UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
POSGRADO EN FILOSOFÍA DE LA CIENCIA

Cognición, mecanismos y computación. Una investigación sobre la filosofía de las explicaciones científicas de la mente

TESIS
QUE PARA OPTAR POR EL GRADO DE:
MAESTRO EN FILOSOFÍA DE LA CIENCIA

PRESENTA:
ANDRÉS FERNANDO GIRALDO SÁNCHEZ

Directora de tesis:
Dra. Nydia Lara Zavala
Facultad de Filosofía y Letras

CIUDAD DE MÉXICO, NOVIEMBRE 2019

Contenido

AGRADECIMIENTOS

Este trabajo fue escrito dentro del Programa UNAM-DGAPA-PAPIIT “Epistemología de la tecnología y el desarrollo del conocimiento científico”.
Agradezco especialmente a Nydia Lara Zavala por su apoyo generoso y oportuno durante los meses de desarrollo de esta investigación. Agradezco también a Mario Gómez Torrente, cuyo seminario sobre los teoremas de incompleción de Gödel me permitieron comprender de manera más precisa el vínculo entre la teoría matemática de la computación y los problemas filosóficos relativos a la explicación de la cognición. Doy gracias a Atocha Aliseda por su revisión detallada del manuscrito final, así como a Ángeles Eraña y Miguel Ángel Sebastián.
Me siento especialmente agradecido por la generosidad de México, sus instituciones y su gente. Sin el apoyo del programa de Becas Nacionales de CONACYT este trabajo habría sido imposible.
Finalmente, agradezco a mis padres y a mi abuela por su amor y respaldo incondicional.

INTRODUCCIÓN

El surgimiento a finales de la década del cincuenta del siglo pasado de la iniciativa institucional denominada "ciencia cognitiva" representa uno de los últimos episodios en la historia de los esfuerzos por darle estatus científico el estudio de las causas de la conducta inteligente. Algunos comentaristas han interpretado la adopción del marco de referencia en el que tomó forma esta iniciativa como un evento "revolucionario": la revolución cognitiva (Gardner 1985; Boden 2006; Martel y Emeling 1997). A partir de los años cincuenta, de acuerdo con esta interpretación, la investigación científica de la mente habría empezado a efectuarse alrededor de un nuevo "paradigma".

De acuerdo con Thomas Kuhn, en cada momento histórico, los practicantes competentes de una disciplina científica exhiben un acuerdo general sobre cuáles problemas merecen investigación, cómo ha de proceder esa investigación y qué cuenta como una resolución de esos problemas. Este acuerdo general resulta de la adopción implícita de una serie de presupuestos de tipo teórico, experimental y metodológico. Conjuntamente, estos presupuestos conforman un "paradigma". El trabajo efectuado en el seno de un paradigma constituye una forma de ciencia "normal". Cuando en virtud de crisis estructurales en la historia de la ciencia los presupuestos que conforman un paradigma pierden credibilidad es posible sin embargo que nuevos presupuestos se impongan, dando lugar a "revoluciones" (Kuhn, 1996). La idea de que un episodio de este tipo tuvo lugar en el ámbito de la investigación de la mente resulta verosímil cuando se sitúa el surgimiento del "cognitivismo" fundacional de la ciencia cognitiva sobre el trasfondo del conductismo dominante en psicología en Norteamérica en la primera mitad del siglo XX. Cuando el marco de referencia es en cambio el esfuerzo de *integración* del estudio de ciertas capacidades biológicas en la empresa teórica de la ciencia moderna, el surgimiento del cognitivismo puede interpretarse como un episodio de ciencia normal: un capítulo del desarrollo del mecanicismo biológico.

El "cognitivismo" representa la convicción de que la conducta inteligente es la manifestación o el resultado causal de factores internos a ciertos organismos y la estrategia consiguiente consistente en investigar las capacidades cognitivas de estos mediante la postulación de estructuras y procesos "internos". En este sentido, el cognitivismo no supoone una innovación reciente. Desde el siglo XVII al menos, diferentes autores y tradiciones de pensamiento adoptaron agendas de investigación que se proponían desentrañar las estructuras internas responsables de las habilidades características de los organismos inteligentes. Descartes, para empezar, desarrolló un tipo de cognitivismo en términos de la cual propuso hipótesis sobre la naturaleza de fenómenos como la percepción, la memoria y las emociones (Isaac

2019). La verdadera novedad del cognitivismo contemporáneo reside en el tipo de estructuras y procesos internos que postula como explicativos de la cognición.

Desarrollos técnicos en el campo de las matemáticas y progresos tecnológicos fundados en estos desarrollos convencieron a un sector de la comunidad científica en los años de la posguerra de que las capacidades cognitivas en su conjunto podían modelarse y explicarse científicamente en términos mecánicos (algo que Descartes juzgaba imposible). La innovación técnica crucial fue el desarrollo de la teoría matemática de la computación por parte de Alan Turing, Kurt Gödel, Alonzo Church y otros matemáticos en la década de los treinta (Sieg 2009). En particular, un formalismo propuesto por Turing en este contexto para definir con rigor la noción matemática intuitiva de *algoritmo* –las "Máquinas de Turing"– tuvo una influencia decisiva en el desarrollo del cognitivismo.

El formalismo de las Máquinas de Turing inspiró el diseño y construcción de los primeros artefactos físicos capaces de efectuar tareas complejas de computación de manera automática (Copeland 2006). Antes de la construcción de estos artefactos resultaba difícil imaginar la posibilidad de un sistema mecánico capaz de realizar de manera flexible tareas que en el caso de los seres humanos requería inteligencia. La idea general subyacente en el cognitivismo que se impuso por estos años en un sector de la comunidad científica fue que el cerebro es un sistema físico que, al igual que una máquina de Turing universal, posee capacidades de cómputo potencialmente ilimitadas y que, en analogía con este tipo de artefactos, su funcionamiento puede explicarse mediante la descripción de su configuración mecánica "interna". El elemento novedoso en el cognitivismo contemporáneo es así la adopción de un tipo de *computacionalismo* sobre la cognición. El computacionalismo extiende la vieja hipótesis desarrollada por Hobbes y Leibniz de acuerdo con la cual el razonamiento puede modelarse como un tipo de computación en la hipótesis de que *todas* las capacidades cognitivas pueden explicarse como un tipo de computación física.

La hipótesis computacional y la promesa que ofrece de mecanización de la cognición han recibido diferentes interpretaciones desde su formulación en los años cuarenta. Todas las formas de computacionalismo concuerdan en que los cerebros son sistemas computacionales, en el sentido específico de que sus capacidades distintivas pueden *explicarse* como el resultado de la operación de mecanismos de computación. Qué tipo de cosa es un mecanismo de computación, cómo pueden emplearse estos mecanismos para describir el funcionamiento del cerebro y, en particular, en qué sentido pueden ser *explicativos* de sus capacidades son sin embargo cuestiones sobre las que no ha habido consenso.

La interpretación dominante del computacionalismo a partir de los años sesenta, alrededor de la cual se desarrolló la ciencia cognitiva "clásica", afirma que el cerebro está funcionalmente organizado como el *hardware* de un computador digital sobre el que pueden ejecutarse programas diferentes. De acuerdo con esta forma de

computacionalismo "funcionalista", el cerebro procesa algorítmicamente información de acuerdo con secuencias de instrucciones almacenadas en componentes funcionales del mismo, igual que los computadores digitales ejecutan programas almacenados en sus componentes de memoria. La cognición consiste y puede explicarse en este sentido como una forma de procesamiento computacional de información y la "mente", el componente cognitivo del cerebro, puede concebirse como un tipo de "software" (Block 1995). Alrededor de esta concepción de la cognición, la iniciativa institucional de la ciencia cognitiva se articuló a finales de los años cincuenta como un campo interdisciplinar. Esta iniciativa fue etiquetada deliberadamente en forma singular ("ciencia" y no "ciencias" cognitivas) para reflejar el compromiso de sus promotores con el logro de un programa de investigación integrado (Núñez et. al 2019).

La hegemonía y el entusiasmo generado por este programa de investigación duraron relativamente poco. Aunque desde el comienzo hubo voces disidentes, a partir de los años ochenta empezó a hablarse de la disolución de la ciencia cognitiva clásica y de la emergencia de programas alternativos (Bechtel et. al 1998, Boden 2006). Una de las alternativas que empezó a contemplarse fue el uso de formalismos matemáticos diferentes en la descripción de los procesos cognitivos. La teoría de sistemas dinámicos fue empleada para estos efectos a partir de los años noventa (Van Gelder 1995; Faries y Chemero 2019). De acuerdo con el dinamicismo anti-computacionalista que empezó a adquirir notoriedad por estos años, la cognición habría de describirse y explicarse en términos de variables dinámicas vinculadas por "leyes" matemáticas (conjuntos de ecuaciones diferenciales) que caracterizan diferentes tipos de interacciones entre los agentes y el entorno. Estas variables no hacen referencia siempre a factores neurológicos ni en general "internos" a los organismos por lo que el dinamicismo representa también un tipo de anti-cognitivismo.

El dinamicismo se desarrolló en abierta controversia con la forma de computacionalismo fundacional de la ciencia cognitiva clásica. Otra de las alternativas programáticas surgidas a partir de los años ochenta fue el "conexionismo". Al igual que el dinamicismo, este programa surgió en oposición con la plataforma teórica del programa clásico pero, a diferencia de este, el conexionismo no representa una ruptura con el computacionalismo sino una forma diferente del mismo. Los conexionistas se sirvieron de formalismos computacionales diferentes a los usados por el computacionalismo funcionalista y defendieron una concepción diferente de la manera en que este tipo de formalismos podían emplearse en la explicación de las capacidades cognitivas del cerebro (Churchland y Sejnowski 1992; Piccinini 2008c).

Como resultado de la emergencia de estas alternativas al computacionalismo funcionalista y a la ciencia cognitiva clásica y de las respuestas que computacionalistas y clasicistas han propuesto, en las últimas décadas ha tenido lugar

en el ámbito de la filosofía de la mente y de las ciencias cognitivas el resurgimiento y desarrollo de discusiones *fundacionales* relativas a la naturaleza de la cognición y a la forma en que habría de tomar el estudio científico de la misma (Akagi 2017; Marrafa y Paternoster 2012). Estas discusiones hacen referencia a cuestiones que la adopción del programa clásico daba por resueltas y que su socavamiento vuelve a poner en la agenda de los filósofos y teóricos de la cognición. ¿Es el cerebro un sistema computacional? En caso de serlo, ¿qué tipo de sistema computacional es y en qué forma toma la explicación de sus capacidades? ¿Conlleva el computacionalismo un tipo de representacionalismo de acuerdo con el cual las capacidades cognitivas son capacidades de procesamiento de información? ¿Puede la cognición implicar procesos no representacionales? ¿Pueden estructuras corporales ajenas al sistema nervioso ser parte de los factores explicativos de la cognición?

El presente trabajo se sitúa en el contexto de estas discusiones fundacionales sobre la naturaleza de la cognición y su estudio científico; en particular, en el contexto de la reconsideración que la hipótesis computacional ha recibido en las últimas décadas como consecuencia de la emergencia de programas no computacionalistas. Su propósito general es exponer y discutir críticamente el computacionalismo como una hipótesis sobre el tipo de explicaciones científicas adecuadas de las capacidades cognitivas y además los presupuestos filosóficos sobre la ciencia en que el contexto de los cuales esta hipótesis se ha presentado y defendido.

Dos aspectos caracterizan la perspectiva desde la que este trabajo se aproxima a este objeto de discusión. En primer lugar, creo que la historia del computacionalismo puede ser esclarecedora en el debate actual. Reconstruir los debates sobre la hipótesis computacional en el contexto de su propia historia intelectual puede contribuir al progreso en la comprensión de la relación entre los cerebros y los mecanismos computacionales y de la manera en que el uso de las herramientas formales desarrolladas por la teoría matemática de la computación puede ser útil en la explicación de las capacidades cognitivas.

En segundo lugar, este trabajo aborda la discusión del computacionalismo desde la perspectiva de lo que cabe llamar la "filosofía general de la ciencia". La filosofía general de la ciencia es el ámbito en el que la imagen científica del mundo es sintetizada y en la cual la estructura abstracta de la ciencia como actividad teórica humana deviene objeto de investigación (Psillos 2016). El rasgo distintivo de esta disciplina es su orientación meta-teórica: su carácter de teorización de segundo orden sobre una actividad teorizadora de primer nivel. El filósofo general de la ciencia se ocupa de dos funciones diferentes pero complementarias: una función dilucidadora y una función crítica. En el primer caso el objetivo es esclarecer los conceptos meta-teóricos generales que son empleados por las diferentes disciplinas científicas en su tarea de generación de conocimiento (*predicción*, *ley*, *mecanismo*, *explicación*, *teoría*, *experimento*, *modelo*, *hipótesis*). La función crítica en cambio tiene por propósito

evaluar las diferentes propuestas generales sobre la naturaleza de la actividad científica, así como la comprensión de los métodos, recursos y objetivos de la misma. El concepto meta-teórico central del presente trabajo es el de *explicación*. ¿En qué consiste explicar científicamente un fenómeno natural? ¿Qué tipos de cosas son objetos de explicación científica, qué tipo de factores son explicativos de estos objetos y cuál es la forma que toma la presentación de la relación entre unos y otros? ¿En qué medida el tipo de explicaciones ofrecidas por los diferentes sectores de la ciencia representan un factor unificador de la misma? La consideración de estas preguntas y de las respuestas que han recibido en la reflexión filosófica sobre la ciencia en general y en la reflexión metateórica sobre el proyecto de una ciencia de la cognición en particular recorre los diferentes capítulos que conforman este trabajo.

En los dos primeros capítulos presentaré la concepción mecanicista de la ciencia y de las explicaciones científicas desarrollada a partir del siglo XVII como el contexto teórico en el que tuvo lugar el surgimiento del computacionalismo sobre la cognición. La falta de una explicitación meta-teórica clara de los conceptos de *causalidad, explicación* y *reducción* empleados por los mecanicistas motivó, como expondré en el tercer capítulo, el desarrollo de una filosofía de la ciencia anti-mecanicista en términos de la cual se interpretó el computacionalismo a partir de la segunda mitad del siglo pasado. En el cuarto capítulo presentaré en términos generales la interpretación funcionalista del computacionalismo y argüiré que el marco de referencia anti-reduccionista en la que se funda conduce a un tipo de "nihilismo computacional" conforme al cual el uso de modelos computacionales en la investigación de la cognición carece en último término de fuerza explicativa. En el último capítulo expondré el resurgimiento de la filosofía mecanicista en el contexto de las discusiones post-positivas sobre la naturaleza de las explicaciones causales características de las ciencias "especiales". El trabajo realizado por los nuevos mecanicistas, según defenderé, aporta las explicitaciones meta-teóricas ausentes en el mecanicismo clásico y ofrece un análisis de las explicaciones científicas en términos del cual resulta posible eludir las dificultades reseñadas del funcionalismo y justificar el potencial explicativo del uso de modelos computacionales en la investigación de las capacidades cognitivas.

CAPÍTULO 1

1.1. Introducción

Los organismos vivos, entendidos como segmentos del mundo natural, pueden estudiarse de acuerdo con diferentes marcos de referencia teóricos y a diferentes niveles de abstracción: como colecciones de partículas, como sistemas dinámicos, como entidades biológicas, como partes de ecosistemas, entre otros. Una de estas perspectivas pretende estudiar un subconjunto de los organismos vivos como sistemas cognitivos. *Cognición* es una concepto teórico genérico que aspira a designar tanto un conjunto de capacidades peculiares de ciertos de organismos vivos como un nivel de abstracción en el que pueden estudiarse estos organismos en tanto objetos del mundo natural; en particular, en tanto objetos del mundo natural que exhiben las capacidades en cuestión. El estudio de la cognición puede entenderse en este sentido como el estudio de un segmento de la realidad natural desde cierta perspectiva teórica.

La perspectiva que define el estudio cognitivo de ciertos organismos es la que concibe a estos como sistemas o entidades capaces de conducta *inteligente*. Lo que en último término busca establecer el estudio de la cognición es la naturaleza de los factores en virtud de los cuales ciertos organismos pueden comportarse de maneras "inteligentes"; esto es, en términos generales, de maneras adecuadas para alcanzar sus fines dadas sus circunstancias. La búsqueda de estos factores define una agenda teórica: la de la búsqueda de las causas de la conducta inteligente. ¿Puede concebirse esta búsqueda como el objetivo de una disciplina científica? ¿Representa el estudio de la cognición una perspectiva *científica* de estudio de un sector del mundo natural? Responder afirmativamente a estas preguntas supone afirmar la posibilidad de una *ciencia de la cognición*.

De acuerdo con una historia bien conocida, uno de los logros de la ciencia contemporánea fue precisamente el establecimiento de la manera de estudiar de manera científica las capacidades cognitivas. De manera más precisa, el establecimiento de un programa científico de investigación de la cognición. Este pretendido logro de la ciencia contemporánea se entendió como un episodio suficientemente disruptivo en la historia de la ciencia para ser denominado una "revolución": la "revolución cognitiva" (Boden 2006, Miller 2003). Sirviéndose de desarrollos técnicos en el campo de las matemáticas relativos a los conceptos de *computación* e *información* y progresos tecnológicos fundados en estos desarrollos, un grupo de teóricos provenientes de diferentes campos anunciaron a finales de los años 50 del siglo pasado la formación de una nueva disciplina científica. De manera más exacta, estos autores pretendieron dar forma a un nuevo campo interdisciplinario bien integrado, con un objeto de investigación coherente, preguntas

de investigación comunes, y métodos complementarios. Su pretensión era estar dando acta de nacimiento a una "nueva ciencia de la mente": la ciencia cognitiva (Gardner 1985).

El carácter revolucionario de esta nueva empresa estaba vinculado en primer lugar con la posibilidad de integrar el estudio de la mente al proyecto global de la ciencia moderna. Esta era una vieja aspiración que, a pesar de los esfuerzos de diferentes filósofos y científicos desde el siglo XVII, seguía sin cumplimiento al comenzar el siglo pasado. Los intentos de *naturalizar* el estudio de las capacidades cognitivas de los seres humanos y de otros organismos –esto es, de acometer esta tarea de acuerdo con las herramientas y categorías de la ciencia moderna– se oponían a una dificultad que, de acuerdo con una importante tradición del pensamiento moderno, resultaba por principio insuperable; a saber, la tradición vinculada con el dualismo cartesiano.

