en duda la afirmación de que el placer actitudinal implica una creencia. Podemos esclarecer este punto si reflexionamos más cuidadosamente sobre los objetos de nuestro placer en la ficción. Supongamos que usted está disfrutando de la película *Forrest Gump*. Se complace especialmente en una escena en la que Forrest Gump se encuentra con John Lennon. En este caso, ¿cree usted que Forrest Gump realmente se reunió con John Lennon? La respuesta, por supuesto, es "no". Más bien, si alguien está disfrutando de la película lo más probable es que diga algo como esto: "Parece que Forrest Gump podría haber conocido John Lennon. Me divierte y me complace tener esta oportunidad de ver cómo habría sido si Forrest Gump hubiera conocido a John Lennon". Por lo tanto, hay cosas que nos causan placer creerlas aun cuando no sean verdaderas. No es necesario que no disfrutes de nada que no creas que sea verdadero.

c. *La transparencia del placer*. ¿Debe una persona identificar correctamente el objeto de su placer? ¿Podría una persona reconocer que está disfrutando de algo, pero estar confundido o equivocado acerca del estado de cosas en el que se complace? Feldman (2004, p. 60) sostiene que sí. Una persona podría estar experimentando placer en un determinado estado de cosas, se da cuenta de que está disfrutando de algo e identificar erróneamente el objeto de su placer. Por ejemplo, usted podría pensar que disfruta del sabor del Chianti, pues se ha informado sobre este tipo de vino y tiene una punzante curiosidad por paladearlo, pero en realidad lo que está disfrutando es de un Oporto. Por lo tanto, el placer actitudinal no siempre va de la mano con el placer sensorial, y probablemente en algunas ocasiones factores psicológicos añejos a la predisposición del sujeto puedan obnubilar la identificación del objeto de placer.

d. *Placer y verdad*. ¿Podemos disfrutar o complacernos en un estado de cosas que no ocurre? Nuestras formas comunes de hablar de disfrute podrían sugerir que esto es imposible, pero una reflexión más profunda sugiere que las cosas son más complicadas. Supongamos que soy húngaro y he podido leer algunos poemas de *Trilce* y por error creo que han sido publicados en el año 2022 y

pienso erróneamente que conoceré pronto a César Vallejo. Asumamos que estoy encantado con esto ya que es evidente que ello me alegra. Parece incorrecto decir que lo que me alegra es el hecho de que creo que conoceré a César Vallejo. Parece mejor decir que me alegro de que voy a conocerlo, aunque, en realidad, no voy a conocerlo porque César Vallejo está muerto. Luego, algo que es falso podría gatillar un placer actitudinal en el sujeto por lo cual se afirma que el placer actitudinal no implica verdad. Caso contrario ocurre con el placer sensorial donde el objeto de placer no se puede desvincular del placer sensorial.

e. *La pluralidad del placer*. Una persona puede ostentar un placer actitudinal intrínseco en diversos estados de cosas al mismo tiempo. Por ejemplo, una persona puede estar complacida de estar en Perú, de disfrutar su gastronomía, de recibir la calidez de su gente, de visitar los lugares turísticos, de conocer el legado cultural milenario de sus pueblos, etc. Por lo tanto, según Feldman (ibidem) una persona puede disfrutar múltiples placeres actitudinales superpuestos. Pues cuando esta persona planificó venir a Perú pudo haber tomado su decisión por los factores antes descritos o incluso por más.

f. *La orientación temporal del placer*. Existe simpatía con la idea de que los objetos del placer actitudinal deben ser simultáneos con el placer sensorial, mientras que los objetos de deseo deben estar relacionado con el deseo a futuro. Como señala Sumner (2003, p.128), los deseos forman una especie de actitud propositiva, identificada por su intencionalidad y marcada orientación hacia el futuro. Estas características distinguen el placer de otra especie de actitudes como disfrutar algo, encontrarlo placentero o agradable. A diferencia de los deseos, estas actitudes sólo pueden dirigirse a estados temporales: puedo disfrutar sólo de lo que ya tengo (o de lo que ya he disfrutado). Por ejemplo, para tener una actitud placentera hacia la paz, debo, de alguna manera, haber experimentado algunos de sus beneficios. En cambio, el deseo de que haya paz no está ligado a que haya existido esa eventualidad. Un niño judío que hubiera crecido en un campo de concentración nazi probablemen-

te no haya experimentado la paz, pero sí es admisible que la desee. Sin embargo, también es viable que placer actitudinal y deseo vayan de la mano.

4.2. Formulación del hedonismo actitudinal intrínseco

El hedonismo actitudinal diferencia, además, entre placeres actitudinales intrínsecos y extrínsecos, y considera la posibilidad de su cuantificación. Los placeres actitudinales extrínsecos se dan cuando la persona encuentra agrado en un estado de cosas, como por ejemplo en la sensación placentera al paladar de degustar un Chianti en Italia; pero solo debido al placer actitudinal intrínseco que halla en otro estado de cosas relacionado con el primero, por ejemplo, degustar el mismo Chianti en Perú. En este caso, el placer actitudinal intrínseco es hacia degustar Chianti, siendo indiferente, por ejemplo, en qué país se lleve a cabo el acto. Respecto al problema de la cuantificación, Feldman (2004, p. 62) sostiene que los placeres actitudinales se pueden medir de manera análoga a como se hace con los placeres sensoriales. En ambos casos, sostiene Feldman, la cantidad de placer está dada por el producto de dos propiedades de los estados placenteros: la duración y la intensidad. La diferencia entre ambos tipos de cuantificación es que en la determinación de la cantidad de placer actitudinal no se evalúa la duración e intensidad de una sensación placentera, sino la duración y la intensidad de la actitud involucrada (2004, p. 63-66). Por ejemplo, si yo tengo una predisposición prolongada e intensa hacia las bebidas espiritosas como el Chianti en detrimento de otras como el Oporto, se puede evidenciar en una cuantificación de estas preferencias en el tiempo (años, lustros) y en la intensidad (constante, prioritaria, imprescindible). Por lo tanto, la formulación de Feldman del hedonismo actitudinal intrínseco (HAI) sostiene que los portadores fundamentales del valor intrínseco son episodios en los que alguien siente placer o dolor actitudinal intrínseco hacia algo, en lugar de episodios en los que alguien siente placer o dolor sensorial. Así:

Hedonismo actitudinal intrínseco (HAI)

i. Cada episodio de placer actitudinal intrínseco es intrínsecamente bueno; cada episodio de dolor actitudinal intrínseco es intrínsecamente malo.
ii. El valor intrínseco de un episodio de placer actitudinal intrínseco es igual a la cantidad de placer contenido en ese episodio; el valor intrínseco de un episodio de dolor actitudinal intrínseco es igual a la cantidad de dolor contenido en ese episodio.
iii. El valor intrínseco de una vida está enteramente determinado por los valores intrínsecos de los episodios de placer y dolor actitudinales intrínsecos contenidos en esa vida, de tal manera que una vida es intrínsecamente mejor que otra si y sólo si la cantidad neta de valor intrínseco de placer actitudinal en una es mayor que la cantidad neta de ese tipo de placer en la otra. (p. 66):

El HAI es una forma de hedonismo actitudinal puro, ya que implica que los placeres y dolores actitudinales son las únicas cosas que contribuyen de la manera más fundamental al valor de una vida. Supongamos que existe una persona que nunca siente ningún placer actitudinal intrínseco hacia nada. Entonces, pase lo que pase en su existencia, el valor de su vida para él, según el HAI, no puede superar cero. No importa cuánto conocimiento, virtud, honor, riqueza, salud, longevidad, relaciones amorosas, etc. pueda tener, si no se complace actitudinalmente con nada, no hay base para atribuir valor intrínseco positivo a su vida según la HAI. Se puede resumir este punto diciendo que, si el HAI es cierto, una vida sin actitudes intrínsecamente placenteras o dolorosas es una vida sin valor (sin sentido).