El dualismo cartesiano es una tesis metafísica según la cual las entidades y propiedades "materiales" conforman un reino ontológico separado de las entidades y propiedades "mentales", que pueden interactuar con aquellas pero son cualitativamente distintas e irreducibles a las mismas. Este dualismo metafísico conlleva, de manera relevante para la presente discusión, un tipo de dualismo *metodológico*. De acuerdo con este último, mientras que los objetos "materiales" o "físicos" son susceptibles de investigación y explicación de acuerdo con las herramientas de la ciencia natural, los objetos "mentales" se resisten a esta teorización y exigen un tratamiento distinto. El estudio de la mente y de sus propiedades habría de constituir así una empresa teórica distinta a la del estudio del mundo natural. Aunque el dualismo cartesiano sea en primer lugar una tesis metafísica, las razones que respaldaron su introducción resultan de consideraciones que podrían denominarse "metacientíficas", relativas a lo que para Descartes constituían los límites de la aproximación científica al estudio de la naturaleza. Descartes, dicho en otros términos, propuso su dualismo metafísico sobre la base de reflexiones acerca de la naturaleza y posibilidades de la ciencia de su tiempo.

En la sección 2 del presente capítulo examino en términos generales el desarrollo de la concepción de las explicaciones científicas como explicaciones mecánicas en la filosofía de Descartes y las razones por las que de acuerdo con este autor la mente y sus capacidades distintivas no podrían ser objeto de explicación científica. Estas razones configuran un desafío teórico a la posibilidad de una ciencia de la mente que llamaré en este y en los capítulos sucesivos el "reto de Descartes". En la sección 3 presento el mecanicismo como una filosofía general de la ciencia. Después de una breve mención de las objeciones de Newton a esta filosofía, en la segunda parte de la sección paso revista a las discusiones teóricas y filosóficas que inspiró la aplicación del mecanicismo en el ámbito de las ciencias biológicas y al desarrollo de una biología mecanicista. En la sección 4 considero dos proyectos no mecanicistas de darle estatus

científico al estudio de la cognición: el introspeccionismo y el conductismo. El fracaso reconocido de ambos proyectos, según indicaré, obedeció a su compromiso tácito con una concepción cartesiana de la cognición y a su consecuente incapacidad de responder al reto de Descartes. La respuesta a este reto exigía una mecanización de la cognición del tipo efectuado por la biología del siglo XIX con respecto a otros fenómenos biológicos. Las herramientas necesarias para esta mecanización, sin embargo, no estaban aún disponibles.

1.2. El reto de Descartes a una ciencia de la cognición

La aproximación científica al estudio de la naturaleza dependía para Descartes de la posibilidad de concebir a esta en términos "mecánicos". La visión "mecanicista" de la naturaleza y de su estudio surgió en la mente de Descartes y de otros científicos de la época en oposición a la visión aristotélica, de acuerdo con la cual existen diferentes tipos de sustancias y de causas en la naturaleza y diferentes tipos de explicaciones de hechos y fenómenos naturales.

Aristóteles concebía la explicación científica de hechos y fenómenos naturales en términos *causales*, como la exhibición de los factores en virtud de los cuales un hecho existe y tiene sus propiedades distintivas. Estos factores eran entendidos como causas (*aitía*) y conformaban diferentes tipos: "materiales", "formales", "eficientes" y "finales" (Falcon 2019). Si bien la explicación satisfactoria de un hecho específico dependía para Aristóteles de la posibilidad de exhibir la manera en que diferentes tipos de causas daban lugar al mismo, su concepción de la ciencia sancionaba como legítimas explicaciones de fenómenos naturales basados meramente en la determinación de causas "finales" (Psillos 2007). Una causa final es "aquello en virtud de lo cual algo es hecho o existe", su "fin" (*telos*). La causa final de una acción es el "fin", entendido como la razón o el propósito, por el cual la acción es efectuada. La causa final de un objeto o de un hecho es del mismo modo el "fin", aquello para lo cual existe. La postulación de causas finales y la proposición de explicaciones basadas en ellas, representaba para Descartes y para los primeros teóricos de la ciencia moderna una práctica epistémica viciosa, que motivaba la formulación de hipótesis infundadas, teóricamente vacuas o imposibles de corroborar. Retrospectivamente puede explicitarse la objeción mecanicista al planteamiento de hipótesis "teleológicas" en el sentido aristotélico como la "objeción de la virtud dormitiva". Esta objeción denuncia la ilegitimidad teórica de hipótesis que identifican las causas de un fenómeno mediante una re-descripción de los efectos que buscan explicarse. ¿Por qué una sustancia específica causa somnolencia? Porque tiene la capacidad de inducir el sueño o está hecha de manera tal que su fin es adormecer. Las hipótesis y explicaciones de este tipo son verdaderas pero vacuas: indican que existe algo que causa un efecto dado pero no aportan ninguna información sobre cuál podría ser esa causa. La nueva

ciencia habría de prescindir en primer lugar de cualquier tipo de explicación teleológica en este sentido.

Para Descartes, igual que para Aristóteles, un componente central de la actividad científica era la búsqueda de causas. A diferencia de Aristóteles, sin embargo, Descartes reconocía únicamente la legitimidad de las llamadas causas "eficientes". Una causa eficiente es el "principio" por el cual un objeto o hecho se genera o produce (como algo opuesto al *fin* en virtud del cual existe). De acuerdo con Descartes, la ciencia habría de recurrir solo a causas eficientes y la aproximación científica al mundo natural consistiría en concebir estas causas de manera "mecánica". Al tratar de explicar o describir un fenómeno, la nueva ciencia procedería a determinar las causas eficientes del mismo en términos de "mecanismos". Un mecanismo era concebido como un sistema físico cuya operación depende de interacciones *locales* entre sus partes o componentes (Milkowski 2018). Al postular un mecanismo como la causa y por tanto la explicación de un fenómeno era innecesario recurrir a ninguna noción teleológica en el sentido aristotélico. Una explicación de un fenómeno es "mecánica" si en la presentación de las causas del mismo recurre únicamente a la postulación de entidades físicas ("cuerpos extensos") y a la interacción directa entre ellos; esto es, en los términos empleados por Descartes, a "movimiento" y "contacto físico directo".

La oposición a la concepción aristotélica de la ciencia cristalizó en una nueva filosofía general de la ciencia que Robert Boyle bautizó tiempo después como "filosofía mecánica" o "filosofía corpuscular" y en el seno de la cual pronto se ofrecieron explicaciones mecánicas de todo tipo de fenómenos, desde la gravedad y el magnetismo hasta la circulación de la sangre (Roux 2018). En oposición a la concepción aristotélica, la nueva filosofía enfatizaba la importancia crucial para el trabajo científico de la experimentación, tanto al momento de formular conjeturas sobre el mundo natural como a la hora de poner a prueba estas conjeturas. Al concebir el mundo natural en términos mecánicos y negar la existencia de fines y propósitos en la naturaleza, la investigación de las causas subyacentes a los fenómenos de interés para la ciencia resultaba por principio realizable. De la misma manera en que los componentes que conforman un artefacto mecánico y los principios que rigen su funcionamiento, por complejo que sea, pueden por principio determinarse, los componentes y principios constitutivos del funcionamiento de los fenómenos naturales podrían determinarse a partir de un trabajo teórico y experimental sistemático. Para efectos del trabajo científico no existiría ninguna distinción fundamental entre el ámbito de lo artificial o mecánico y lo natural. La naturaleza se concebiría como una entidad compleja pero homogénea, regida siempre por el mismo tipo de principios. En analogía con los artefactos mecánicos, la ciencia mecanicista aspiraría a descubrir en la naturaleza principios precisos, regulares y predecibles. La analogía era crucial también para fundamentar la dimensión

experimental de la nueva ciencia. Así como es posible intervenir en los componentes de un artefacto complejo para estudiar la manera en que este funciona, sería posible intervenir en la naturaleza para estudiar su funcionamiento y eventualmente poner a prueba las conjeturas formuladas (Allen 2018).

Para los mecanicistas la interacción entre los componentes de un mecanismo estaba determinada por leyes cuya especificación requería en algunos casos del uso del lenguaje matemático. En este sentido, las matemáticas hacían parte también de la concepción cartesiana-mecanicista de la ciencia como un instrumento indispensable en la descripción y explicación de los fenómenos naturales.

¿En qué sentido esta nueva concepción de la ciencia excluía de acuerdo con Descartes el estudio de la mente?

Descartes estudió en términos mecanicistas algunos fenómenos que hoy se denominarían típicamente "cognitivos" o "psicológicos". En particular, propuso descripciones mecánicas de fenómenos como la percepción, la memoria y las emociones. Al tratar de explicar estas capacidades de manera mecanicista se opuso a los modelos aristotélicos basados en nociones teleológicas y pretendió extender el alcance de la nueva ciencia a un ámbito en el que el recurso a la teleología parecía inevitable. Su esfuerzo puede entenderse como el del establecimiento de un programa de investigación de conductas biológicas complejas por analogía con artefactos mecánicos (Isaac 2019). Descartes aspiraba a representar y explicar el funcionamiento de organismos vivos de la misma manera en que los ingenieros de su época representaban y explicaban el funcionamiento de artefactos mecánicos (Roux 2018).

Todas las explicaciones ofrecidas por Descartes en el curso del desarrollo de este programa están basadas en un modelo mecánico del cerebro y el sistema nervioso. De acuerdo con este modelo, el sistema nervioso es concebido como un sistema hidráulico de túbulos diminutos que vehiculan el movimiento de fluidos especiales denominados "espíritus animales". Estos fluidos son producidos en la sangre y ejercen influencia directa sobre diferentes partes del organismo. Diferentes tipos de movimientos de estos fluidos a través del sistema tienen el potencial de producir conductas complejas; de la misma manera, según Descartes, en la que el flujo del agua a través de conductos ocultos en las grutas de los jardines reales de su tiempo tenía el potencial de producir "conductas" complejas en los artefactos instalados en los mismos (Descartes 2011, p. 683). El principio de operación básico de este sistema es el denominado "arco reflejo": los estímulos ambientales activan mecanismos en el sistema nervioso que causan la apertura de pequeñas válvulas en el cerebro como resultado de lo cual se liberan espíritus animales que viajan hacia diferentes músculos para causar movimientos o conductas diferentes. Este modelo general del sistema nervioso permitió a Descartes conjeturar hipótesis mecánicas específicas de la manera en que, además de la conducta refleja, operan diferentes capacidades que

hoy denominaríamos "cognitivas", como la percepción, la memoria y las emociones. En el *Tratado del hombre*, Descartes atribuye al mecanismo descrito funciones como:

> La percepción de la luz, de los sonidos, de los olores, de los sabores, del calor y de todas las demás cualidades en los órganos de los sentidos externos; la impresión de las ideas de todas estas cualidades en el órgano del sentido común y de la imaginación, la retención o huellas de esas ideas en la memoria, los movimientos internos de los apetitos e inclinaciones o pasiones; y, finalmente, los movimientos externos de todos los miembros… (Descartes 2011, pp. 736).

Estas conjeturas mecánicas no son importantes para los presentes propósitos más que como ejemplos de la manera en que para Descartes la ciencia de su tiempo podía dar cuenta de las causas de conductas complejas en un sistema físico como el cuerpo humano. Esto ejemplos, por otra parte, resultan ilustrativos de las razones por las que estimó que las capacidades cognitivas más distintivas del cerebro humano (o de los seres humanos, entendidos como sistemas físicos) eludían cualquier explicación mecánica y exigían la postulación de un tipo de entidad no material que diera sentido a su existencia. Esta entidad es la "mente" o el "alma", concebida como una sustancia perteneciente a un ámbito ontológico diferente al de los cuerpos extensos y caracterizada por sus poderes "racionales".

Para Descartes, la mente tenía un rol distintivo en la producción de la conducta que no podía reducirse a "movimiento" o "contacto físico directo" y que él asociaba primordialmente con la voluntad y la razón. Dos signos o evidencias conductuales señalaban en su opinión la presencia de una mente en un sistema físico. El primero de estos signos es la capacidad de manipular símbolos significativos, principalmente en el uso del lenguaje:

> Si bien se puede concebir que una máquina esté de tal modo hecha que profiera palabras, y hasta que las profiera a propósito de acciones corporales que causen alguna alteración en sus órganos, como, v. g., si se la toca en una parte, que pregunte lo que se quiere decirle, y si en otra, que grite que se le hace daño, y más cosas por el mismo estilo, sin embargo, no se concibe que ordene en varios modos las palabras para contestar al sentido de todo lo que en su presencia se diga, como pueden hacerlo aun los más estúpidos de entre los hombres (Descartes 2011, pp. 138-139).

El segundo signo conductual de la presencia de una mente en un sistema físico, que puede entenderse como una versión general del primero, es la capacidad de producir conductas "novedosas", entendidas como conductas "apropiadas" a cada circunstancia específica posible en la que pueda encontrarse. Los sistemas mecánicos tenían por definición para Descartes un rango limitado de conductas u *outputs* potenciales, causadas por estímulos o *inputs* específicos. Esta propiedad de los

mismos era justamente la que hacía posible su explicación causal. La mente, en cambio, en virtud de sus poderes racionales, es un "instrumento universal". Esta universalidad es la marca distintiva de la operación de la mente y puede determinarse a partir de tres propiedades diferentes: libertad con respecto a estímulos, infinidad potencial y oportunidad o pertinencia. Esto es, ningún estímulo externo parece desencadenar causalmente una conducta inteligente específica puesto que dado un *input*, un sistema dotado de una mente puede producir un conjunto indefinido y potencialmente ilimitado de *outputs* diferentes; todos los cuales además pueden ser oportunos para la situación en cuestión (McGilvray 2017).

La conclusión de Descartes es categórica: "[De lo anterior se sigue que] es imposible que haya tantas y tan varias disposiciones en una máquina que puedan hacerla obrar en todas las ocurrencias de la vida de la manera como la razón nos hace obrar a nosotros" (Descartes, 2011, p. 139). El resultado es que la mente no es un fenómeno susceptible de estudio a través de las herramientas de la filosofía mecanicista que él, junto con otros como Galileo, Gassendi y Boyle, instituyó en el siglo XVII y que constituye el fundamento de la ciencia moderna. La mente, a diferencia de los fenómenos del mundo natural, opera de acuerdo con principios diferentes, no mecánicos, y queda fuera del ámbito teórico de la ciencia.

Esta conclusión podría denominarse el "reto de Descartes" a una ciencia de la cognición y hasta el siglo XX, a pesar del esfuerzo de diferentes corrientes teóricas, supuso un reto sin respuesta satisfactoria. Antes de exponer las razones precisas que motivaron la convicción por parte de un conjunto de autores en la primera mitad del siglo XX de que finalmente existían los recursos teóricos necesarios para responder al reto conviene sin embargo, con el propósito de situar adecuadamente el contexto intelectual en el que surgió esta convicción, repasar algunos aspectos del desarrollo de la filosofía mecanicista de la ciencia después del siglo XVII.

1.3. El mecanicismo como una filosofía general de la ciencia

La filosofía mecanicista puede entenderse como un programa con dos componentes: un componente metafísico y uno metodológico o epistemológico. El mecanicismo es en este sentido tanto una filosofía de la naturaleza como una filosofía de la manera de estudiar científicamente la misma (Roux 2018). En su sentido metafísico, el mecanicismo representa una concepción materialista del mundo como un ámbito compuesto de entidades con propiedades definidas en virtud de cuya interacción regular tienen lugar procesos constitutivos de sistemas complejos que exhiben poderes causales (Allen 2018). En su sentido metodológico, el mecanicismo representa una tesis sobre la manera adecuada de estudiar y explicar el mundo natural desde una perspectiva científica: los constituyentes del mundo material y su organización pueden representarse para efectos descriptivos y explicativos como

mecanismos. Otra manera de expresar el mismo punto es afirmar que el mecanicismo supone una *heurística* de la investigación científica.

Una heurística es un conjunto de principios y estrategias empleadas en la investigación que indican a los investigadores qué tipo de preguntas plantear y qué tipo de características exhiben las respuestas satisfactorias a esas preguntas. El objetivo de la investigación de acuerdo con esta heurística es ofrecer explicaciones mecánicas de fenómenos naturales de interés científico. Una explicación mecánica muestra el *porqué* de un fenómeno al exhibir el *cómo* de ese fenómeno. Dado que los fenómenos naturales de interés para la ciencia están usualmente vinculados con el "comportamiento" de sistemas físicos complejos, exhibir el *cómo* de un fenómeno implica investigar la manera en que el sistema está compuesto por diferentes tipos de componentes y la manera en que estos componentes interactúan. Antes de poder explicar mecánicamente un fenómeno, dicho de otra manera, es preciso investigar los sistemas físicos en los que tiene lugar mediante dos estrategias complementarias: (a) una estrategia de *descomposición* del sistema en sus componentes y de las tareas efectuadas por el sistema en sub-tareas; y (b) una estrategia de *localización* de cada sub-tarea en un componente del sistema. El cumplimiento adecuado de estas estrategias aporta las herramientas necesarias para exponer cómo un sistema da lugar a un fenómeno; esto es, para explicar mecánicamente el fenómeno (Bechtel y Hamilton 2007).

La exposición de cómo los componentes de un sistema generan un fenómeno de acuerdo con estas directrices tiene lugar a través de la articulación de un *modelo*. Los vehículos de presentación de las explicaciones y en general del conocimiento científico son para el mecanicismo en primer modelos y no lo que en la tradición newtoniana y empirista que discutiré con detalle en el tercer capítulo se denominan "teorías".

Un modelo es una herramienta representacional; un dispositivo simbólico mediante el cual un científico o un grupo de científicos representan aspectos del mundo. El segmento del mundo que es el objeto de la modelación es típicamente un fenómeno o sistema físico. El vínculo entre el modelo y el sistema o fenómeno modelado tiene lugar a través de una *hipótesis teórica*, de acuerdo con la cual el modelo se *parece* al sistema en ciertos grados y respectos. Algunos modelos son modelos *mecánicos*. Este tipo de modelos representan *mecanismos* que se presuponen explicativos del fenómeno o sistema objeto del modelo. En este sentido, estos modelos se oponen a modelos meramente fenoménicos, que se proponen no exhibir las causas de un fenómeno sino solamente describirlo para efectos predictivos o en general heurísticos en la investigación. Los mecanismos que los modelos mecánicos buscan representar pueden entenderse como colecciones de entidades cuyas actividades e interacciones están organizados de tal manera que generan o producen causalmente un fenómeno. Un modelo mecánico comporta así dos componentes: una descripción

de un fenómeno y una descripción de un mecanismo que da lugar o genera ese fenómeno. Un modelo mecánico satisfactorio muestra cómo el fenómeno es causado y está constituido por un mecanismo. La explicación mecanística de un fenómeno consiste por tanto en la determinación de su origen causal a través de la postulación de mecanismos (Glennan 2017).

El éxito reiterado de la heurística mecanicista en la descripción y explicación de diferentes fenómenos del mundo natural, puesta a prueba en procesos de experimentación y validada por las posibilidades que ofrece de manipulación e intervención en el mundo natural, es el factor fundamental que otorga credibilidad a la visión del mundo que representa el componente metafísico de la filosofía mecanicista.

La heurística mecanicista supone por otra parte una concepción general de la actividad científica con implicaciones *normativas*. De acuerdo con esta concepción, ciencia cumple sus propósitos en la medida en que describe la naturaleza en términos de procesos y sistemas en los que un conjunto específico de interacciones localizadas de partes componentes de esos sistemas y procesos conducen o dan lugar a resultados predecibles o regulares. Esta concepción tiene una ambición de generalidad irrestricta: con prescindencia de la disciplina y del fenómeno estudiado, el estudio y la explicación científica de la naturaleza *debe* tomar siempre la misma forma. Una hipótesis o explicación que no se ajuste a estas restricciones metodológicas, que no tome la forma de la explicitación de un mecanismo causalmente responsable de un fenómeno, sería defectuosa o no haría parte de la ciencia. De esta manera el mecanicismo enfatiza el carácter *integrado* de la empresa científica. En contra de las divisiones aristotélicas, la nueva ciencia exhibiría unidad tanto en términos de la naturaleza de su objeto de estudio como en términos de la metodología empleada.

1.3.1. Newton contra la heurística mecanicista

La unidad pretendida de las dimensiones metafísica y epistemológica de la filosofía mecanicista y de la nueva ciencia se reveló sin embargo muy pronto problemática y dio lugar a discusiones que recorren el desarrollo de la ciencia moderna y que, en cierto sentido, como señalaré en los próximos capítulos, siguen presentes en la filosofía de la ciencia actual. En el ámbito de la física, Newton puso de manifiesto que, en contra de uno de los principios fundacionales del mecanicismo, la formulación de hipótesis mecánicas no era necesaria en la determinación de los principios que rigen el comportamiento de los cuerpos físicos. Newton enfatizó el aspecto experimental y predictivo de la concepción mecanicista de la ciencia y articuló una concepción de la explicación científica de acuerdo con la cual esta consiste en la subsunción de fenómenos empíricos bajo leyes universales expresadas matemáticamente y en la cual es innecesaria la determinación de mecanismos causales. Para Newton las

hipótesis y modelos mecánicos podían desempeñar una función heurística en el trabajo científico pero eran en último término prescindibles en la determinación de los principios fundamentales que regían la naturaleza del mundo físico (Janiak 2014, Psillos 2007). Esta desestimación newtoniana de la función de las hipótesis mecánicas tuvo una influencia decisiva en la filosofía empirista de Hume y, como expondré con detalle en el tercer capítulo, en la concepción de la ciencia dominante en el momento del surgimiento de la revolución cognitiva.

Aunque la mecánica newtoniana desplazó a la mecánica cartesiana y se impuso como el marco de referencia de la física moderna, muchos mecanicistas estimaron que constituía una teoría incompleta y que las regularidades y leyes descritas por Newton eran susceptibles de explicación mecánica. Así, a pesar del impacto del trabajo de Newton y del desarrollo incipiente de un tipo de newtonianismo en la comprensión de la ciencia, el mecanicismo siguió representando el marco de referencia de la comprensión de la empresa científica en un segmento importante de la comunidad científica (Boden 2006). Un reto diferente para el mecanicismo provendría sin embargo de otro ámbito de la ciencia.

1.3.2 La explicación mecánica de fenómenos biológicos

Si bien las directrices mecanicistas parecían en el peor de los casos compatibles con la metodología y el conocimiento acopiado en la ciencia física, lo mismo no cabía decir, al menos durante un periodo significativo del desarrollo de la ciencia moderna, con respecto a los fenómenos vinculados con la biología. El creciente reconocimiento a partir de finales del siglo XVIII de la complejidad propia del funcionamiento de los organismos vivos ponía de manifiesto retos que, en opinión de algunos, eran insuperables para la descripción mecánica (Allen 2005).

1.3.2.1. Mereología, reducción y holismo

Aunque desde una perspectiva metafísica materialista pudiera aceptarse que los sistemas biológicos eran sistemas físicos, el tratamiento científico de aquellos imponía dificultades teóricas especiales. Supuesta una metafísica materialista, los principios que rigen los procesos físicos son principios de aplicación universal. Con prescindencia de cualesquiera características distintivas adicionales que exhiba un sector específico de la realidad, su constitución material lo hace parte del conjunto de elementos descritos y explicados por los principios de la física. Este carácter universal de los principios físicos contrasta con el carácter "local" de los principios que, al menos a primera vista, rigen el comportamiento de los sistemas biológicos. Estos sistemas son porciones del mundo material que exhiben complejidad "interna", en el sentido de que están compuestos por subsistemas en diferentes "niveles" de organización cuya interacción obedece a principios específicos en muchas ocasiones no generalizables. Lo que es verdadero de un tipo de célula o tejido, las

generalizaciones teóricas que permiten entender su funcionamiento, por ejemplo, pueden no ser verdaderas para otro tipo de célula o tejido.

Este tipo de complejidad parece hacer ineludible una especie de mereología científica en la explicación de los procesos biológicos. Una mereología de este tipo debe especificar de manera sistemática criterios que permitan entender las relaciones entre las "partes" que conforman el "todo" de un sistema y determinar de manera igualmente sistemática esas partes. Aunque el mecanicismo comportaba una mereología tal, su corrección fue objeto de intensa discusión a partir del desarrollo de la biología desde finales del siglo XVIII. El foco de esta discusión fue el carácter *reductivo* de la mereología mecanicista.

Dado que los sistemas biológicos son sistemas complejos compuestos de subsistemas en diferentes "niveles de organización" –en virtud de lo cual lo que en un nivel puede ser un componente simple en otro puede constituir un sistema en sí mismo–, las directrices mecanicistas toman en este caso la forma de la búsqueda de mecanismos en virtud de los cuales un sistema en un nivel "superior" de complejidad consiste en o *se reduce a* los componentes que lo constituyen. La idea, expresada en términos programáticos, es que el "todo" de un sistema no puede ser más que la suma de las "partes" que lo componen, organizadas de ciertas maneras. La motivación metodológica fundamental de esta estrategia es la convicción de que, aun si en los procesos biológicos el funcionamiento de cada parte es indisociable de sus vínculos con los otros componentes del mismo en diferentes niveles de organización, el "aislamiento" teórico de cada parte es la mejor manera de estudiar sus propiedades y su función en el proceso. El supuesto tácito detrás de esta convicción es que una vez que las características de cada componente aislado son conocidas, su relación con los otros componentes y con el todo del proceso resultaría evidente.

La mereología mecanicista es reductiva por otra parte dado que en los procesos biológicos las partes y los mecanismos en virtud de los cuales estas se organizan son concebidos como entidades y procesos de naturaleza química y física. La idea subyacente era que la comprensión del funcionamiento de los organismos depende del reconocimiento de su descomposición en sistemas de órganos; la comprensión del funcionamiento de los órganos del reconocimiento de su descomposición en tejidos; la de estos, de su descomposición en células; la de estas en moléculas, hasta alcanzar los niveles fundamentales de organización de la materia (Allen 2018).

No resultaba claro sin embargo cómo estas directrices metodológicas, a pesar de su atractivo metafísico, permitían dar cuenta de ciertas características y capacidades distintivas de los organismos vivos, que carecían de equivalencia en el resto del mundo físico: la auto-replicación, la respuesta discriminada a los estímulos ambientales, capacidades de autorregulación sofisticadas y de eficiencia en la transducción de energía. Al igual que Descartes no reconocía la posibilidad de que un sistema mecánico diera lugar a capacidades mentales complejas y "universales",

diferentes tradiciones de pensamiento en la biología anterior al siglo XX no advertían la manera de dar cuenta mecánicamente de estas características de los organismos vivos y vinculaban la insistencia de los partidarios de la estrategia mecanicista con una concepción simplista e ingenua de los procesos biológicos y de los organismos vivos. Así, mientras que Newton sostenía que el instrumental descriptivo de la filosofía mecanicista era innecesario para dar cuenta de una realidad que obedecía a principios más simples y uniformes de lo que los mecanicistas creían, las críticos provenientes de la biología estimaban que el instrumental descriptivo del mecanicismo era insuficiente para dar cuenta de una realidad que se revelaba cada vez más compleja.

Esta percibida insuficiencia de la ciencia mecanicista, en un movimiento reminiscente de los razonamientos que llevaron a Descartes a postular su dualismo metafísico, estimuló el desarrollo de doctrinas anti-mecanicistas que, para dar cuenta de las características distintivas de los organismos vivos, postulaban entidades y procesos que obedecían a principios no mecánicos. Estas doctrinas tuvieron especial protagonismo en biología desde mediados del siglo XIX y pueden agruparse de manera retrospectiva bajo la etiqueta de "holismo" (Allen 2005).

Lo que vincula a las diferentes variedades del "holismo" es el rechazo de las directrices metodológicas del mecanicismo en virtud de su percibido vínculo con una visión reduccionista e ingenua de los organismos. Dadas sus características distintivas, los organismos vivos no podrían entenderse a partir de procesos mecánicos –en particular, de procesos químicos simples– y debían estudiarse no como una agregación aditiva de componentes sino como "totalidades". Los holistas pusieron especial énfasis en el carácter local e idiosincrásico de los principios que regían el comportamiento de los sistemas biológicos en cada uno de sus niveles de organización. En su opinión, este carácter local e idiosincrásico hacía que estos principios fueran impredecibles a partir del conocimiento de los principios que regían otros niveles y en consecuencia irreductibles a los mismos (Allen 2018).

La versión más radical del holismo, que cabe concebir como la versión del dualismo cartesiano en la biología, es el vitalismo. Los vitalistas estimaban que las dificultades del mecanicismo eran por principio insuperables en el estudio de los organismos vivos. Para los vitalistas, los organismos vivos desafían cualquier descripción en términos puramente físico-químicos. La caracterización de los mismos, en su opinión, exigía la postulación de fuerzas no materiales y no mensurables. La propiedad distintiva de estas fuerzas sería una dimensión teleológica irreducible mecánicamente. Es en virtud de esta teleología intrínseca que tales fuerzas podrían dar cuenta de la complejidad de los organismos. El vitalismo representa así también el resurgimiento de estrategias aristotélicas de explicación en el ámbito de la biología.

El problema con el vitalismo, al igual que con la concepción aristotélica de la explicación, es que fomenta y legitima prácticas epistémicas viciosas de acuerdo con los estándares de la ciencia moderna. Las hipótesis en el sentido de que diferentes sistemas vivos están organizados por entidades y fuerzas no físicas y no químicas son objeto de la "objeción de la virtud dormitiva" mencionada antes: determinan una pretendida causa pero no aportan información sobre la naturaleza de esa causa que permita poner a prueba empíricamente la hipótesis. Las hipótesis vitalistas, en general, no son susceptibles de prueba experimental ni de corroboración empírica, por lo que carecen de valor científico.

El vitalismo representó sin embargo nada más que una respuesta extrema y descaminada a las limitaciones percibidas en las directrices mecanicistas con respecto al estudio de los procesos biológicos. Las variedades menos radicales del holismo se mantuvieron en último término dentro de las márgenes de la concepción mecanicista de la ciencia. De manera retrospectiva, es posible interpretar el holismo biológico del siglo XIX no como un movimiento revolucionario sino reformista del mecanicismo. Su trabajo puso de manifiesto la inadecuada simplicidad de los modelos mecánicos iniciales desarrollados por Descartes y los científicos mecanicistas durante los siglos XVII y XVIII (basados en analogías con artefactos simples como bombas de agua, relojes y molinos) de cara a la creciente complejidad conocida de los procesos biológicos que subyacen al funcionamiento de la vida. El énfasis holista en la complejidad de los organismos en sus diferentes niveles de organización y en los principios de organización que rigen esos niveles fomentó la división y la especialización del trabajo en el seno de las ciencias biológicas y en último término contribuyó al desarrollo de la ciencia mecanicista. A pesar de que diferentes aspectos del pensamiento holista, en particular su énfasis en la incorrección de las ambiciones reduccionistas del mecanicismo, persistieron en la filosofía de la ciencia del siglo XX, cabe decir que el holismo biológico del siglo XIX terminó siendo subsumido por el mecanicismo.

1.3.2.2. El triunfo de la biología mecanicista

La creciente especialización y sofisticación del trabajo en el estudio de los procesos biológicos produjo a partir del último tercio del siglo XIX grandes logros para la ciencia mecanicista. Claude Bernard en Francia y Ludwig von Helmholtz y Emil du Bois-Reymond en Alemania le dieron estatus científico al campo de la fisiología al efectuar contribuciones decisivas en la comprensión de la naturaleza de los mecanismos que rigen el funcionamiento del páncreas y el hígado, en un caso, y del sistema nervioso, en el otro. En este último caso, el trabajo experimental desarrollado por los miembros de la escuela de Berlín vinculados con von Helmholtz y du Bois-Reymond sentó las bases del desarrollo de la neurobiología que décadas más tarde encontraría fundamentos mecanísticos sólidos en el trabajo de Santiago Ramón y

Cajal (Allen 2018). El logro fundamental del mecanicismo en el ámbito de las ciencias biológicas en el siglo XIX provino sin embargo del desarrollo de otra disciplina: la embriología.

Ningún fenómeno resultaba más problemático para la biología del siglo XIX y para el desarrollo del programa mecanicista que la fecundación. El proceso de fertilización de un óvulo por un espermatozoide y la sucesión subsiguiente de eventos de desarrollo embrionario que se desencadena a partir de la formación del nuevo cigoto representaba un bastión de especulaciones vitalistas y metafísicas. El evento de iniciación de un nuevo organismo supone después de todo un criterio de distinción fundamental entre la materia viva y la no viva y su comprensión suponía un reto decisivo para las aspiraciones teóricas de la biología (Allen 2018). Entre finales del siglo XIX y principios del XX, el fisiólogo alemán Jacques Loeb desarrolló un programa de trabajo experimental mediante el cual demostró que la fecundación en los erizos de mar podía inducirse de manera artificial a través de estímulos físico-químicos. El trabajo de Loeb mostró que los científicos podían manipular materiales químicos en un laboratorio para inducir las etapas iniciales de un organismo vivo. Su objetivo fundamental era probar que la vida podía entenderse empleando los conceptos de la física y la química. De manera más precisa, Loeb quería demostrar que los procesos biológicos asociados con el surgimiento de la vida seguían principios físicos y químicos y que a través de investigación experimental estos procesos podían ser manipulados. Para Loeb, el carácter científico de la biología dependía de su éxito en la empresa de ofrecer explicaciones a través de la postulación de mecanismos cuyos componentes y principios de organización son físico-químicos (Loeb 1912). Su propio trabajó demostró que esto era posible y supuso un impulso decisivo al desarrollo de la concepción bioquímica de la vida, uno de los últimos grandes logros de la ciencia mecanicista. Con el desarrollo de este marco de referencia de comprensión de la vida se cumplía un viejo anhelo mecanicista: la unificación de los principios científicos y de la comprensión de los procesos orgánicos e inorgánicos de la materia.

Esta conquista intelectual puso así de manifiesto que, a pesar de los reparos de los holistas del siglo XIX, los procesos biológicos fundamentales eran susceptibles de explicación en términos mecánicos y físico-químicos y que la biología no difería por tanto de la física y la química ni en la naturaleza de sus objetos de estudio ni en la naturaleza de los recursos metodológicos y conceptuales de los que se servía en su trabajo. En las primeras décadas del siglo XX estos avances en la biología (aunados al desarrollo de la mecánica cuántica, que ponía de manifiesto la unidad imprevista de la física y la química) estimularon la reaparición en el panorama intelectual del ideal mecanicista de una ciencia "unificada" (Chomsky 1997).

Un obstáculo, sin embargo, se oponía aún en el camino del cumplimiento de este ideal: el reto de Descartes a una explicación científica de las capacidades cognitivas superiores.

1.4. Dos proyectos de una ciencia no mecanicista de la cognición

El reto de Descartes representa la manifestación de una perplejidad teórica: ¿cómo es posible que las capacidades cognitivas del cerebro –un tipo de sistema biológico– surjan en un mundo constituido por procesos físico-mecánicos? Descartes pensaba que era imposible y esta convicción motivó su proposición de un dualismo metafísico. De acuerdo con este dualismo, la mente es la sustancia en la que el pensamiento "reside" (la mente es "sustancia pensante") y el rasgo esencial del pensamiento es su accesibilidad "inmediata"; esto es, para Descartes, su accesibilidad mediante reflexión o introspección (Descartes 2011, pp. 270-271). En consecuencia, el ámbito de lo mental se identifica para Descartes con el ámbito de las experiencias subjetivas accesible a través de introspección. Algunas de estas experiencias, por otra parte, tienen contenido representacional –se refieren a estados de cosas "externos" a la mente– y constituyen el punto de acceso de la mente a la realidad y una de sus fuentes fundamentales de conocimiento del mundo.

La tesis metafísica de Descartes y su caracterización de la mente motivó una serie de acertijos relativos a la manera en que esta sustancia pensante puede alcanzar conocimiento del mundo material. Estos acertijos definieron un segmento importante de la agenda teórica de la filosofía moderna (Rorty 1979). El trabajo desarrollado en esta tradición (salvo contadas excepciones) no tuvo ambiciones de cientificidad y se entendió en cambio como un intento de "fundamentar" desde una perspectiva exterior a la ciencia todo el conocimiento humano, incluyendo el científico. De acuerdo con la convicción de Descartes, las investigaciones sobre los principios que rigen el funcionamiento de la mente y de sus capacidades cognoscitivas se efectuó de manera no científica. Las tesis resultantes de estas indagaciones se tomaban como tesis *sui generis*, en cuya evaluación no intervenían consideraciones empíricas ni experimentales.

1.4.1. Introspeccionismo

A finales del siglo XIX, sin embargo, en medio del clima de optimismo intelectual generado por algunos de los logros científicos descritos en la sección previa, el fisiólogo alemán William Wundt quiso darle estatus de cientificidad a esta tradición de investigación de la mente. En contra de las directrices cartesianas, Wundt pensó que las experiencias subjetivas eran susceptibles de estudio empírico y fundó un laboratorio con el propósito de investigar experimentalmente los principios que pretendidamente las regían (Boden 2006). Wundt y sus seguidores estaban interesados en determinar los elementos y las estructuras básicas de las experiencias subjetivas, de la misma manera, en su opinión, en que la ciencia química de la época buscaba determinar los elementos y estructuras básicas de los compuestos químicos. De manera poco sorprendente, el método que dirigía estas investigaciones era la introspección. Wundt pensaba que podía lograr sus objetivos entrenando a

individuos para que, en el curso de pruebas experimentales cuidadosamente elaboradas, analizaran sus propias experiencias, determinaran sus componentes y estructuras básicas y elaboraran reportes de las mismas. En una de estas tareas, por ejemplo, se le presentaba a un grupo de participantes el nombre de dos animales, se les pedía juzgar cuál de los dos era típicamente mayor en tamaño y luego se les asignaba la tarea de elaborar reportes introspectivos detallados de lo que en su opinión había ocurrido en sus mentes en el intervalo entre la presentación de los nombres y la producción del juicio (Weisberg y Reeves 2013). De esta manera, se presumía, podrían elaborarse hipótesis y modelos empíricos sobre la manera en la que la mente desplegaba sus capacidades.