El HAI podría definirse como una forma de "estatismo mental". El valor de una vida depende, según esta teoría, sobre hechos de los estados mentales de la persona. Esto se desprende naturalmente de la plausible suposición de que los episodios de placer y dolor actitudinales intrínsecos son estados mentales, previos e independientes de la experiencia sensorial. Una implicación de ello es que, si dos vidas ostentan los mismos estados mentales, entonces son

iguales con respecto a los estados mentales y el tipo de valor relevante que eligen. En otras palabras, si dos personas son indiscernibles con respecto a sus estados mentales (intenciones, preferencias, predisposiciones), entonces el valor en sí mismo para la vida de la primera persona es igual al valor en sí mismo para la vida de la segunda persona. El principio no tiene por qué ser tan amplio. Podemos acotarlo: si dos personas son indiscernibles con respecto a los placeres y dolores actitudinales intrínsecos, entonces sus vidas tienen el mismo valor intrínseco para ellos. Esto último resulta controvertido y la vuelve vulnerable a la crítica.

Lo antes expuesto parece implicar que la experiencia de placer actitudinal es una condición necesaria para la existencia de placer sensorial. Al menos en lo que respecta a su definición del placer sensorial, considera Carrasco (2018, p. 66), la teoría de Feldman puede ser vulnerable a ciertas objeciones. A partir de un ejemplo de Harry Frankfurt, que es presentado con otros fines metodológicos, se puede esclarecer la cuestión. En "Freedom of the Will and the Concept of a Person", Frankfurt evalúa el caso particular de un drogadicto que vive su adicción de manera conflictiva (1971, p. 12). La polémica nace entre dos deseos de primer orden, uno (sensorial) que lo induce a consumir la droga y otro (actitudinal) que lo aparta de su consumo. Esta persona tiene la capacidad de formar ciertos deseos de segundo orden, dirigidos precisamente a encaminar los deseos de primer orden que están en controversia. De esta manera, el adicto consigue alcanzar cierto orden en su estructura volitiva, identificándose solo con el deseo (actitudinal) de no consumir la droga. Sin embargo, puede suceder que, aun identificándose con el deseo (actitudinal) de no consumirla, no logre aplacar el deseo (sensorial) que lo estimula al consumo. Así pues, el caso del drogadicto incontinente que materializa su deseo adictivo muestra que es posible experimentar placer sensorial aun cuando no nos identifiquemos con los deseos cuya satisfacción da lugar a un aumento de este placer sensorial, pues discrepamos, en términos de placer actitudinal, de los mismos. Nuestro drogadicto, desde el punto de vista del placer actitudinal, no deseaba consumir

la droga, pero tuvo placer sensorial al consumirla. Luego, no sería del todo correcto afirmar que el placer actitudinal es condición necesaria para que exista placer sensorial.

4.3. Hedonismo actitudinal intrínseco: el objeto de la crítica de Nozick[11]

Sea:

(a) el hedonismo actitudinal parece ser vulnerable a ciertas objeciones, que provienen de resaltar/ relevar el papel del estado cognitivo (creencias, preferencias) en la determinación del valor de una vida.
(b) la relación que se establece entre los placeres actitudinales y los placeres sensoriales es cuestionable, dado que es posible imaginar casos en los que la experiencia del placer sensorial no implica la emergencia de placer actitudinal.