A pesar de grandes esfuerzos invertidos en estas investigaciones, ni Wundt ni ninguno de sus seguidores pudieron acopiar un cuerpo significativo de resultados experimentales generalmente aceptados y la ambición de cientificidad de este programa de investigación fue pronto duramente cuestionada. Dos problemas centrales aquejaban al programa. En primer lugar, la formulación de hipótesis suponía una tarea relativamente sencilla pero su confirmación o contrastación parecía imposible. Sobre la base de una serie de reportes introspectivos recogidos en pruebas experimentales del tipo descrito un investigador podía llegar a una hipótesis *A* relativa a la manera en la que la mente realiza alguna de sus tareas. Esta tarea se entendía como la búsqueda de patrones en los procesos mentales descritos por los sujetos experimentales y no estaba sujeta a protocolos estrictos. Ahora bien, sobre la base de una serie distinta de reportes introspectivos recogidos en instancias diferentes de la misma prueba un investigador diferente podía llegar a una hipótesis *B* sobre el mismo fenómeno mental. ¿Cómo decidir entre ambas hipótesis? Los partidarios de este programa confiaban en que pruebas adicionales con sujetos mejor entrenados en la capacidad de la introspección llevarían a reportes convergentes pero esta confianza nunca se vio de hecho reflejada en los resultados presentados. Ni Wundt ni sus seguidores pudieron en último término ofrecer controles ni criterios estrictos sobre las tareas de observación introspectiva de manera que la utilidad y legitimidad científica de las mismas pudiera verificarse (Martel Johnson 1997).

El segundo problema que aquejaba al programa es de naturaleza más conceptual. Como los críticos del programa enfatizaron pronto, muchos tipos de procesos y fenómenos mentales no están asociados a experiencias conscientes y no son accesibles mediante introspección. Un individuo no puede, por ejemplo, meramente prestando atención a sus experiencias subjetivas, distinguir de manera fiable alguna de las creencias –o actitudes proposicionales– que interviene en la economía de su vida mental de otra. La razón es que en general estas actitudes no están vinculadas con experiencias subjetivas. Por otra parte, un individuo no puede acceder introspectivamente a los procesos en virtud de los cuales su mente efectúa, por ejemplo, operaciones aritméticas, de procesamiento lingüístico o de recuperación de

recuerdos. Así, este programa, que aspiraba a que mediante estudios experimentales basados en la introspección pudiera desentrañarse la manera en la que la mente funciona, parecía dejar por principio fuera muchas de sus capacidades distintivas. La mente entendida como el asiento de las experiencias subjetivas no parecía a fin de cuentas un objeto susceptible de estudio científico.

1.4.2. Conductismo

El fracaso del programa introspeccionista se interpretó como evidencia de la inviabilidad de cualquier aproximación "mentalista" que intentara explicar la conducta inteligente mediante la postulación de procesos y fenómenos "internos". La ciencia buscada debía seguir el ejemplo de las ciencias establecidas y, de acuerdo con una popular interpretación de lo que estos ejemplos dictaban, enfocarse nada más que en fenómenos observables y directamente mensurables y manipulables. En el primer tercio del siglo XX emergió así un programa diferente que prometía dar estatus de cientificidad a la psicología: el conductismo.

De acuerdo con los conductistas, la ciencia psicológica debía ocuparse únicamente de dos variables: los estímulos ambientales que afectan a un organismo y sus respuestas conductuales. Cualquier alusión a procesos y entidades "mentales", "internas", quedaba proscrita de las hipótesis de la nueva ciencia, cuyo objetivo expreso era la predicción y el control de la conducta. La noción de *conducta*, por otra parte, debía entenderse en términos no mentalistas, como un conjunto de eventos físicos cuya descripción e individuación no requería aludir procesos internos al organismo sujeto de la conducta (Harnish 2002). El foco de atención de los conductistas en el comportamiento estaba inspirado en el modelo del "arco reflejo", entendido en este caso como la relación directa, no mediada por ningún "procesamiento interno" teóricamente relevante, entre los estímulos ambientales de un organismo y sus respuestas a los mismos. La conducta, así, debía entenderse en último término como un tipo de "reflejo". Aunque las versiones menos radicales del conductismo no negaban la existencia de procesos internos que median entre estímulos y conducta, se suponía que estos procesos podían tratarse como una "caja negra" y carecían de importancia tanto teórica como práctica (Martel Johnson 1997). La "mente", en caso de que hiciera falta usar el concepto por razones prácticas, debía entenderse no como algo "oculto" e "interno" sino como algo "abierto" y "externo", y sus propiedades podían conocerse mediante observación ordinaria y no a través de observación introspectiva.

La convicción que subyacía al programa conductista es que existían leyes o regularidades no accidentales que regían el vínculo entre estímulos ambientales y respuestas conductuales. La idea central detrás de esta convicción era que cada evento conductual estaba asociado de manera no accidental con un estímulo detectable y mensurable empíricamente y que a su vez cada estímulo producía de

manera no accidental un repertorio preciso de respuestas conductuales. Dado esto, debía ser posible analizar la conducta en un grado tal de precisión que, para cada evento conductual específico, el psicólogo pudiera determinar de manera exacta el estímulo o conjunto de estímulos con los que estaba vinculado etiológicamente y que, a su vez, para cada estímulo específico pudiera determinarse –esto es, predecirse– qué tipo de respuesta conductual generaría (Weisberg y Reeves 2013). Otra manera de entender el objetivo de la nueva ciencia psicológica anunciada por los conductistas era, de acuerdo con esto, la de determinar experimentalmente las causas ambientales ("externas") de la conducta. En los términos preferidos por B. F. Skinner, uno de las figures más prominentes de este programa de investigación, el organismo debía entenderse como un "espacio de variables": un lugar donde variables ambientales previas ("variables independientes") producen conducta ("variables dependientes") (Harnish 2002). El estudio experimental de las relaciones funcionales entre estas variables es el ámbito de la psicología conductista.

El "método" de la psicología entendida en términos conductistas puede resumirse en un programa de seis pasos (Bechtel et al 1998): (1) Obsérvese la conducta de un organismo; (2) Selecciónese descripciones no mentalistas de la misma; (3) Selecciónese descripciones igualmente no mentalistas del entorno en el que tiene lugar la conducta (en el sentido de descripciones que no supongan capacidades mentales en el organismo); (4) Obsérvese que ciertos aspectos no mentales de la conducta (como, en particular, su frecuencia de ocurrencia) están correlacionados con estímulos ambientales presentes en el entorno de la conducta; (5) Manipúlese y modifíquese de manera experimental variables de los estímulos ambientales recibidos por el organismo y determínese así de manera precisa la clase de eventos ambientales y la clase de conductas cubiertas por la correlación; (6) Conclúyase que la conducta es una función del entorno: la relación entre los estímulos ambientes y las respuestas conductuales es una relación funcional.

Mediante la aplicación de estas directrices los conductistas pretendían explicar las diferentes capacidades características de los organismos inteligentes como tipos de hábitos adquiridos a través de procesos de *condicionamiento* externo. La idea básica era que conductas inicialmente aleatorias en el organismo suponen consecuencias provechosas o nocivas para el bienestar del mismo y que estas consecuencias determinan el crecimiento o disminución en la frecuencia e intensidad de ese tipo de conducta. Al largo plazo las conductas recompensadas ("reforzadas", en la jerga conductista) por el ambiente devienen *hábitos*. El arsenal de hábitos adquiridos de esta manera en el intervalo de desarrollo de un organismo constituyen sus capacidades "cognitivas". El lenguaje, por ejemplo, era concebido por Skinner en términos conductistas como un conjunto de hábitos adquiridos por medio de condicionamiento (Flanagan 1991). Todo lo que hay "en la mente", como se diría en idioma mentalista, es el resultado de procesos de condicionamiento y proviene así de

la interacción entre el organismo y su entorno. Y toda conducta estable, que es en último término el objeto de interés de la psicología, puede asimismo explicarse como resultado de la historia de la relación del organismo con su entorno. El objetivo de los psicólogos, los científicos de la conducta, es proponer descripciones, fundados en un trabajo experimental cuidadoso y susceptibles de corroboración empírica, de la manera en que esta fijación de hábitos tiene lugar en los organismos. Para los conductistas, todo lo que importaba en este trabajo era la determinación de regularidades funcionales no accidentales y, en la medida en que estas dieran lugar a predicciones acertadas y permitieran intervenir y controlar la conducta, cualquier recurso a hipótesis causales adicionales relativas a mecanismos internos en los organismos era no solo innecesaria sino anti-científica. Tal era la lección aprendida del fracaso del programa introspeccionista. El conductismo representó así un tipo de Newtonianismo, de acuerdo con el cual la postulación de hipótesis mecánicas no es un componente esencial del estudio científico de la conducta (Boden 2006).

Las directrices de investigación conductistas podían aplicarse no solo en el estudio de la conducta de los seres humanos sino de cualquier organismo que exhibiera una respuesta discriminada al ambiente y motivara en ese sentido una atribución de inteligencia y capacidades cognitivas. Para los conductistas, el comportamiento humano difería del comportamiento del resto de animales únicamente en virtud de su mayor complejidad; esto es, únicamente en virtud de la mayor complejidad de las relaciones funcionales que intervienen en su descripción, predicción y control (Harnish 2002). Esta característica del programa era percibida como una virtud, en la medida en que aportaba un tratamiento unificado de la conducta inteligente y contribuía así a poner en contacto a la psicología con la biología y el resto de las ciencias naturales.

La concepción conductista de la psicología ejemplificaba así una serie de virtudes que se esperaban de una ciencia de la mente en la atmósfera de optimismo intelectual de finales del siglo XIX y principios del XX, todas las cuales pueden entenderse en relación con el ideal de una ciencia unificada. A diferencia del programa introspeccionista, el conductismo enfatizaba la continuidad metodológica de la psicología con respecto al resto de las ciencias naturales: la nueva ciencia procedería de acuerdo con observaciones empíricas, trabajo experimental basado en protocolos rigurosos y formulación de hipótesis y modelos fundados experimentalmente y susceptibles de corroboración empírica. Por otra parte, el conductismo, también en contraste con el programa introspeccionista, mostraba la continuidad de la psicología con el resto de las ciencias también en términos de su objeto de estudio. La nueva ciencia se ocupaba de un segmento bien delimitado del mundo natural, estudiado desde otras perspectivas y niveles de abstracción por diferentes ciencias, y no de oscuras entidades y procesos subjetivos, aislados de la jurisdicción teórica del resto de la empresa científica. La preservación de estas virtudes sería en adelante un

desiderátum de cualquier programa de investigación científica de la mente y la conducta inteligente.

El conductismo no pudo sin embargo superar el estatus de promesa teórica y convertirse en un programa de investigación con resultados robustos. Aunque el trabajo experimental con animales no humanos fructificó parcialmente en la articulación de regularidades funcionales que permitían modelar y predecir algunos rasgos interesantes de su conducta, los conductistas nunca pudieron acopiar un conjunto de leyes funcionales que dieran cuenta de las capacidades cognitivas distintivas de los seres humanos, como el lenguaje, la memoria, el aprendizaje o el razonamiento práctico. Ni siquiera en el ámbito del estudio experimental del comportamiento de animales no humanos, el bastión teórico más prometedor del programa, los resultados fueron invariablemente favorables a las convicciones conductistas. Una serie de estudios experimentales desarrollados en los años treinta por el psicólogo conductista Edward Tolman con ratones arrojó resultados ruinosos para el programa.

Un principio nuclear del conductismo era que todo aprendizaje es el resultado de *condicionamiento*, entendido como un proceso de reforzamiento de conductas favorables para el organismo. Tolman mostró que los ratones exhibían una capacidad que denominó "aprendizaje latente" (Bermúdez 2014). El objetivo de sus estudios era investigar la manera en que los ratones aprenden a dirigirse en laberintos artificiales. En uno de los experimentos ideados por Tolman, un conjunto de ratones era dividido en tres grupos diferentes y situados en momentos diferentes en un laberinto cuya configuración podía modificarse de diferentes maneras. El primer grupo de ratones recibía una recompensa cada vez que podía sortear exitosamente el laberinto, el segundo grupo nunca recibía una recompensa y el tercer grupo empezaba a recibir una recompensa solo diez días después de empezar a dirigirse en el laberinto. En un primer momento, como la doctrina conductista predecía, el grupo inicial de ratones aprendió a sortear el laberinto rápidamente, mientras que los otros dos grupos exhibieron una conducta errática en sus movimientos por el mismo. El hecho crucial del experimento, sin embargo, fue que una vez que el tercer grupo de ratones empezaba a recibir recompensas aprendía a sortear el laberinto con una rapidez significativamente mayor de lo que lo había hecho el primer grupo. Este hecho, corroborado por Tolman cuidadosamente, representaba para él evidencia de que los ratones tenían la capacidad de almacenar "información" (susceptible de "uso" posterior en la conducta) aun sin la mediación de ningún proceso de reforzamiento; esto es, capacidad de "aprendizaje latente". Esto representaba evidencia experimental en contra de un principio fundacional del conductismo y condujo a Tolman a cuestionar las formas más radicales del mismo representadas por Skinner, dominantes en su momento. Trabajos experimentales adicionales inspiraron a Tolman a ir aún más lejos y a sostener que los ratones tienen la capacidad de formar

"representaciones" de alto nivel de su entorno, denominadas por él "mapas cognitivos" (Boden 2006). Al postular la existencia de representaciones de este tipo como la mejor manera de dar cuenta del perfil conductual de los ratones, Tolman proponía una hipótesis que, aun cuando estuviera basada en evidencia experimental y no recurriera a entidades ni procesos subjetivos del tipo empleado por los introspeccionistas, quedaba fuera del ámbito doctrinal del conductismo. El trabajo experimental inspirado por el programa llevaba de esta manera a una socavación empírica de los principios teóricos del mismo.

Más allá de este socavamiento "experimental" del conductismo, el programa fue abandonado en último término en virtud de su esterilidad teórica. Los conductistas aspiraban a determinar las causas de la conducta inteligente; esto es, a desentrañar la naturaleza de los factores que determinaban las capacidades típicas de los organismos inteligentes, en particular de los seres humanos. Los esfuerzos por modelar conductistamente el lenguaje, la memoria o el razonamiento no llevaron a resultados robustos, como muchos conductistas reconocieron pronto (Boden, 2006). El esfuerzo de Skinner por extender el programa conductista a la lingüística y de explicar las habilidades linguísticas de los seres humanos como un conjunto de hábitos generados en procesos de condicionamiento fue severamente examinado y desacreditado por Noam Chomsky (Chomsky, 1959) en una reseña que pronto fue interpretada por la comunidad científica como el documento de defunción del programa conductista. Las habilidades linguísticas de los seres humanos, de acuerdo con Chomsky, exhiben una complejidad y sistematicidad cuya modelación y explicación elude las categorías conductistas y exige, en la línea descubierta por Tolman a partir de su trabajo experimental, la postulación de procesos y estructuras internas al organismo. La caja negra, cuya naturaleza y funcionamiento carecía de importancia de acuerdo con los conductistas, debía abrirse y examinarse cuidadosamente, so pena de dejar sin explicación los fenómenos de mayor interés para la ciencia de la conducta.

1.5. Una reconceptualización necesaria

Un elemento fundamental unía los proyectos conductistas e introspeccionistas de darle estatus de cientificidad al estudio de la conducta inteligente. Ambos proyectos, de diferentes maneras, estaban anclados a la concepción cartesiana de la mente como sustancia pensante no material; una concepción de la mente entendida, en palabras de Daniel Dennett, como un teatro de experiencias subjetivas. Esta sustancia pensante, de acuerdo con Descartes, está de alguna manera vinculada con la conducta inteligente –es su origen subyacente–, aun cuando la naturaleza de este vínculo y en particular de los principios por los que opera la mente no sea susceptible de tratamiento científico. Wundt y sus seguidores pensaban que a pesar de la

interdicción cartesiana era posible descubrir a través de trabajo experimental los principios que regían el funcionamiento de esta sustancia.

El fracaso de este programa convenció a los conductistas de que cualquier esfuerzo de modelar la conducta como la manifestación de factores internos al organismo y en general cualquier intento de formular hipótesis mecánicas sobre la naturaleza de estos factores carecía de cientificidad. La razón de esta convicción anti-mentalista y anti-mecanicista era fundamentalmente que el horizonte de comprensión conductista de estas variables "internas" de la conducta se regía por la idea cartesiana de la mente como un ámbito accesible solo a través de la introspección. Su renuncia radical a apelar a factores internos al organismo, sin embargo, no ofreció los frutos esperados y convenció a la comunidad científica de que, después de todo, la comprensión científica de la conducta exigía la postulación de complejos procesos y estructuras internas. Lo que hacía falta era una reconceptualización de las nociones de "mente", "cognición" y de los factores y procesos "internos" que dejara atrás el dualismo cartesiano. Una serie de innovaciones técnicas en los campos de la ingeniería y las matemáticas aportarían finalmente los elementos necesarios para llevar a cabo esta reconceptualización.

CAPÍTULO 2

2.1. Introducción

Para la filosofía mecanicista desarrollada por Descartes, Boyle y Gassendi la nueva ciencia surgida a partir del siglo XVI constituía una empresa de generación de conocimiento sobre el mundo físico integrada metafísica y metodológicamente. De acuerdo con Descartes, uno de los fundadores de esta filosofía la mente no entraba sin embargo dentro de la jurisdicción de la ciencia. Las razones que motivaron esta exclusión fueron reseñadas en el capítulo previo. Estas razones entrañan una perplejidad teórica: ¿cómo puede un sistema físico dar lugar al tipo de conducta flexible y sistemática característica de los organismos inteligentes, en particular de los seres humanos? La superación de esta perplejidad, entendida como una condición indispensable de la integración de la teorización sobre las causas de la conducta inteligente en el ámbito de la ciencia, fue lo que denominé "el reto de Descartes".

Los programas introspeccionista y conductista en psicología representaron intentos fallidos de efectuar esta integración. La razón subyacente del fracaso de ambos programas, según indiqué, fue su compromiso tácito con el dualismo cartesiano; esto es, con una concepción de la mente entendida, en palabras de Daniel Dennett, como un teatro de experiencias subjetivas. De diferentes maneras, ambos programas supusieron intentos de soslayar el reto cartesiano en la cientifización de la psicología. Una lección del fracaso de estos programas fue el reconocimiento de la necesidad de una reconceptualización sustancial en la comprensión de la naturaleza de la mente y en particular de los procesos "internos" que constituyen las causas de la conducta inteligente. Sin esta reconceptualización el reto cartesiano resulta de hecho insoluble y una ciencia de la mente imposible. La reconceptualización necesaria y la superación de este reto solo fueron posibles a través del desarrollo de ideas en el ámbito de la ingeniería y las matemáticas y del subsiguiente reconocimiento de su relevancia para la comprensión de las causas de la conducta inteligente.

En la sección 2 paso revista al desarrollo de la idea de un *mecanismo de propósito general*, partiendo del diseño de una máquina de tejer automática y programable hasta la formulación de un modelo matemático preciso de un mecanismo de computación programable: una máquina de Turing universal. En la sección 3 considero en términos generales el uso de esta herramienta formal en el proyecto de mecanización de la cognición, entendido como una respuesta al reto de Descartes a la posibilidad de una ciencia de las capacidades cognitivas. En la sección 4 expongo el modelo computacional del cerebro desarrollado por el neurólogo Warren McCulloch y el lógico Walter Pitts a partir de las ideas de Turing. En la sección 5, por último,

examino los diferentes sentidos en los que el computacionalismo sobre la cognición desarrollado por estos autores representa una hipótesis mecanicista.

2.2. De Jacquard a Turing: mecanismos de propósito general y computación

Descartes y los filósofos mecanicistas del siglo XVII habían tenido en mente una concepción restringida de lo que constituía un mecanismo. De acuerdo con esta concepción temprana, un mecanismo es un sistema físico ("extenso") con componentes definidos ("cuerpos") cuya operación depende de interacciones locales entre estos componentes (Roux 2018). Dado un tipo específico de entradas o estímulos, un mecanismo habría de producir, de acuerdo con principios mecánicos definidos que determinaban una sucesión de movimientos en los componentes del sistema, un repertorio restringido de salidas o "conductas" definidas. Esta concepción de los mecanismos, inspirada en los artefactos conocidos en la época, como indiqué antes, se mostró excesivamente simple para la modelación de los procesos biológicos y exigió un ensanchamiento de la categoría en la biología del siglo XIX. Si bien esta revisión del mecanicismo dio lugar a modelos mecánicos más complejos, con diferentes niveles de organización regidos por principios diferentes, la dificultad fundamental señalada por Descartes con respecto a la modelación mecánica de procesos mentales o cognitivos no recibía solución con esta complejidad añadida.

La dificultad residía en la imposibilidad de concebir cómo un mecanismo, que por definición producía un repertorio circunscrito de salidas dado un repertorio circunscrito de entradas, podía dar lugar a capacidades "universales"; esto es, de acuerdo con Descartes, capacidades definidas por exhibir un repertorio no circunscrito de entradas y salidas. Los mecanismos concebidos por los primeros mecanicistas eran sistemas especializados o "de propósito específico" mientras que la mente parecía ser un sistema "de propósito general". Había así una dificultad de principio en concebir cómo en el mundo físico constituido por mecanismos de propósito específico podía ocurrir un sistema con las características de la mente humana. Nada en el desarrollo del mecanicismo en el campo de la biología resolvía esta dificultad de principio. Los sistemas descritos por los biólogos mecanicistas, aun con toda la complejidad añadida, seguían siendo mecanismos "de propósito específico". La innovación requerida para superar esta dificultad empezó a gestarse en el lugar menos esperado.

2.2.1. El telar de Jacquard

A principios del siglo XIX, el tejedor y comerciante francés Joseph Marie Jacquard ideó un mecanismo que permitía automatizar el funcionamiento de los telares de la época y sobre la base de esta idea diseñó y construyó un tipo especial de telar mecánico, conocido como el "telar de Jacquard" (Isaac 2019). Un telar es un dispositivo para

tejer telas y tapices. La acción de tejer requiere entrelazar un hilo ("trama") con un conjunto de hilos diferentes dispuestos en serie ("urdimbre"). Al alternar cuáles de estos hilos están delante o detrás del hilo de trama en cada turno del proceso, el tejedor crea patrones diferentes en el tejido. El tiempo y trabajo requerido para producir patrones complejos motivó el desarrollo de innovaciones tecnológicas que hicieran más eficiente el proceso. Una innovación decisiva fue la introducción de un sistema para codificar mediante agujeros perforados en tarjetas rígidas la secuencia de las posiciones de la urdimbre en el telar que definen un patrón. A través de las tarjetas perforadas pasaban agujas que movían los hilos, de manera que cada tarjeta perforada podía corresponder a una línea del diseño, y su disposición en una serie ordenada de tarjetas determinaba el patrón que el telar tejía. Jacquard perfeccionó este sistema de tarjetas perforadas y diseñó un artefacto que operaba de manera bastante eficiente con el mismo (Boden 2006).

El aspecto más relevante de este artefacto para efectos de la presente discusión es que las tarjetas con las que operaba el telar podían cambiarse, de manera que diferentes patrones podían generarse a partir de los mismos hilos. Las tarjetas funcionaban como "instrucciones" para el mecanismo mediante el que funcionaba el telar. Esta propiedad de este artefacto tendría consecuencias de largo alcance. Lo que el telar de Jacquard ponía de manifiesto era la posibilidad de que un sistema mecánico exhibiera un comportamiento "flexible" a través de un expediente simple: el de almacenar instrucciones que regían el comportamiento del sistema en un medio independiente del sistema, permitiendo que este produjera diferentes efectos ante los mismos estímulos o entradas dependiendo de cuáles instrucciones estuviera "siguiendo" en cada caso (Isaac 2019).

El artefacto de Jacquard no exhibía sin embargo capacidades que cupiera denominar "inteligentes" en ningún sentido relevante para resolver el reto de Descartes. El reconocimiento de que a pesar de todo aportaba los principios necesarios para empezar a resolverlo vino del trabajo que inspiró en el seno de una tradición diferente de la filosofía mecanicista. De acuerdo con esta tradición, el pensamiento humano podía modelarse provechosamente como un proceso matemático; en particular, como un proceso de computación.

2.2.2. Leibniz: pensamiento y computación

La idea de que el pensamiento humano consistía o podía describirse como la efectuación de computaciones, originalmente concebida con precisión por Hobbes, fue desarrollada por Leibniz a finales del siglo XVII. Leibniz compartía la convicción cartesiana de que los procesos mentales no podían explicarse a través de interacción mecánica e igual que Descartes desarrolló una teoría metafísica inspirada en esta convicción. A pesar de esto, creía que parte de los poderes racionales de la mente humana podían describirse matemáticamente y que su naturaleza podía entenderse

mediante un análisis (cuasi-mecanicista) que aislara sus componentes básicos y las "reglas" que regían su organización en unidades más complejas.

Para Leibniz, el pensamiento humano era susceptible de descomposición en secuencias de pasos formales simples, regidos por reglas formales, y se empeñó en formular una teoría del pensamiento entendido en estos términos como un tipo de computación. La forma que tomó este proyecto fue el del desarrollo de un lenguaje simbólico para el razonamiento general, la famosa "characteristica universalis" (Sieg 2009). Leibniz aspiraba a que la culminación de su proyecto aportara las herramientas teóricas necesarias para diseñar y construir un artefacto mecánico capaz de efectuar computaciones, de la manera en la que la mente humana lo hacía. Al margen de la evaluación de las virtudes del proyecto leibniziano como una teoría del pensamiento humano, el punto importante para los presentes propósitos es que Leibniz construyó de hecho un artefacto mecánico capaz de efectuar computaciones. El artefacto de Leibniz era capaz de realizar de manera simple las cuatro operaciones aritméticas básicas y mediante una serie de pasos adicionales ligeramente más complejos podía también calcular raíces cuadradas (Boden 2006). En términos actuales, el artefacto computacional de Leibniz puede entenderse como un tipo de calculadora. Esta calculadora supuso una prueba concreta de que un artefacto mecánico podía exhibir de manera fiable capacidades que en el caso de los seres humanos requieren el empleo de capacidades racionales.

A pesar del logro intelectual y técnico indudable que supuso el diseño y la construcción de este artefacto, la calculadora de Leibniz no dejaba de ser una máquina con capacidades restringidas y probablemente no hubiera sorprendido a Descartes. La mente humana es capaz de muchas más cosas que solo las operaciones aritméticas básicas de las que era capaz el artefacto de Leibniz y no había razones para pensar que la manera en que funciona este arrojese mucha luz sobre la manera en que funciona aquella.

2.2.3. La máquina analítica de Babbage

En la primera mitad del siglo XIX, sin embargo, el matemático inglés Charles Babbage se inspiró en el diseño del telar de Jacquard para llevar un paso adelante el proyecto técnico de Leibniz. Babbage diseñó un artefacto mecánico capaz no solo de calcular las funciones matemáticas básicas sino cualquier función matemática susceptible de solución efectiva: la llamada "Máquina analítica". De acuerdo con Babbage, su diseño establecía las condiciones que permitían a una "máquina finita" efectuar "cálculos de una extensión ilimitada" (Copeland 2006). El diseño de la máquina de Babbage, al igual que el del telar de Jacquard, permitía que el artefacto fuera "programado" para efectuar diferentes operaciones dados los mismos *inputs*. Ambos artefactos podían ser modificados (a través de intervenciones manuales) para que realizaran secuencias largas de operaciones diferentes de una manera que dependía de la

modificación efectuada. En ambos casos, además, la "programación" de la máquina dependía del mismo procedimiento: el empleo de tarjetas perforadas. Mientras que el telar de Jacquard es el primer artefacto mecánico "programable" del que se tiene noticia, el diseño de Babbage encarnaba la primera idea precisa de un artefacto mecánico computacional "programable"; esto es, la primera idea precisa de lo que denominados ahora un "computador" (Piccinini 2008a).

El artefacto de Babbage entraña una arquitectura funcional que distingue entre un mecanismo encargado de efectuar operaciones de computación simples y de almacenar temporalmente valores relevantes para estas operaciones y, por otro lado, un mecanismo encargado de almacenar permanentemente instrucciones complejas que guían las operaciones efectuadas por el primer mecanismo. Esta diferenciación funcional del diseño de Babbage fue redescubierta en el diseño computacional del siglo XX e incorporada en la llamada "arquitectura von Neumann", que constituye la base de casi todos los computadores digitales modernos (Isaac 2019).

La construcción de la máquina de Babbage conllevaba grandes dificultades técnicas y un considerable costo financiero y nunca se llevó a cabo[1]. En parte por esto y en parte en virtud de la novedad radical que suponían, las ideas encarnadas en el diseño de Babbage no tuvieron resonancia en la comunidad científica y fueron parcialmente olvidadas. De manera retrospectiva, por otra parte, puede reconocerse la presencia de una deficiencia teórica en el planteamiento de Babbage que, aunque excusable, permite entender parte de la falta de interés la comunidad científica en sus ideas.

Como indiqué, Babbage pensaba que su diseño exhibía la manera en que mediante medios finitos un procedimiento automático podía efectuar "cálculos de una extensión ilimitada". Sus ideas al respecto, sin embargo, se fundaban en nociones intuitivas relativas al tipo de funciones matemáticas que eran susceptibles de solución "efectiva" o "algorítmica" y carecían de un tratamiento matemático riguroso. El reconocimiento de la importancia de cubrir esta laguna y las primeras formulaciones precisas de las dificultades que entrañaba definir con precisión las nociones de *procedimiento efectivo* y *función computable* solo tuvo lugar a finales del siglo XIX y, en particular, en la primera parte del siglo XX, alrededor de la figura del matemático alemán David Hilbert (Sieg 2009).

2.2.4. Las máquinas de Turing

En el curso de su trabajo en uno de los problemas delimitados por el programa de fundamentación de las matemáticas de Hilbert, el matemático inglés Alan Turing reformuló en términos generales la idea de Babbage y le dio la fundamentación matemática de la que carecía. El problema específico en el que Turing estaba

[1] En la actualidad está en curso un proyecto que pretende construirla para el aniversario 150 de la muerte de Babbage, en el 2021.

trabajando era el conocido "problema de la decisión". El problema consistía en determinar si existía un procedimiento efectivo (un "algoritmo") que pudiera determinar para cada enunciado posible de un sistema lógico de primer orden (es decir, un enunciado de un cálculo de predicados) si el enunciado es susceptible de prueba dentro del sistema. La principal dificultad en la resolución de este problema estribaba en la articulación de una definición general rigurosa de lo que significaba "procedimiento efectivo". En términos intuitivos, un procedimiento de este tipo (un "algoritmo") es un conjunto finito de instrucciones simples para resolver un problema matemático. Dado que el "seguimiento" de un procedimiento este tipo no requiere inteligencia y se supone "automático", los algoritmos se denominaban "procedimientos mecánicos" (Piccinini 2018).

En un artículo escrito en 1936 Turing resolvió el problema de la decisión al demostrar que no existía un procedimiento efectivo que "decidiera" con respecto a cualquier enunciado de un sistema de primer orden si era susceptible o no de prueba dentro del sistema. En su demostración, Turing se sirvió del concepto de "máquina" para ofrecer una delimitación precisa de lo que cuenta como un procedimiento efectivo. Las "máquinas de Turing" son dispositivos computacionales abstractos que manipulan símbolos impresos sobre una cinta idealmente ilimitada dividida en cuadrículas de acuerdo con una serie precisa de instrucciones. Una máquina de Turing simple computa una función de acuerdo con una lista de instrucciones y, entendida como un dispositivo abstracto, es individuada de manera inequívoca por esa lista de instrucciones (Piccinini 2008a). Un procedimiento efectivo es en este sentido una lista de instrucciones constitutiva de una máquina de Turing.

Uno de los resultados sustantivos del artículo de Turing es precisamente que cualquier procedimiento efectivo adecuadamente codificado puede ser concebido como una máquina de Turing, de manera que estos dispositivos permiten delimitar el conjunto de funciones matemáticas computables mediante procedimientos efectivos. Esta conclusión se conoce como la "tesis de Turing" (o bien la "tesis de Church-Turing", dado que el mismo año de la publicación del artículo de Turing, el matemático Alonzo Church llegó de manera independiente a un resultado equivalente) (Sieg 2009). Como se verá en el tercer capítulo, esta tesis y su interpretación posterior tuvo un significado crucial en la historia del desarrollo de las ideas fundacionales de la ciencia cognitiva durante el siglo XX.

Dada por sentada esta tesis, Turing demostró que la pregunta constitutiva del "problema de la decisión" tiene una respuesta negativa: no existe un procedimiento efectivo que "decida" cualquier fórmula posible de un sistema lógico de primer orden. En términos de la tesis, Turing mostró que existen más funciones matemáticas que máquinas de Turing, por lo que muchas de estas funciones no son computables por una máquina de Turing: son, dando por sentada la corrección de la tesis, funciones no-computables. El problema de la decisión es justamente una de estas funciones

(Sieg 2009). Los detalles de la demostración son complejos y carecen de importancia para la presente discusión. Una innovación técnica empleada por Turing en su demostración es sin embargo de importancia crucial para la misma.

Turing mostró cómo podía construirse una única máquina de Turing capaz de simular el comportamiento de cualquier máquina de Turing simple. A estas máquinas las denominó "universales". La idea que subyace a la concepción de una máquina de Turing universal es simple pero, al igual que el diseño de Babbage, tendría consecuencias de largo alcance en el diseño de artefactos computacionales a partir de la segunda guerra mundial (Copeland 2006). Una máquina de Turing simple, como indiqué, se individúa inequívocamente a partir de una lista de instrucciones, entendidas como instrucciones para manipular símbolos impresos en una cinta dividida en cuadrículas. Cada máquina de Turing simple recibe sus instrucciones de una "tabla" que determina el comportamiento del mecanismo de manera exacta en cada paso de su funcionamiento. Cada tabla, esto es, define qué manipulaciones efectúa el mecanismo sobre los símbolos impresos en la cinta de manera secuencial, a partir de un punto de partida y hasta concluir las instrucciones. Los símbolos impresos en la cinta al comienzo del proceso constituyen el *input* de cada máquina y los símbolos impresos al final constituyen el resultado del proceso, que es interpretado en su conjunto como un proceso de computación. Cada máquina, así, es capaz de seguir un único conjunto de instrucciones diseñado para computar una función matemática precisa dado un *input* cualquiera aportado en la cinta de computación. La idea de Turing consistió en contemplar la posibilidad de codificar las instrucciones constitutivas de la tabla de una máquina simple de manera que pudieran aportarse a otra máquina en el mismo formato de los *inputs*, como símbolos impresos en una cinta. Una máquina de Turing universal es una máquina de Turing capaz de simular cualquier máquina de Turing simple, en el sentido de que puede tomar como *input* el conjunto de instrucciones de cualquier máquina de Turing simple y efectuar los mismos procedimientos que esta efectuaría (De Mol 2018).

La idea de una máquina universal de este tipo concebida por Turing puede entenderse, al igual que la idea de Babbage, como el diseño de un *artefacto computacional programable*. Un artefacto programable es un mecanismo capaz de un comportamiento flexible; esto es, capaz de generar de manera sistemática un repertorio amplio de *outputs* diferentes dados *inputs* idénticos. Un artefacto computacional programable es por tanto un mecanismo capaz de efectuar computaciones matemáticas diferentes dados los mismos *inputs* (Piccinini 2008a). Una máquina universal de Turing, en particular, es un artefacto capaz de efectuar todas las computaciones matemáticas susceptibles de ser "mecanizadas". Uno de los logros de Turing fue así mostrar con rigor matemático la posibilidad que Babbage había contemplado un siglo antes: cómo una máquina finita puede desplegar capacidades de cálculo potencialmente ilimitadas.

2.3. El cognitivismo y la mecanización de la cognición

La capacidad de calcular valores de funciones matemáticas con la flexibilidad con la que una máquina universal de Turing puede hacerlo es una capacidad especial. Esta es una de las capacidades que Descartes estimaba imposible en sistemas físicos y que motivó su postulación de una sustancia pensante ontológicamente separada del mundo material. Los artefactos ideados por Babbage y Turing, a diferencia de los artefactos que Descartes y los filósofos mecanicistas tenían en mente al formular sus analogías y modelos de la realidad física, pueden entenderse por tanto, en la medida en que exhiben habilidades pretendidamente exclusivas de los seres humanos, como máquinas "inteligentes". Si bien Babbage nunca tuvo en mente las implicaciones de sus ideas en relación con el reto de Descartes (Boden, 2006), Turing reconoció pronto la importancia de sus ideas matemáticas para la comprensión de la mente humana. De hecho, pensó desde el comienzo en sus máquinas universales como modelos matemáticos de procesos de computación efectuados por humanos (Proudfoot y Copeland 2019). Aunque inicialmente esta analogía entre mente y máquina estaba al servicio de la solución de los problemas matemáticos reseñados, pocos años después de la publicación de su famoso artículo Turing empezó a pensar en la posibilidad de modelar y *explicar* los procesos mediante los que opera la mente en términos de los procesos mecánicos cuyos fundamentos matemáticos había establecido. Turing, dicho en otros términos, empezó a trabajar en la posibilidad de responder de manera directa el reto de Descartes.

Estimulado en parte por la construcción de los primeros computadores electrónicos, cuyos diseños estaban fundamentados en buena medida en sus propias ideas (Copeland 2006), Turing contempló la posibilidad de que todas las operaciones mentales efectuadas por el cerebro humano fueran susceptibles de modelación y explicación mediante máquinas de Turing y consideró el proyecto de construir un cerebro artificial (Proudfoot y Copeland 2019). Aunque sus ideas a este respecto carecían de la precisión de sus ideas matemáticas, la intuición de Turing era que la organización física interna que permitía al cerebro exhibir "inteligencia" debía estar regida por mecanismos similares a los que regían los dispositivos que había ideado. Estas ideas quedaron en estado programático (debido en parte a la muerte prematura de Turing) pero convencieron a un sector importante de la comunidad científica de que el estudio de los procesos mentales en términos mecánicos constituía una empresa teórica viable. La idea de que el cerebro podía ser después de todo un sistema mecánico y que la explicación de sus capacidades podía darse en términos del descubrimiento de los mecanismos internos que regían su funcionamiento parecía por primera vez algo más que una improbable especulación.