Basados en los puntos (a) y (b), se puede afirmar plausiblemente que el objeto de la crítica al hedonismo de parte de Nozick con el experimento mental de la máquina de experiencias es una forma de hedonismo actitudinal. Efectivamente, a partir de (a) se colige ciertas limitaciones epistémicas, a saber, que en el hedonismo actitudinal el sujeto puede estar sesgado y dirigido por creencias y estados internos que le hagan preferir incluso creencias no justificadas. Retomando el *puzzle* de Nozick, es plausible imaginar un sujeto que defienda el hedonismo actitudinal cuya máxima sea "la mejor opción a la que debo aspirar es aquella donde puedo maximizar mi placer en VR y reducir mis sensaciones de dolor en VR". Al no haber mayor dictamen que su propio fuero interno, resultaría engorroso un cuestionamiento. Aun cuando la máquina de experiencias le garantiza la adquisición de estas sensaciones placenteras su hedonismo actitudinal lo empujaría a no entrar para siempre, pues por una

11 Si bien Feldman (2011, p. 77-78) considera al hedonismo actitudinal como posible candidato a la refutación del *puzzle* de Nozick estima que no resultaría satisfactorio.

especie de sesgo puede preferir la VR que una desconocida VME. Evidentemente, cualquiera podría objetar esta elección en función de no ser una alternativa sensata en términos de acceso a experiencias sensoriales placenteras, pero ello solo evidenciaría un sesgo epistémico de quien elige, mas no habría posibilidad de injerencia en lo que respecta a su decisión pues, como se ha indicado, las diferentes formas de hedonismo privilegian el fuero de la subjetividad. Por otra parte, la elección tampoco nos dice nada sobre el supuesto mayor valor intrínseco de VR sobre VME.

En lo que concierne a (b) también surgen problemas en torno a la naturaleza de las experiencias placenteras de quien se encuentra en el escenario de Nozick: ¿las experiencias placenteras de quien se encierra en la máquina de experiencias son sensoriales? Según la evaluación de Nozick, no constituirían experiencias reales, en el sentido de que no se corresponden con un estado de cosas del mundo, resultarían una especie de inputs proporcionados por la máquina y que estimularían nuestro sistema nervioso. Sin embargo, como se ha analizado en párrafos previos para el hedonismo actitudinal no es primordialmente relevante el placer sensorial, o su sucedáneo, sino, por el contrario, las disposiciones (creencias, preferencias) para obtener el placer. En el caso del *Gedankenexperiment* de Nozick, es plausible que nuestro hedonista actitudinal rechace conectarse a la máquina de experiencias pues prefiere un vínculo real con el mundo y no las experiencias placenteras privadas que nos ofrece el artilugio. Luego, el argumento de Nozick contra el hedonismo encajaría más precisamente contra el hedonismo actitudinal y vendría dado de la siguiente manera:

1. Si estuvieses en el escenario de Nozick y supieses, sin duda alguna, que VME podría ser más placentera que VR, y fueses un hedonista actitudinal que valora por sobre todo una vida en VR, no te conectarías para obtener una VME. Permanecerías en VR.
2. Si se da (1), entonces VME no tiene un valor de bienestar mayor que VR.

3. Si VME no tiene un valor de bienestar mayor que VR, entonces el hedonismo ético es falso.
4. Por consiguiente, el hedonismo ético es falso.

A partir de (1) se colige que nuestro hedonista actitudinal ostenta ciertas preferencias en su valoración del mundo, pues prefiere la VR a la VME. Aparte, no considera que la mera experiencia privada placentera basta para alcanzar la felicidad ni asume que la mejor manera de alcanzar la misma es preprogramando sus experiencias tal como le ofrece la máquina de experiencias. Respecto a (3) se entiende a partir del contacto con el mundo real que ofrece VR en comparación con VME. Si bien VME puede brindar mayor cantidad de experiencias privadas placenteras, que es lo que usualmente se asume que persigue el hedonismo, el hedonista actitudinal no lo acepta. El presupuesto del hedonismo actitudinal de privilegiar más la disposición ordenadora del placer por encima de su experiencia sensible encajaría en la estructura del argumento de Nozick y otorga plausibilidad a su conclusión sobre el carácter falso del hedonismo. Del mismo modo, esta postura es susceptible de ser enlazada con una perspectiva perfeccionista que se trabajará a continuación.