En los años posteriores a la segunda guerra mundial, la convicción de que la mente y su funcionamiento podía entenderse en analogía con los mecanismos computacionales descritos por Turing se conjugó con los hallazgos experimentales y

teóricos que mostraban la inviabilidad teórica del conductismo para rehabilitar la estrategia de *explicar* las causas de la conducta mediante la postulación de procesos y mecanismos "internos". La rehabilitación de esta estrategia, fundada principalmente en las razones indicadas, es la fuerza propulsora de la llamada "revolución cognitiva". El "cognitivismo" (un término acuñado solo décadas después pero útil para describir el núcleo de las convicciones compartidas por diferentes investigadores de la época) representa justamente la convicción de que la conducta inteligente es la manifestación o el resultado de factores internos a ciertos organismos y la estrategia consiguiente consistente en modelar y buscar explicar las capacidades mentales de estos organismos mediante la postulación de mecanismos y procesos "internos" a los mismos. La idea general subyacente en el cognitivismo de la posguerra era que el cerebro es un sistema físico que, al igual que una máquina de Turing universal efectivamente construida (o bien un computador electrónico de los tipos que empezaron a construirse por esos años), posee capacidades de cómputo potencialmente ilimitadas, "universales", y que, al igual que el comportamiento de este tipo de artefactos podía explicarse mecánicamente mediante la descripción de su configuración "interna". El funcionamiento del cerebro y la naturaleza de sus capacidades debían poder explicarse, esto es, mediante modelos mecánicos en los que se usarían de manera esencial las herramientas de la teoría matemática de la computación. Esta idea general puede entenderse como el proyecto de la *mecanización* del cerebro y de la cognición (Boden 2006).

La reconceptualización de la cognición necesaria para asignar un estatus científico a su estudio resulta por tanto de responder el reto de Descartes de manera directa y concebir el cerebro como un tipo de mecanismo especial –un mecanismo computacional– y la explicación de sus capacidades distintivas como la tarea de desentrañar los procesos de los que depende el funcionamiento de un mecanismo físico. El computacionalismo es en este sentido una hipótesis mecanicista y una ciencia de la cognición fundada en esta hipótesis representa un capítulo de la empresa científica entendida en clave mecanicista. De la misma manera en la que Loeb mostró cómo la fecundación podía explicarse en términos de mecanismos físico-químicos sin necesidad de recurrir a entidades metafísicamente problemáticas, el computacionalismo supondría la posibilidad de explicar la cognición como el resultado causal de procesos físico-químicos susceptibles de modelación mecánica. En ambos casos, la mecanización de los fenómenos de interés conlleva su integración en el ámbito de estudio de la ciencia.

A diferencia del modelo mecánicos desarrollados en la tradición de la biología mecanicista, sin embargo, el computacionalismo imaginado por los cognitivistas de la posguerra suponía solo una hipótesis general sobre cómo podía proceder la mecanización de los fenómenos cognitivos. Una hipótesis que requería justificación y desarrollo en un modelo concreto que incorporara los hechos conocidos sobre el

funcionamiento del sistema nervioso. En términos generales, la analogía general entre el cerebro y los mecanismos computacionales no establece en qué sentido un proceso de computación puede ser explicativo de la cognición. Aun cuando esta analogía suponga la única manera concebible de explicar científicamente las capacidades cognitivas del cerebro, es preciso sustanciar esta analogía mediante un modelo que indique cómo y en qué sentido los mecanismos computacionales abstractos ideados por Turing pueden emplearse para describir los factores causalmente responsables de la cognición.

Los mecanismos descritos por la teoría matemática de la computación son mecanismos *abstractos*. La descripción de un mecanismo de este tipo es la descripción de entidades y procesos matemáticos. ¿En qué sentido la descripción de mecanismos computacionales abstractos puede emplearse para explicar un fenómeno físico como la cognición? Si el computacionalismo ha de servir para *explicar* procesos físicos y no meramente para describirlos en cierto nivel de abstracción, es preciso explicitar cómo se vinculan los formalismos matemáticos con procesos físicos. Por otra parte, ¿es una explicación computacional de la cognición una explicación mecanicista en el mismo sentido en el que son las explicaciones de otros procesos biológicos como la fecundación o la meiosis? Una explicación mecanicista de la capacidad de un sistema biológico, como indiqué en el capítulo anterior, conlleva un tipo de reducción de esa capacidad. Explicar mecánicamente una capacidad como la metabolización de carbohidratos es exhibir *cómo* esa capacidad resulta de y "se reduce" al comportamiento e interacción de las partes de un sistema en último término físico. Dado que en este caso el sistema en cuestión es el cerebro, la pregunta es si la cognición se reduce a la interacción de las partes del cerebro. En otras palabras, ¿conlleva una mecanización de la cognición una identificación de los procesos cognitivos con procesos cerebrales?

El primer intento sustantivo de responder estas preguntas y sustanciara así el proyecto de mecanización inspirado por las ideas de Turing provino de un artículo publicado en 1943 por el neurólogo Warren McCulloch y el lógico Walter Pitts.

2.4. McCulloch y Pitt: el primer modelo computacional de la cognición

El estudio del cerebro y del sistema nervioso había avanzado desde finales del siglo XIX al margen de la teoría de la matemática de la computación. Como indiqué brevemente en el capítulo anterior, a partir del trabajo pionero de Ludwig von Helmholtz y Emil du Bois-Reymond sobre la fisiología de las células nerviosas, la neurobiología se convirtió en un ámbito fértil de trabajo científico en la segunda mitad del siglo XIX.

Sobre la base de una innovación técnica introducida por el médico italiano Camillo Golgi, consistente en teñir células nerviosas con nitrato de plata, Santiago

Ramón y Cajal desarrolló a finales del siglo caracterizaciones detalladas de la estructura y función de diferentes tipos de células nerviosas. La culminación del trabajo de Cajal fue la articulación de la llamada "doctrina de la neurona", de acuerdo con la cual las células neuronales constituyen la estructura y el componente funcional básico del sistema nervioso. Las neuronas son entendidas aquí como entidades celulares discretas, con componentes y estructura interna, que se organizan en redes interconectadas y que se comunican entre sí. Esta comunicación, según se estableció posteriormente, tiene lugar a través de la transmisión de biomoléculas y de la propagación de impulsos eléctricos entre diferentes neuronas. Esta concepción de la estructura del sistema nervioso constituye el modelo fundamental de la neurobiología contemporánea (Harnish 2002).

Para los años de la posguerra en los que se discutía la posibilidad de formar una ciencia de la cognición fundada en la mecanización del funcionamiento del cerebro y de las capacidades cognitivas, estos hechos sobre la estructura y organización del sistema nervioso eran bien conocidos. La discusión daba por sentado que el cerebro es el órgano en el que tiene lugar la cognición, que las células que constituyen este órgano y que efectúan las funciones cognitivas son principalmente las neuronas y que estas células realizan su trabajo al organizarse en redes interconectadas. Lo que faltaba era un modelo que permitiera entender cómo este sistema físico, compuesto de estos elementos organizados de esta manera, podía generar las capacidades cognitivas complejas características de los organismos inteligentes. Esto es, en la jerga mecanicista, *cómo* las capacidades del "todo" que es el sistema nervioso, resultan de la interacción de sus "partes", las redes neuronales. La convicción resultante de la adopción de las ideas de Turing en la reflexión sobre la cognición, como se vio, es que la solución de este problema requería el empleo de las herramientas de la teoría matemática de la computabilidad.

McCulloch y Pitts llevaron a cabo la primera aplicación sistemática de la teoría de la computación en la modelación del cerebro. Su trabajo puede entenderse como un esfuerzo por resolver el problema fundamental de una teoría científica de la cognición: a saber, cómo mecanismos en último término físicos pueden dar lugar a capacidades cognitivas. De acuerdo con McCulloch y Pitts, la solución a este problema dependía de entender el cerebro como una máquina de cómputo del tipo diseñado por Turing. Un presupuesto central de su propuesta era el llamado "principio del todo o nada" sobre el comportamiento de las neuronas, ampliamente aceptado en los años de la posguerra a partir del trabajo experimental del fisiólogo inglés Edgar Adrian (Abraham 2019). Una célula neuronal típica tiene dos tipos de apéndices funcionalmente importantes que emergen de su núcleo: dendritas y axones. Las dendritas modifican sus propiedades eléctricas en virtud de la recepción señales químicas provenientes de otras neuronas de manera que, bajo ciertas circunstancias, estos cambios afectan las propiedades eléctricas de la membrana del núcleo de la

célula. Esta modificación causa un disparo o impulso en el axón de la neurona como resultado del cual esta envía a su vez una señal química a otra neurona. Este proceso general es el núcleo de la actividad del cerebro (Grush y Damm 2012). De acuerdo con el principio del todo o nada, la respuesta de una neurona a un estímulo químico no depende de la "fuerza" de este estímulo. Si el estímulo excede un cierto umbral definido de excitación, la neurona generará un impulso; si no, permanecerá en reposo.

Esta característica de las neuronas hacía posible entender su comportamiento en términos computacionales. Una red de neuronas podía concebirse en virtud de este principio, de acuerdo con McCulloch y Pitt, como un dispositivo computacional simple y el funcionamiento del cerebro en su conjunto como un sistema físico podía entenderse en términos de una máquina de Turing (McCulloch y Pitt 1943).

2.4.1. El cerebro como una máquina lógica de procesamiento de información

La idea subyacente en el modelo propuesto es que la cognición es procesamiento de información. El núcleo de esta idea puede explicitarse en los siguientes términos. Un organismo recibe información del entorno a través de mecanismos de transducción, que convierten estímulos físicos en series de impulsos neuronales. Estos impulsos codifican información en un formato adecuado para su procesamiento por parte del cerebro. Los impulsos nerviosos constituyen en este sentido entidades que vehiculan información sobre los eventos que los causan, de manera que la comunicación entre neuronas puede concebirse como el procesamiento de información sensorial. Un modelo del funcionamiento del cerebro es por tanto un modelo de cómo la información "fluye" a través de redes de neuronas (Piccinini 2004c).

Los *inputs* y *outputs* de una neurona eran entendidos por McCulloch y Pitt en analogía con símbolos escritos en una cinta de una máquina de Turing. En el funcionamiento de una de estas máquinas, tales símbolos son objeto de manipulaciones mecánicas de maneras susceptibles de interpretación semántica. El *input* de una máquina de Turing es una ristra de símbolos y la manipulación de esta ristra de acuerdo con las instrucciones constitutivas de la máquina genera un *output* que es asimismo una ristra de símbolos. Estos *inputs* y *outputs* pueden recibir una *interpretación*, en el sentido de que pueden entenderse como codificaciones de números específicos, de manera que el proceso de modificación mecánica de los símbolos de entrada en los de salida puede interpretarse como la determinación del valor de una función matemática. De la misma manera, conjeturaban McCulloch y Pitt, los impulsos neuronales son entidades que codifican información y su propagación causal en redes de neuronas da lugar a transformaciones "mecánicas" de los mismos que pueden entenderse como procesos computacionales susceptibles de interpretación semántica; esto es, como mecanismos de procesamiento de información. El cerebro es en este sentido una máquina cuyos componentes y funcionamiento son de naturaleza física pero que vehicula contenido semántico

relativo a estados de cosas del mundo con el que el organismo interactúa (McCulloch y Pitt, 1943).

El contenido que codifican los impulsos neuronales era para McCulloch y Pitt contenido *proposicional*. Las señales neuronales, tal era la idea, son formalmente equivalentes a variables de un cálculo proposicional. La información que codifica un disparo neuronal era concebida como información relativa a la ocurrencia de un estado de cosas, por lo que en su modelamiento computacional estos disparos podían interpretarse como variables que toman valores booleanos. Un disparo neuronal, de acuerdo con esto, codifica información que toma uno de dos valores posibles. Este hecho hacía posible entender el comportamiento de las redes neuronales como la computación de funciones lógicas; a saber, funciones que toman como *inputs* y generan como *outputs* proposiciones con valores booleanos (McCulloch y Pitt, 1943).

McCulloch y Pitt desarrollaron una poderosa técnica para diseñar dispositivos abstractos que computan mecánicamente, en el sentido de una máquina de Turing, valores de funciones proposicionales. La idea que motivó el desarrollo de esta innovación fue que el comportamiento de las redes neuronales podía interpretarse como un tipo de evento funcionalmente equivalente al de estos dispositivos. En términos actuales, la innovación técnica de McCulloch y Pitt se conoce como el diseño de "puertas lógicas". Esta innovación fue adoptada años después por John von Neumann en el diseño de los primeros computadores electrónicos y eventualmente las puertas lógicas terminaron constituyendo el componente computacional primitivo de la mayoría de los artefactos computacionales modernos (Piccinini 2008b).

Una puerta lógica es un dispositivo con dos o tres terminales de entrada cuya función es recibir uno o varios *inputs* y generar mecánicamente un único *output*. Los *inputs* y *outputs* de estos dispositivos son denominados booleanos o "binarios", en el sentido de que pertenecen a uno de dos tipos, usualmente denominados "1" y "0". Estos "1" y "0" pueden interpretarse como los valores "verdadero" y "falso", de manera que las operaciones que el dispositivo efectúa sobre sus *inputs* pueden concebirse como la determinación de valores de funciones lógicas. El tipo de *output* generado por estos dispositivos corresponde así a lo que resultaría de aplicar una conectiva de la lógica proposicional a uno o varios valores proposicionales. Esta es la razón por la que se denominan puertas *lógicas*. El comportamiento de un dispositivo de este tipo puede representarse mediante tablas de verdad y mediante ecuaciones escritas en un álgebra booleana (Patterson y Hennesy 2014). Una puerta NO, por ejemplo, toma como *input* un valor binario (un 1 o un 0) y produce como *output* el valor opuesto. Una puerta Y toma como *input* dos valores y genera como *output* un 1 si y solo si estos dos valores corresponden al valor 1 o bien un 0 en cualquiera de los otros casos posibles.

McCulloch y Pitt emplearon este formalismo para modelar el comportamiento de las neuronas y de las redes que conforman en su interconexión. La propagación de

impulsos a través de redes neuronales podía entenderse de acuerdo con este formalismo como la computación de vehículos de información y el cerebro en su conjunto podía concebirse como una máquina constituida por puertas lógicas interconectadas que reciben y procesan información del entorno con el que interactúa un organismo.

2.4.2. El sistema nervioso como un sistema cognitivo

El punto central de la propuesta de McCulloch y Pitt es que el empleo de un formalismo lógico del tipo descrito permite representar el sistema nervioso como un sistema cognitivo; esto es, un sistema cuyo funcionamiento consiste en el despliegue de capacidades cognitivas. Su modelo computacional del funcionamiento del sistema nervioso, dicho en otros términos, aportaría las herramientas necesarias para cerrar el hiato entre la descripción neuronal del cerebro y su descripción psicológica y de esta manera para articular un modelo causal de la cognición. McCulloch y Pitt pretendían que su modelo hacía posible *explicar* mecánicamente la manera en que el cerebro humano efectúa tareas como la percepción de objetos sobre la base de estímulos sensoriales, el razonamiento matemático y el pensamiento abstracto; tareas, esto es, definidas en términos psicológicos antes que neuronales. El principio del que dependía esta unificación entre lo neuronal y lo cognitivo, como he mostrado, era la pretendida equivalencia formal entre cadenas de eventos neuronales y ciertas inferencias lógicas.

Puesto que la cognición era concebida como la recepción y procesamiento de información, lo que una teoría de la cognición debía explicar en primer lugar era cómo el cerebro, un sistema físico organizado a través de redes neuronales, realiza tareas de procesamiento de información. La respuesta de McCulloch y Pitt es que el cerebro efectúa estas tareas mediante mecanismos de computación lógica. La razón por la que la sucesión causal de eventos físicos en los que consiste la propagación de impulsos a través de una red de neuronas puede interpretarse en términos cognitivos como el procesamiento de información es su equivalencia estructural con mecanismos que efectúan inferencias lógicas. Una sucesión de eventos neuronales exhibe de acuerdo con esto una estructura causal que refleja la estructura de un mecanismo que computa funciones lógicas. Esta equivalencia formal era en opinión de McCulloch y Pitt suficiente para asignar a una sucesión de eventos neuronales propiedades cognitivas (Piccinini 2004c).

2.5. El computacionalismo como una hipótesis mecanicista

El de McCulloch y Pitt representa el primer modelo general computacional sistemático del cerebro y sus capacidades cognitivas, por lo que estos autores pueden considerarse los padres de lo que se ha denominado la “teoría computacional de la

mente". Aun cuando Turing contempló la posibilidad de modelar computacionalmente el cerebro, nunca presentó una propuesta concreta de cómo podría hacerse esto y existe evidencia de que sus ideas a este respecto y en particular su proyecto de construir un cerebro artificial tomaron forma solo después de que leyera el artículo de McCulloch y Pitt (Piccinini 2004c). En manos de McCulloch y Pitt, el computacionalismo es una *hipótesis* mecanicista sobre el funcionamiento del cerebro; en particular, sobre la naturaleza funcional específica de los procesos neuronales. De acuerdo con esta hipótesis, el comportamiento de cada red neuronal y por tanto todo lo que hace el cerebro puede describirse como la efectuación de computaciones. Explicar una capacidad cognitiva cualquiera consiste en presentar un modelo mecánico-computacional relativo a cómo el comportamiento de una red neuronal da lugar a la capacidad en cuestión.

El lenguaje, la memoria, el aprendizaje, la percepción y el razonamiento, de acuerdo con el modelo general de McCulloch y Pitt, *habrían* de poder explicarse en este sentido como capacidades resultantes del procesamiento computacional de información por parte de redes neuronales. El éxito de esta empresa permitiría, por otra parte, "reducir" las propiedades y fenómenos mentales a propiedades y fenómenos neuronales y en consecuencia integrar el estudio científico de la cognición con el estudio científico del mundo físico. En la concepción mecanicista de la ciencia compartida por McCulloch y Pitt, la reducción de un fenómeno a componentes y principios en último término físicos, como indiqué en el capítulo previo, se concebía como una condición de su tratamiento científico. Al reducir la cognición a actividad neuronal, el objetivo era desterrar finalmente la amenaza del dualismo cartesiano y su postulación de una sustancia "fantasmal" que residía de alguna manera en la maquinaria del cerebro (McCulloch y Pitt 1943).