5. Perfeccionismo ético

A juicio de Hurka (1993, p.17) el ideal del perfeccionismo ético es un ideal moral ya que es una aspiración que las personas deben perseguir sin considerar si lo desean actualmente o lo podrían desear en una circunstancia hipotética futura e incluso más allá del placer que pueda granjear. Es por ello que Bradford (2017) sostiene que si bien tiene pocos defensores posee un ingente número de detractores debido a que en ocasiones se le asocia con una moral impositiva al propugnar el desarrollo de las capacidades humanas características como objetivo ético. Este talante impositivo es el que comúnmente lo aleja de los ideales liberales clásicos. Por ejemplo, Mill, a juicio de Hurka (2003), considera nociva la idea de que ciertos modelos de desarrollo de las capacidades humanas fuesen impuestos. Mill (2014) afirma que las cualidades que son connaturales al ser huma-

no se ejercen con el fin de realizar una elección para fomentar tales elecciones. Amparado en el principio de perjuicio, recomienda, nuevamente, que no se interfiera en el ámbito privado dejando que los individuos libremente escojan las diferentes opciones de vida que deseen[12]. Sin embargo, Hurka (1993, p. 3-5) sostiene que el perfeccionismo parte de la asunción de una vida humana buena que permita el florecimiento de las propiedades naturales del ser humano. De este modo, sostiene Hurka, el perfeccionismo se ampara en tres presupuestos. Primero, que la bondad del ser humano descansa en la propia naturaleza humana; segundo, que valores, estados o acciones como el conocimiento, la amistad, la consecución de un logro son connaturalmente buenos, sin importar sus consecuencias positivas; y, por último, que el perfeccionismo brinda una sistematización de lo que juzgamos como intrínsecamente bueno. Couto (2014, p. 17), en línea con las afirmaciones de Hurka, señala que el perfeccionismo ético se fundamenta en las siguientes premisas:

(1) Existen bienes y actividades que son objetiva e inherentemente valiosos.
(2) La vinculación con estos bienes promueve el bienestar del sujeto.
(3) Estos bienes desempeñan un papel fundamental en establecer qué es lo moralmente correcto.

Nozick (2013, p. 84-85) esgrime tres objeciones por las cuáles no deberíamos conectarnos a la máquina de experiencias que calzan sobremanera con la posición perfeccionista antes esbozada. En primer lugar, afirma que queremos hacer ciertas cosas en el mundo real y no simplemente tener la experiencia privada de las mismas. Queremos ser grandes escritores en el mundo real y no flotar dentro de la máquina de experiencias alucinando que somos conspicuos novelistas. En segundo lugar, de quienes se encuentran en la

12 Buena parte de la tradición liberal ha abrazado la neutralidad del estado en consonancia con la posición milliana. Cf. Couto, A. (2014, p. 2-4); Raz, J. (2003, p. 105).

máquina de experiencias no se puede evaluar con justicia qué tipo de personas son. No sabemos si son valientes, amables, soñadores, continentes. Para Nozick, las experiencias privadas placenteras que proporciona el artilugio no permitirían forjar una personalidad cimentada del sujeto. Esa personalidad se consolida en el mundo real en el contacto con los otros. Por último, la máquina de experiencias nos condena a una vida preprogramada donde no encontramos desafíos ni oposición a nuestros proyectos, donde no experimentamos el aprendizaje del fracaso ni valoramos la resiliencia. Al parecer para Nozick una conexión con la realidad, aun cuando sea negativa como el fracaso, es más valiosa en sí misma que las experiencias placenteras que nos puede brindar la máquina de experiencias.

Podemos preguntar, afirma Nozick (1983, p. 411) cuál es el mejor tipo de persona o el más valioso, cuál es la mejor manera de ser. Recuérdese que entre las objeciones presentadas en el experimento mental de la máquina de experiencias Nozick, se hace énfasis en la importancia de elegir la mejor manera de ser – la más valiosa. También podemos preguntar cuál es el mejor tipo de vida o la más valiosa y cuál es la mejor manera de vivir. Hacer primaria esta pregunta es dar por hecho que lo que deberíamos tener es el tipo de vida más valiosa a lo largo del tiempo. Lo que queremos es, por tanto:

a. Ser el mejor tipo de persona, como
b. Tener el mejor tipo de vida.

Evidentemente, a partir de (a) y (b) comprendemos mejor por qué Nozick rechaza una vida en la máquina de experiencias en favor de la vida real. La máquina de experiencias no nos permite ni (a) ni (b). Aun cuando la vida real no nos garantiza la cantidad de felicidad (interna) que proporciona la máquina de experiencias, sin embargo, nos da la posibilidad de alcanzar una vida más valiosa y lo valioso, según su apreciación, no se reduce al mero placer. Queremos ser el tipo más valioso de persona (1983, p. 413), tener el tipo de vida más valioso, y más aún, que esa vida contribuya de forma correcta

a que seamos ese tipo de persona. Queremos que no solo estén presentes este tener y ser, sino también interconectarlos adecuadamente. Lo mejor es llevar la vida más valiosa, pero a menudo las condiciones externas lo harán imposible o inalcanzable. La persona quiere ser valiosa, no meramente tener posesión de ese valor por un momento como si dependiese de un factor externo. Por ello, busca que lo valioso este íntimamente conectado con su propia vida.

Sher (2004, p. 219- 220) sostiene que las conclusiones a las que llega Nozick en su análisis de por qué no conectarse a la máquina de experiencias constituyen un ataque directo al hedonismo y a las teorías subjetivistas. En particular, al ser de carácter no teleológico, las teorías hedonistas poseen una raigambre subjetiva pues, tanto el placer sensorial como actitudinal, yacen en las propias experiencias. Por otra parte, el perfeccionismo de Nozick a la par que debe ser analizado a la luz de la máquina de experiencias, también puede ser evaluado a partir de su propuesta de la *unidad orgánica* de la vida. Según el propio Sher (2004, p. 222) con esta propuesta Nozick nos muestra cómo el valor de variados aspectos de la vida, actividades y tipos de relaciones que establecemos en el mundo pueden ser plausiblemente atribuidos a la cantidad de *unidad orgánica* que cada uno despliega[13]. Este es un concepto que el propio Nozick (2013, p. 130) toma de la pintura y la biología. Señala que los teóricos de la pintura sostienen que existe valor estético cuando se logra integrar una amplia diversidad de materiales en una unidad de una manera vívida y sorprendente. Se le calificó, sostiene Nozick, *unidad orgánica* porque se asumía que los organismos en el mundo biológico evidencian la misma unidad, pues su diversos órganos y tejidos se interrelacionan para mantener la vida del organismo. Ahora bien, según Nozick, otras áreas exhiben esta *unidad orgánica*, por ejemplo, la ciencia, donde las leyes de Newton son

13 "My tentative conclusion is that the intrisic value of responsive linkage to reality is not a primitive fact or a special one; instead, it folllows from value as degree of organic unity. Reality is the target with which we can establish the greatest organic unity, given our limited imaginative powers and the fact that only in reality can we (actually?) act" (Nozick, 1983, p. 527).