McCulloch y Pitt no sustanciaron sin embargo su modelo general mediante la formulación de modelos específicos explicativos de capacidades cognitivas precisas. Salvo por un ejemplo muy simple sobre cómo podía explicarse una ilusión perceptual, su propuesta se redujo a explicitar el marco general en el que podían proceder estas explicaciones. Esta renuncia a proponer modelos explicativos específicos respondía a la consciencia del carácter rudimentario del conocimiento disponible de los procesos neurobiológicos responsables de las capacidades del cerebro.

A pesar de los avances de la neurobiología desde finales del siglo XIX, el conocimiento del funcionamiento del cerebro era incipiente y rudimentario en el momento en que McCulloch y Pitt formularon su modelo (como lo es, a pesar de los avances efectuados en las últimas décadas, todavía hoy). El uso por parte de McCulloch y Pitt de la analogía entre cerebro y máquina entrañaba en este contexto una dosis importante de idealización. En contra de evidencia conocida en su momento, por ejemplo, el modelo daba por sentado que ciertos factores químicos que determinan la generación de impulsos en las neuronas y su comportamiento posterior

a la producción de estos impulsos son funcionalmente irrelevantes. De la misma manera, el modelo asume que la estructura de las redes neuronales no se modifica con el tiempo. McCulloch y Pitt sabían que estas presuposiciones representaban falsedades pero creían que su presuposición no afectaba el potencial teórico del modelo.

La presuposición de estas falsedades era necesaria para emplear los formalismos de la teoría de la computación en la modelación del comportamiento de las neuronas. En la medida en que la mecanización de la cognición (la conceptualización de la misma como una manifestación causal del funcionamiento del cerebro) exigía el uso de estas herramientas formales, las simplificaciones e idealizaciones en cuestión estaban justificadas. Concebir el cerebro en términos de *inputs*, *outputs* y procesamiento de información suponía de acuerdo con estos autores la mejor estrategia disponible para entender su funcionamiento como un sistema cognitivo. La analogía haría *posible* formular modelos que describen de manera precisa fenómenos cognitivos en términos de mecanismos físicos y señalaría cómo pueden ponerse a prueba estos modelos en test experimentales precisos. Ninguna otra estrategia conocida aportaba estos rendimientos. McCulloch y Pitt eran conscientes de que los mecanismos específicos que podían postularse a partir de su modelo serían mecanismos meramente posibles (Piccinini 2004c). La propuesta de estos autores, en resumen, era concebida por ellos mismos más como una herramienta para avanzar en el conocimiento de las propiedades funcionales del cerebro que una teoría o un modelo fiel a los mecanismos efectivos que explican estas propiedades.

La postulación de mecanismos es una actividad teórica que puede evaluarse en dos dimensiones diferentes. En primer lugar, la postulación de un mecanismo puede estar mejor o peor respaldada epistémicamente por la evidencia disponible. En segundo lugar, el mecanismo explicativo postulado en un modelo mecánico puede estar descrito con mayor o menor detalle. El ideal orientador de la filosofía mecanicista ha sido la postulación de mecanismos respaldados con la mayor fuerza epistémica y descritos con el mayor detalle posible. La satisfacción de este ideal depende sin embargo en cada caso del conocimiento acopiado por la ciencia con respecto a un ámbito de estudio en un momento dado de la historia. En las disciplinas científicas más desarrolladas se espera la formulación de modelos explicativos máximamente detallados y sustentados por cuerpos robustos de evidencia. En las disciplinas menos desarrolladas en cambio la formulación de modelos esquemáticos o tentativos y respaldados en cuerpos menos robustos de evidencia es admisible y representa una estrategia de desarrollo del conocimiento científico (Glennan 2017).

En términos actuales, este tipo de modelos se denominan modelos mecánicos "*how-possibly*" (Craver 2007; Glennan 2017). Un modelo de este tipo postula un mecanismo *posible*, en el sentido de que representa una conjetura explicativa esquemática para la cual no existe un respaldo evidencial robusto. Estos modelos son

admisibles y valiosos teóricamente en la medida en que abren nuevas avenidas de investigación y, en muchos casos, representan la única conjetura explicativa razonable dado el conocimiento acopiado en un ámbito de estudio científico. Tal es el caso con respecto a la neurobiología y el modelo del cerebro desarrollado por McCulloch y Pitt.

Esta especie de modestia teórica en la formulación de modelos e hipótesis explicativas no se opone en general al espíritu de la filosofía mecanicista. Los mecanicistas clásicos daban por supuesto que el mundo físico estaba constituido por mecanismos como parte de una estrategia heurística de modelación que permitía la formulación de conjeturas empíricas susceptibles de corroboración experimental relativas a las causas de los fenómenos naturales. La analogía entre el cerebro y los dispositivos mecánicos fundados en la teoría de la computación se justifica en este sentido en virtud de las posibilidades que ofrece de articular hipótesis y estudios experimentales precisos (Abraham 2019).

La dimensión reduccionista del modelo de McCulloch y Pitt (a diferencia de otros aspectos del mismo, como de la identificación de las nociones de *computación* y *procesamiento de información*) no fue acogida de manera positiva por los investigadores que dieron perfil institucional a la ciencia cognitiva a finales de los años cincuenta del siglo pasado. Aunque en un primer momento el trabajo de estos investigadores procedió de espaldas a esta cuestión fundacional, eventualmente la nueva "ciencia cognitiva" se organizó alrededor de una filosofía anti-reduccionista.

El modelo computacional desarrollado por McCulloch y Pitt representaba un marco de referencia demasiado esquemático para fundamentar una nueva disciplina científica. El escaso conocimiento de los procesos neurológicos que subyacían a la cognición contrastaba con el entusiasmo generado por los nuevos modelos post-conductistas y con el percibido potencial de la teoría de la computación para entender los procesos cognitivos. Todos estos factores empujaban en la dirección del desarrollo de una filosofía que fundara una disciplina de estudio de la cognición autónoma con respecto al estudio del cerebro. El funcionalismo desarrollado a partir de los años sesenta principalmente por Jerry Fodor supuso la satisfacción de esta necesidad y se convirtió eventualmente en la filosofía oficial de la nueva ciencia cognitiva (Boden 2006; Bechtel et. al 1998).

A pesar de su origen en una tradición de pensamiento mecanicista el computacionalismo terminó interpretándose como una tesis no mecanicista y dando forma a un programa de investigación de la cognición alejado de los ideales de la filosofía mecanicista: la ciencia cognitiva clásica. Calibrar las razones y la forma que tomó este desplazamiento será el objeto del próximo capítulo.

CAPÍTULO 3

3.1. Introducción

El concepto de *reducción* empleado en la descripción del modelo computacional de McCulloch y Pitt y característico de la filosofía mecanicista discutida en los dos últimos capítulos es un concepto meta-teórico amplio y no del todo preciso. ¿Qué tipo de cosa cuenta como una reducción? El mecanicismo se sirve de este concepto para presentar tesis sobre la naturaleza de las explicaciones científicas, la unidad de la ciencia y la estructura de la realidad.

Explicar mecánicamente un fenómeno es exhibir cómo este depende causalmente de las partes de un sistema físico y de la organización de sus interacciones. Esta dependencia causal, como se vio, es interpretada en la filosofía mecanicista como un tipo de reducción fisicalista. En la medida en que de acuerdo con esta filosofía la explicación mecánica de los fenómenos naturales es el objetivo de todas las disciplinas científicas, por otra parte, la unidad de la ciencia resulta de la reducción de los fenómenos estudiados por cada disciplina a mecanismos físicos. Y dado también que las representaciones científicas de la realidad aspiran a ser verdaderas de esa realidad y el conocimiento científico a desentrañar la estructura del mundo, las reducciones efectuadas por las explicaciones mecánicas de los fenómenos exhiben relaciones de dependencia metafísica y no meramente epistémica. Las explicaciones mecánicas aspiran a describir la manera en que el mundo está efectivamente constituido más allá de nuestras representaciones parciales y subjetivas del mismo. El concepto mecanicista de reducción, así, igual que la filosofía mecanicista misma, entraña componentes tanto epistémicos y metodológicos como metafísicos.

Esta mezcla entre dimensiones metafísicas, metodológicas y epistémicas en el concepto mecanicista de *reducción* se percibió como una deficiencia en la reflexión filosófica sobre la ciencia cuando esta empezó a profesionalizarse en la segunda mitad del siglo XIX. En particular, el componente metafísico de la concepción mecanicista de la ciencia resultaba sospechoso dada la orientación empirista y neokantiana de los teóricos que, principalmente en Inglaterra y Alemania, empezaron a ocuparse de la ciencia como un objeto de estudio en derecho propio. Hasta el siglo XIX, la reflexión sobre la naturaleza del conocimiento *científico* y sobre la ciencia como una empresa regida por reglas distintivas había sido principalmente potestad de científicos preocupados por los fundamentos y el carácter de su propia actividad. Cuando la filosofía de la ciencia emergió como una disciplina lo hizo alrededor de agendas epistemológicas empiristas y neokantianas críticas de la postulación de entidades y estructuras en la naturaleza con una existencia robusta e independiente de las actividades cognoscitivas de los científicos (Díez y Moulines 1999).

Los partidarios de estas nuevas filosofías de la ciencia creían que para entender los fundamentos del conocimiento científico era innecesario e indeseable recurrir a conceptos metafísicos. Algunos de estos autores, entusiastas del ideal de unidad de la ciencia, estimaban que el componente metafísico del mecanicismo era un obstáculo y no, como los mecanicistas clásicos pensaban, un componente necesario en la comprensión del mismo. Aunque los mecanicistas estaban en lo correcto en su adopción de un tipo de fisicalismo y en el énfasis en la importancia de la noción de *reducción*, hacía falta purgar a estos conceptos de su dimensión más metafísica.

En este contexto surgió el positivismo lógico, la filosofía de la ciencia dominante en el momento en el que el computacionalismo sobre la cognición tomó forma en manos de Turing, McCulloch y Pitt. El proyecto intelectual de este movimiento puede entenderse justamente como la formulación de una vision exenta de consideraciones metafísicas de la ciencia como una empresa unificada. Para la corriente principal del positivismo lógico, esta visión de la ciencia se articulaba alrededor de una comprensión no metafísica de las ideas asociadas con el fisicalismo y la noción de *reducción*.

En la sección 2 expongo en líneas generales la tradición newtoniana en filosofía de la ciencia como una tradición alternativa a la filosofía mecanicista discutida en los dos capítulos previos. En esta tradición, como se verá, las nociones de *explicación* y *reducción* se conciben en términos epistémicos como formas de subsunción bajo regularidades. En la sección 3 presento la filosofía de la ciencia del positivismo lógico como una sistematización de la tradición newtoniana. La concepción nomológico-deductiva de las explicaciones científicas y la idea positivista de unidad de la ciencia como resultado de una forma de reducción inter-teórica serán el centro de atención de esta sección. En la sección 4 examino una serie de objeciones a esta idea positivista de unidad de la ciencia y el consecuente surgimiento a partir de los años sesenta de una filosofía anti-reduccionista fundada en la noción de *realizabilidad múltiple*. En la sección 5 me ocupa de la forma que este anti-reduccionismo tomó en la filosofía de la mente: el funcionalismo. En la sección 6, por último, analizo el maridaje efectuado por Jerry Fodor entre el funcionalismo y el computacionalismo sobre la cognición. Esta interpretación del computacionalismo, contraria a la tradición mecanicista discutida en el capítulo anterior, se convertiría en la interpretación dominante y en el fundamento de la llamada ciencia cognitiva clásica a partir de los años sesenta.

3.2. Leyes, teorías y explicación: la tradición newtoniana en la filosofía de la ciencia

La concepción de la ciencia desarrollada por el positivismo lógico es heredera del anti-mecanicismo newtoniano descrito brevemente en el primer capítulo. En esta tradición newtoniana, las nociones de *ley* y *teoría*, más que la de *mecanismo* y *modelo* son las categorías meta-teóricas fundamentales; esto es, las nociones más importantes mediante las que el teórico sobre la ciencia describe la actividad científica.

De acuerdo con Newton, explicar un fenómeno natural consiste en exhibir principios máximamente generales de los cuales se sigue de manera deductiva la ocurrencia de ese fenómeno. Estos principios generales eran para Newton *leyes* de la naturaleza. Al mostrar que un fenómeno dado cae bajo una ley, al "subsumirlo" bajo una ley, la "razón" de su ocurrencia queda para efectos científicos establecida (Psillos 2007). Las leyes de la naturaleza son cierto tipo de regularidades susceptibles de determinación experimental o, de manera alternativa, enunciados que expresan estas regularidades. Estos enunciados son enunciados generales del tipo "todos los *X* son *Y*" o "siempre que ocurre tal cosa ocurre tal otra".

El enfoque de Newton era principalmente matemático y cuantitativo y las regularidades de importancia científica eran para él regularidades expresables matemáticamente. El propósito de la ciencia, de acuerdo con lo que cabría llamar la filosofía newtoniana de la ciencia, consiste en *reducir* los fenómenos naturales a un conjunto restringido de leyes naturales máximamente generales expresables en términos matemáticos. La validación epistémica de estas leyes depende de que las consecuencias inferidas deductivamente de las mismas sean susceptibles de corroboración experimental. Una vez que una ley muestra su poder deductivo en este sentido, cualquier hipótesis adicional sobre las posibles causas en virtud de las cuales la ley tiene aplicación empírica resulta innecesaria. Para Newton, en oposición con los mecanicistas, el intento de explicar las leyes naturales validadas experimentalmente mediante la postulación de mecanismos suponía una pretensión teórica superflua (Janiak 2014; Boden 2006).

La postulación de un mecanismo es en este sentido lo que Newton llamaba una "hipótesis". Las "hipótesis" eran para Newton afirmaciones sobre un fenómeno natural no deducidas de una ley o establecidas mediante prueba experimental. En su opinión, este tipo de afirmaciones no tenían lugar en la ciencia: "For whatever is not deduced from the phenomena is to be called a hypothesis, and hypotheses, whether metaphysical or physical, whether of occult qualities or mechanical, have no place in experimental philosophy" (citado por Psillos 2007, p. 108).

Para los mecanicistas de la época, en cambio, las leyes no podían ser las causas últimas de los fenómenos naturales y exigían ellas mismas una explicación causal en términos de mecanismos físicos. La renuencia de Newton a ofrecer "hipótesis" sobre

las causas de las leyes naturales generó especial controversia en relación con su ley de la gravitación universal. Para mecanicistas como Leibniz, la gravedad newtoniana equivalía a una "cualidad oculta": una propiedad postulada para explicar algo pero carente ella misma de explicación. En el marco teórico del mecanicismo la postulación para efectos explicativos de una propiedad carente de explicación mecánica suponía una jugada teórica ilegítima. No así en la filosofía newtoniana de la ciencia. Newton concedía que no conocía las causas de la gravedad pero afirmaba que esto no era problemático. Una manera de resumir la respuesta de Newton a Leibniz es que aun cuando no pudiera establecer las causas de la gravedad, su trabajo había establecido que la gravedad *era* causal: "to us it is enough that gravity does really exist and act according to the laws which we have explained, and abundantly serves to account for all the motions of the celestial bodies and of our sea" (citado por Psillos 2007, p. 109). En la medida en que una ley permita deducir y predecir la ocurrencia de ciertos fenómenos de maneras sistemáticas, es una ley causal. Su naturaleza causal se reduce a las posibilidades que ofrece de determinar factores empíricamente constatables de los que *depende* la ocurrencia de estos fenómenos. Para Newton, esta constituía la única manera de razonar científicamente sobre causas naturales. Si un evento *a* está relacionado causalmente con un evento *b* es por tanto algo que no requiere de la existencia de un mecanismo que conecte ambos eventos (Harper 2002).

No todas las regularidades empíricas verdaderas, sin embargo, son leyes. Aunque el carácter *legal* de una regularidad no dependa de su fundamentación en un mecanismo, es preciso especificar de alguna manera las razones en virtud de las cuales ciertas regularidades tienen el estatus de leyes y pueden emplearse en consecuencia para explicar y establecer relaciones causales entre fenómenos naturales.

En términos generales, las leyes son regularidades que tienen ciertas propiedades y pueden ser usadas para ciertos propósitos. La propiedad fundamental que distingue a estas regularidades es su llamado carácter "modal". Si una variable X está conectada con una variable Y por una regularidad que es una ley, entonces es "necesario" que si algo es *X* sea también Y. Este carácter modal diferencia a las leyes de las regularidades *accidentales*. Aun cuando todas las esferas de oro en el universo tengan menos de una milla de diámetro, no es necesario que así sea. "Todas las esferas de oro tienen menos de una milla de diámetro" es por tanto un enunciado que expresa una regularidad probablemente verdadera pero no una ley. Que todas las esferas de uranio tienen menos de una milla de diámetro es en cambio una ley: si algo es una esfera de uranio, entonces necesariamente tiene menos de una milla de diámetro. El carácter modal de las leyes se evidencia, como se dice, en su apoyo a enunciados contrafácticos. Si algo *fuera* una esfera de uranio, entonces podría inferirse con seguridad que *tendría* menos de una milla de diámetro. Si se *calentara* un trozo de metal *x*, entonces podría inferirse con seguridad que *x* se *dilataría*. El carácter "necesario" de las leyes está

estrechamente vinculado con el hecho de que puedan emplearse para ciertos propósitos en la práctica científica; a saber, como indiqué, para hacer predicciones, explicar fenómenos y establecer relaciones causales (Díez y Moulines 1999).

¿Qué tipo de *necesidad* es la que está en juego acá? La filosofía de la ciencia de Newton no ofrece una respuesta a esta pregunta. Para Newton, era suficiente con que una regularidad pudiera establecerse de manera experimental, fuera expresable matemáticamente y pudiera emplearse en la deducción de fenómenos de interés científico. De acuerdo con David Hume, sin embargo, las directrices empiristas que la filosofía natural de Newton representaba aportaban las herramientas para dar una respuesta satisfactoria a la pregunta en cuestión.

3.2.1. **Hume sobre la necesidad de las leyes naturales**

Sobre la base del supuesto empirista de que todo conocimiento y en general todos los contenidos de la mente tienen su origen en la experiencia sensorial, Hume se planteó la cuestión del origen de la idea de *necesidad* asociada con las leyes. Esta necesidad, argumentó famosamente, no es algo que *percibamos* a través de los sentidos ni que podamos *deducir* de la evidencia empírica.

Cualquier intento de mostrar con base en la experiencia que una regularidad constatada en el pasado *debe* seguir teniendo lugar en el futuro desemboca en un razonamiento circular e infundado. Para concluir de la constatación hasta un tiempo t de una regularidad empírica que la regularidad en cuestión seguirá presentándose en un tiempo $t+1$ es preciso dar por sentado un "principio de uniformidad de la naturaleza". Pero este principio no puede justificarse a su vez ni de manera *a priori* ni de manera empírica. Es posible que las "regularidades" constatadas en el pasado sean accidentales o no sean regularidades en absoluto; puede ser, esto es, que el mundo se comporte en el futuro de manera diferente a la que esperamos y que nos lleva a efectuar ciertas inferencias. Nuestra confianza en que el mundo se comporta de manera uniforme no se funda en la percepción de ningún tipo de necesidad en las entidades y estados de cosas que son objeto de la experiencia ni en un principio que permita inferir con certeza deductiva este hecho, sino, de acuerdo con Hume, únicamente en un hábito o costumbre de la mente. La necesidad de las leyes, por tanto, es algo que "está" en la mente, no en los objetos. Una ley es una regularidad observada que, en virtud de factores psicológicos, *proyectamos* hacia el futuro, *esperamos* que siga igual (Hume 1999).