estimadas por su simplicidad, consistencia y capacidad de sintetizar teorías previas, aparentemente disímiles como las de Galileo y Kepler, en una unidad armónica. Si bien el propio Nozick (2013, p. 131) considera una labor ardua mensurar la *unidad orgánica* en sus diferentes grados brinda algunos alcances, intuitivos y algo rústicos, de cómo medirla. Afirma que cuanto mayor sea la diversidad unificada, mayor será la *unidad orgánica* al igual que cuanto mayor sea la capacidad de constreñimiento de la unidad a la diversidad mayor también será la *unidad orgánica*. Por ejemplo, supongamos que he decidido realizar actividades de voluntariado en favor de las personas más necesitadas económicamente. Posteriormente, considero que también es necesario que apoye a las poblaciones más vulnerables, como por ejemplo las mujeres, por las condiciones injustas y deplorables que deben afrontar. Al mismo tiempo, tomo conciencia de la necesidad de comprometerme con la vulneración de los derechos de los animales no humanos pues son seres sintientes que constantemente sufren explotación y maltrato. Estas son actividades disímiles, pero que pueden encontrar una unidad orgánica. Podría sostener que la máxima moral en la vida será la de reducir, en la medida de posible, el dolor y sufrimiento en el mundo. En este sentido, mi máxima otorga unidad a las diversas actividades que realizó, y otras que probablemente realizaré en el futuro, pues da sentido a facetas aparentemente disímiles en la vida de un modo más que significativo. Del mismo modo, si yo asumo que la máxima que le otorgue *unidad orgánica* a mi vida será la de reducir el sufrimiento en el mundo, resultaría incongruente que me declare a favor del maltrato animal, por ejemplo, pues provocaría una contrariedad semejante a la de un cuadro donde no se aprecie armonía entre los elementos.

A juicio de Sher (2004, p. 223) Nozick pretende sostener que diversos aspectos valiosos de la vida de las personas como el conocimiento, la autonomía, la autoevaluación, etc., muestran altos grados de unidad orgánica. Si bien esta propuesta de Nozick es débil al momento de determinar si hay un criterio único para la valoración de la unidad orgánica, Sher considera que más relevante es

reconocer la importancia que atribuye a la relación entre elección y unidad orgánica. Al permitir que el valor de la unidad orgánica dependa de la elección, Nozick introduce un elemento teleológico que amenaza con tragarse el resto de su teoría, enfatiza Sher (2004, p. 224).

6. Conclusiones

El experimento mental de la máquina de experiencias de Nozick no es concluyente en demostrar la falsedad del hedonismo ético, tal como lo sostiene Feldman, en ninguna de los tres primeros argumentos. Pero, si el experimento mental de la máquina de experiencias de Nozick pretende atacar al hedonismo, aunque no refutar categóricamente, será a una versión débil del mismo. A nuestro juicio, esta versión feble es el hedonismo actitudinal propuesto por Feldman.

El escenario que presenta Nozick en este *puzzle* conduce a recusar al hedonista actitudinal a ingresar a la máquina de experiencias, pues al no preponderar las meras experiencias sensoriales placenteras sino una predisposición positiva a vivir en el mundo real no considerará atractiva la VME.

Las objeciones de Nozick a la posibilidad de conectarnos de por vida en la máquina de experiencias lo presentan como un representante del perfeccionismo ético, pues asume que hay formas de ser y vivir más valiosas que otras. De este modo, vivir en el mundo real, aun con fracasos y limitaciones, es mucho más valioso que la vida en la máquina de experiencias.

Referencias

Aristoteles (2014). *Ética Nicomáquea*. Política. Editorial Gredos.

Baier, K. (2004) El egoismo. En *Compendio de ética*. Singer, P. (Ed.)(pp. 281- 290). Alianza Editorial.

Bentham, J. (2000). *An introduction to the principles of moral and legislation*. Batoche Books.

Bradford, G. (2017). Problems for Perfectionism. En *Utilitas*, 29, pp. 344- 364.

Brentano, F. (2009) *The origin of our knowledge of right and wrong*. Routledge.

Carrasco Meza, C. (2018). Bienestar prudencial en la ética de Epicuro. *Ideas y Valores*, 67 (167), pp. 57-80. DOI: http://doi.org/10.15446/ideasyvalores.v67n167.57399

Couto, A. (2014) *Liberal Perfectionism. The reasons that goodness gives*. De Gruyter.

Epicuro (2005) *Obras completas*. Cátedra.

Feldman, F. (2004). *Pleasure and the good life. Concerning the Nature,Varieties, and Plausibility of Hedonism*. Oxford University Press.

Feldman, F. (2011) What we learn from the experience machine. En Barder, R. y Meadowcroft, J. (Ed). *The Cambridge Companion to Nozick's Anarchy, State and Utopia*. (pp. 59- 86). Cambridge University Press

Frankena, W. (1988) *Ethics*. Preentice Hall International, INC.

Frankfurt, H. G. (1971) Freedom of the Will and the Concept of a Person. *The Journal of Philosophy*, 68(1), pp. 5-20. https://doi.org/10.2307/2024717

Harman, G. (2000) *Explaining value and others essays in moral philosophy*. Oxford University Press.

Hurka, T. (1993) *Perfectionism*. Oxford University Press.

Hurka, T. (2003) Perfectionism, selections. En Wall, S.; Klosko, G. (Eds.).*Perfectionism and Neutrality. Essays in liberal theory*. (pp. 125- 133). Rowman & Littlefield Publishers, INC.

Kagan, S. (1998) *Normative ethics*. Westview Press.

Mill, J. S. (2014) *EL utilitarismo*. Alianza Editorial.

Mackie, J.L. (2000) *Ética. La invención de lo bueno y lo malo*. Gedisa.

Moore, G.E. (2005) *Ethics. The nature of moral philosophy*. Oxford University Press.

Nagel, T. (2003) *¿Qué significa todo esto? Una brevísima introducción a la Filosofía.* FCE.

Nozick, R. (1983) *Philosophical Explanations*. Harvard University Press.

Nozick, R. (2013) *Meditaciones sobre la vida*. Gedisa.

Nozick, R. (2017) *Anarquía, Estado y Utopía*. FCE.

Parfit, D. (2004) *Razones y Personas*. Machado Libros.

Raz, J. (2003) The morality of freedom. En Wall, S.; Klosko, G. (Eds.) .*Perfectionism and Neutrality. Essays in liberal theory*. (pp. 103- 124). Rowman & Littlefield Publishers, INC.

Rawls, J. (2010) *Teoria de la Justicia*. FCE.

Ross, W. D. (1930) *The right and the good*. Oxford University Press.

Sidgwick, H. (1962) *The Methods of Ethics*. Macmillan and Company.

Shea, M. y Kintz, J (2022) A Thomistic Solution to the Deep Problem for Perfectionism. En *Utilitas*, 34, pp. 461- 477.

Sher, G. (1997) *Beyond neutrality. Perfectionism and Politics*. Cambrdige University Press.

Sumner, L. W. (2003) *Welfare, Happiness and Ethics*. Oxford University Press.

Tännsjö, T. (1998). *Hedonistic Utilitarianism*. Edinburgh University Press.



Esta obra aspira a ofrecer al público hispanohablante una aproximación a una de las tradiciones filosóficas más influyentes del pensamiento contemporáneo. Los dos volúmenes de Introducción a la Filosofía Analítica ofrecen una primera inmersión integral en esta manera de hacer filosofía, caracterizada por el análisis conceptual, la precisión argumentativa y su estrecho diálogo con las ciencias. En ellos, el lector encontrará aproximaciones a problemas de metafísica, epistemología, ética, filosofía del lenguaje, filosofía de la mente y filosofía de la ciencia. En una época marcada por los desafíos de la inteligencia artificial, comprender la tradición filosófica que moldeó buena parte de sus fundamentos conceptuales deja de ser un ejercicio meramente académico para convertirse en una necesidad intelectual.

ISBN: 978-9907-840-01-8

UNIVERSIDAD CENTRAL DEL ECUADOR
FUNDADA EN 1826
QUITO

Facultad de Ciencias Sociales y Humanas
Universidad Central del Ecuador

EDITORIAL UNIVERSITARIA

filosófica
EDITORIAL