De esta manera, el concepto de *ley* concebido por Newton como el punto de referencia de toda la actividad científica fue desprovisto por Hume de cualquier dimensión "metafísica". En este contexto, la metafísica es concebida como un ámbito de especulación sobre las causas subyacentes de los fenómenos naturales deslindado de la experiencia y por tanto teóricamente ilegítimo. Dado que en este marco de referencia empirista la idea de ley delimita el contenido de los conceptos de

causalidad y *explicación*, la búsqueda científica de causas y la explicación de fenómenos naturales fue también en último término deslindado por Hume de toda dimensión metafísica (Psillos 2007). Puesto que la noción de *mecanismo* empleada por los mecanicistas estaba atada a un tipo de teorización sobre las causas subyacentes de los fenómenos naturales no susceptible en muchos casos de validación empirista, para Hume, al igual que para Newton, no tenía cabida en la actividad científica. En el capítulo 5 examinaré con mayor detalle la concepción humeana de la causalidad en el contexto del escrutinio crítico al que esta concepción y en general el marco empirista de la filosofía newtoniana de la ciencia han sido sometidos en la filosofía de la ciencia de las últimas décadas.

3.2.2. **La matematización de la física en el siglo XIX**

A finales del siglo XIX, los progresos de la física parecían respaldar inequívocamente la concepción newtoniana de la ciencia. En particular, el trabajo de James Maxwell puso de manifiesto el potencial del tratamiento matemático de regularidades observadas experimentalmente para dar cuenta de fenómenos físicos fundamentales. El gran logro de Maxwell fue la formulación de unos pocos principios matemáticos mediante los cuales, en el espíritu de la filosofía natural de Newton, podían subsumirse el comportamiento de todos los fenómenos "electromagnéticos". Las ecuaciones de Maxwell permitieron concebir las regularidades relativas a la electricidad, el magnetismo y la luz como manifestaciones de unos pocos principios fundamentales. El desarrollo de geometrías no euclidianas y su empleo en la teoría de la relatividad, por otra parte, aportaron evidencia adicional del carácter matemático y abstracto de las leyes y principios que permitían describir científicamente el mundo natural. La física del siglo XIX parecía respaldar así una idea de la ciencia de acuerdo con la cual la tarea de los científicos consistía en interpretar los fenómenos empíricos en términos teóricos abstractos a través del uso de métodos matemáticos especializados, en los que los modelos mecánicos ideados por la ciencia mecanicista tenían poca o ninguna cabida (Bailer-Jones 2009). Más que modelos, serían *teorías* matemáticamente formuladas los vehículos mediante los cuales la ciencia representaría el mundo natural. Pierre Duhem resumió de manera penetrante la concepción resultante de la naturaleza de las teorías entendidas como los vehículos de expresión del conocimiento científico: "[Una teoría física] es un sistema de proposiciones matemáticas, deducidas de un número pequeño de principios, que aspiran a representar de la manera más simple, completa y exacta posible un conjunto de leyes experimentales" (citado por Bailer-Jones 2009, p. 86). Las "leyes experimentales" son en este contexto regularidades empíricas establecidas experimentalmente. En este clima intelectual dominado por una imagen de la ciencia enfocada en la matematización de la física surgió el positivismo lógico.

3.3. **El positivismo lógico como una filosofía general de la ciencia**

El positivismo lógico puede entenderse como un esfuerzo por reconciliar dos aspectos aparentemente antagónicos de la imagen newtoniana de la ciencia a la luz del progreso de la física durante el siglo XIX y la primera mitad del XX. De acuerdo con esta manera de concebir el trabajo científico, las deducciones efectuadas a partir de las leyes matemáticas básicas debían ajustarse a las regularidades empíricas establecidas experimentalmente: las llamadas por Duhem "leyes experimentales". Este acuerdo, junto con la capacidad de las leyes de vehicular predicciones novedosas exitosas, habría de otorgar legitimidad empírica a los formalismos matemáticos empleados en la tarea científica. Por otra parte, sin embargo, muchas de las entidades a las que se hacía referencia en la interpretación física de las leyes matemáticas básicas eran entidades postuladas sin el respaldo de datos empíricos explícitos. Las ondas electromagnéticas constituyen un ejemplo especialmente claro de la manera en que la exploración matemática de una teoría física llevó a la postulación de una entidad que no había sido observada de manera directa. Dentro del marco de referencia de la epistemología empirista, sin embargo, solo aquello que está basado, de manera directa o indirecta, en experiencias sensoriales cuenta como conocimiento empírico. Así, si una teoría física postula la existencia de una entidad no susceptible de observación directa, se plantea la cuestión de cómo el postulado está atado a datos observacionales. Muchas de las afirmaciones de la física matematizada parecían carecer sin embargo de cualquier vínculo con datos observacionales.

El llamado "método de la reconstrucción racional" fue la herramienta empleada por el positivismo lógico para resolver esta dificultad; esto es, la de explicitar la manera en la que los postulados teóricos más abstractos de la física –los principios o "leyes" mencionadas por Newton– están conectados a pesar de todo con datos observacionales y pueden así ser legitimados en términos empiristas (Bailer-Jones 2009). Este método constituye el núcleo de la filosofía general de la ciencia desarrollada y defendida por este movimiento intelectual.

3.3.1. El método de la reconstrucción racional

De acuerdo con la famosa distinción introducida por Hans Reichenbach, para los positivistas el ámbito de la reflexión filosófica sobre la ciencia era el "contexto de justificación" y no el "contexto de descubrimiento" del conocimiento (Bailer-Jones 2009). La manera consagrada por el positivismo para explorar este contexto de justificación consistió en enfocarse exclusivamente en una de las fuentes de evidencia fundamentales a disposición de un intérprete de la actividad científica: las afirmaciones y dispositivos representacionales en los que se formula en primer lugar el conocimiento. En este sentido, el objetivo de la reflexión filosófica sobre la ciencia era, de acuerdo con Rudolf Carnap, la articulación de una "reconstrucción racional" del *lenguaje* de la física, el paradigma y la disciplina fundamental de la ciencia

moderna (Uebel 2007). El propósito de la filosofía de la ciencia era antes que nada la explicitación de los fundamentos epistemológicos de las *teorías* de la física a través de la construcción de lenguajes formales que desvelaran la estructura subyacente a estas teorías (Carnap 1991).

Para estos efectos, la filosofía positivista desarrolló una sofisticada interpretación de las teorías científicas, entendidas en la tradición newtoniana como el vehículo representacional por excelencia del conocimiento científico. De acuerdo con esta interpretación, las teorías científicas habrían de reconstruirse como *cálculos*, en el sentido metamatemático desarrollado por la lógica contemporánea: esto es, como axiomatizaciones de conjuntos de enunciados. La noción clave era la de un *cálculo interpretado*. Cada teoría empírica podía describirse de acuerdo con esta concepción a través de dos componentes: un cálculo axiomático abstracto y un componente que vincularía las expresiones del cálculo con los fenómenos naturales objeto de la teoría. Mientras que el primer componente comprendía los enunciados más generales de la teoría –sus leyes--, además del instrumental matemático y formal que la teoría emplea en la descripción de sus fenómenos y en la deducción de consecuencias, el segundo elemento está conformado por enunciados o "reglas" que conectan los términos del formalismo con términos observacionales que hacen referencia a entidades, propiedades y relaciones observadas directamente por los científicos. Estos enunciados, las llamadas "reglas de correspondencia", tenían la función de aportar "contenido empírico" al cálculo, en sí mismo puramente formal (Díez y Moulines 1999).

Al glosar las teorías científicas de esta manera, al "reconstruirlas racionalmente", los empiristas lógicos pretendían fundamentarlas epistémicamente, en el sentido de explicitar las razones por las que resulta justificado creer en sus afirmaciones. Esta fundamentación se daba en dos momentos: al describir las teorías como sistemas axiomáticos se evidenciaría cuáles de las aserciones que las componen podían derivarse de otras y ser en ese sentido más fundamentales epistémicamente; y, en segundo lugar, al explicitar "reglas de correspondencia", se vincularía los términos teóricos del formalismo con situaciones directamente observables, validando así las teorías en términos empíricos. Este segundo componente buscaba revelar, en concordancia con los compromisos empiristas de estos autores, que todo conocimiento científico, por distanciado que parezca del testimonio directo de los sentidos, procede y descansa en último término en la experiencia directa de una realidad teóricamente neutral (Carnap 1963).

El método de reconstrucción racional permitió a los positivistas desarrollar una concepción empirista de las explicaciones científicas y, sobre la base de esta, un modelo de las relaciones teóricas existentes entre las diferentes teorías que conforman la empresa científica. Así como Hume llenó un hiato en la imagen newtoniana de la ciencia con su explicitación de la naturaleza de las ideas de *ley* y

causalidad, en el seno del positivismo lógico se desarrolló con un nivel de precisión no disponible previamente una concepción empirista de la naturaleza de las explicaciones científicas y de sus relaciones.

3.3.2. **La concepción nomológico-deductiva de las explicaciones científicas**

En términos generales, una explicación constituye una respuesta a una pregunta sobre el "por qué" de un evento, hecho o regularidad; sobre las "razones" por las que algo tiene lugar en el mundo natural. El evento, hecho o regularidad cuya explicación se requiere es comúnmente denominado un "explanandum", mientras que aquello que explica el "explanandum" es denominado un "explanans". Típicamente, un explanandum de interés científico es un evento o estado de cosas que motiva algún tipo de perplejidad teórica.

Para que un explanans explique un explanandum debe haber algún tipo de relación de *dependencia* del segundo con respecto al primero: la ocurrencia de un explanandum *depende* en cierto sentido de la ocurrencia o existencia de un explanans. Una buena explicación consiste en una representación de esta relación de dependencia que es adecuadamente informativa para remover el estado de perplejidad teórica que motivó la indagación sobre el explanans. Una buena explicación aporta por tanto *comprensión* sobre la naturaleza de un segmento del mundo natural (Halina 2018). Ahora bien, ¿qué tipo de *dependencia* es aquella que subyace al vínculo entre un explanans y un explanandum?

El autor que mejor articuló el concepto de *explicación* resultante de la concepción positivista de la actividad científica es Carl G. Hempel. De acuerdo con Hempel, la relación de dependencia en cuestión en el caso de las explicaciones ofrecidas por las ciencias es primordialmente una relación *epistémica*. Explicar científicamente un explanans *e* es mostrar o poner de manifiesto cómo la "inesperabilidad" de *e*, la sorpresa o perplejidad que conlleva su ocurrencia, desaparece si se tienen en cuenta tanto las leyes que gobiernan su ocurrencia como ciertas "condiciones iniciales". Hempel denominó a esta característica "esperabilidad nómica" (*nomic expectability*) y, en el espíritu de la idea positivista de la reconstrucción racional, propuso una reconstrucción lógica del concepto de explicación de acuerdo con la cual la relación entre explanandum y explanans es un cierto tipo de relación *deductiva*. Un explanans hace "esperable" un explanandum en la medida en que del explanans se *infiere* deductivamente el explanandum. La concepción resultante se conoce como el "modelo nomológico-deductivo" o "modelo de cobertura legal" de la explicación científica (Hempel 1965).

Las explicaciones científicas, de acuerdo con este modelo, son argumentos deductivos. En una explicación, el explanandum es la conclusión de un argumento deductivo en el que el explanans está representado por una serie de premisas. No todos los argumentos deductivos, por supuesto, son explicaciones. La característica

distintiva de los argumentos deductivos que son explicaciones es que entre sus premisas hay al menos un enunciado expresivo de una ley. Considérese el siguiente ejemplo simple como una ilustración del modelo hempeliano. Un explanandum posible es la rotura de las cañerías de una casa. Este es un evento que puede conllevar sorpresa o perplejidad, cuya disolución requiere de la proposición de un explanans adecuado. Una explicación adecuada posible de este evento de acuerdo con el modelo nomológico deductivo es un argumento deductivo del siguiente tipo. Dado que: (a) tuvo lugar un descenso extremo de la temperatura en un intervalo de tiempo *t* en las inmediaciones de las cañerías *c*, (b) las cañerías estaban llenas de agua en *t*, (c) el descenso extremo de la temperatura congela el agua y (d) el congelamiento del agua contenida en una cañería lleva a su rompimiento; se sigue deductivamente que: (e) las cañerías en cuestión se rompieran. En este argumento, los enunciados (a) y (b) expresan hechos singulares –denominados por Hempel "condiciones iniciales"– y los enunciados (c) y (d) expresan regularidades empíricas con propiedades modales; esto es, leyes (Díez y Moulines 1999). Aunque las leyes en cuestión en este caso son regularidades en cuyo establecimiento y formulación no es preciso recurrir a formalismos matemáticos y no entrañan ningún interés científico especial, el ejemplo sirve para exhibir la forma que, de acuerdo con Hempel, tomarían *todas* las explicaciones científicas. El punto crucial del modelo hempeliano es que toda explicación científica conlleva forzosamente el empleo de leyes. Una pretendida explicación en la que no intervengan leyes es una explicación defectuosa o no es una explicación en absoluto (Psillos 2007). Este es el componente *nomológico*. Las leyes en cuestión, por otra parte, deben intervenir en la *deducción* de los explanandum de interés científico. Este el componente *deductivo*. La marca distintiva de las explicaciones científicas de acuerdo con este modelo es por tanto, en el mismo sentido en el que lo era para Newton, la *subsunción* de los fenómenos naturales bajo leyes.

3.3.3. **Reducción interteórica y unidad de la ciencia**

El modelo de cobertura legal tiene una ambición de generalidad irrestricta con respecto al ámbito de todas las posibles explicaciones científicas de fenómenos naturales. El modelo está diseñado para dar cuenta de dos tipos o formas de explicaciones que, de acuerdo con la concepción empirista de la ciencia, agotan el ámbito de las posibles explicaciones científicas del mundo natural: la explicación de eventos espacio-temporalmente delimitados y la explicación de regularidades empíricas o leyes. Este segundo tipo comprende la explicación de ciertas regularidades en términos de regularidades más "básicas" o fundamentales. Un ejemplo típico de este segundo tipo de explicación es la explicación de las regularidades del movimiento planetario establecidas por Kepler en términos de las leyes más básicas de la mecánica Newtoniana. Un conjunto de regularidades *x* es más

básica que un conjunto *y* en este contexto en la medida en que las regularidades que conforman *y* pueden ser *derivadas* a partir de *x* mientras que las regularidades que conforman *x* no pueden ser derivadas a partir de *y*. Las regularidades keplerianas de movimiento planetario pueden explicarse –esto es, subsumirse– en este sentido por las leyes de la mecánica newtoniana. Dado que para los empiristas lógicos las *teorías* científicas se individúan en términos de un conjunto de leyes, la derivación de un conjunto de leyes a partir de otro conjunto de leyes más básicas es un tipo de *reducción interteórica* (Kaplan 2017).

El modelo nomológico-deductivo fue empleado en este sentido por los positivistas no solo como una explicitación adecuada de la naturaleza de la explicación de fenómenos naturales, sino también como una manera adecuada de explicitar las relaciones entre las diferentes teorías que conforman la empresa científica. En el mismo sentido en el que un evento o una regularidad empírica, una teoría dada pueden ser el "explanandum" de una explicación. En ambos casos, la explicación consiste en la subsunción bajo leyes a través de una derivación deductiva. Cuando el explanans es una teoría, sin embargo, la subsunción en cuestión es un tipo de *reducción*. Para los positivistas lógicos, así, la noción meta-teórica de *reducción* hace referencia a un tipo de relación explicativa entre teorías (Díez y Moulines 1999).

La conjunción de esta idea de reducción con la convicción en la existencia de un conjunto reducido de leyes básicas explicativas de la totalidad de las regularidades empíricas susceptibles de tratamiento científico constituye el núcleo del modelo positivista de *unidad* de la ciencia.

El desvelamiento de las leyes fundamentales sería un objetivo de las ciencias físicas. El progreso de estas durante el siglo XIX, como indiqué, estuvo atado a una estrategia de matematización creciente en el tratamiento de sus fenómenos de interés. El sorprendente éxito predictivo de esta estrategia persuadió a los positivistas de que eventualmente los físicos podrían determinar un conjunto restringido de ecuaciones o leyes matemáticas que lograrían con respecto a la realidad física en su conjunto lo que las famosas cuatro ecuaciones de Maxwell habían logrado con respecto a los fenómenos electromagnéticos. Dada por supuesta esta posibilidad, las leyes constitutivas de todas las disciplinas científicas habrían de poder explicarse en términos de estas leyes básicas de manera que la ciencia en su conjunto se revelaría como una empresa unificada. Así, la unidad de la ciencia dependía de la posibilidad de reducir eventualmente todas las teorías científicas a teorías de la física. La fundamentación de esta idea de unidad reductiva de la ciencia en el modelo de cobertura legal, con su énfasis en el carácter deductivo de las explicaciones, hacía innecesarias las especulaciones metafísicas que, en opinión de los positivistas, contaminaban las ideas mecanicistas de reducción y unidad de la ciencia El concepto de *reducción*, empleado de manera confusa en la tradición mecanicista, podía aclararse de esta manera como un tipo de relación deductiva entre teorías

diferentes. Una teoría *reduce* a otra en la medida en que tanto las leyes que conforman la segunda como los fenómenos empíricos que se derivan de estas (los hechos particulares que explica la teoría) pueden deducirse de los enunciados de la primera teoría (Bechtel y Hamilton 2007).

En un artículo publicado en 1958, Paul Oppenheim y Hilary Putnam dieron forma sistemática a estas ideas a través de la formulación de un modelo de reducción global de la ciencia a la física. De acuerdo con este modelo, las disciplinas que conforman la empresa científica están organizadas en una jerarquía. A pesar de la renuencia positivista a recurrir a categorías metafísicas, Oppenheim y Putnam vinculaban esta jerarquía de disciplinas con una jerarquía de "niveles de organización" de la realidad. En un extremo de la jerarquía se encuentran las disciplinas que estudian las sociedades humanas y en el otro extremo la disciplina que estudia las propiedades fundamentales de la materia. La idea era que a través de una serie progresiva de reducciones (denominadas por ellos "micro-reducciones"), todas las teorías científicas podían reducirse a teorías físicas. Así, las teorías sociológicas debían poder reducirse a teorías psicológicas, estas a teorías biológicas, estas a teorías químicas, hasta llegar a las teorías más fundamentales de la ciencia. Oppenheim y Putnam citaron en su artículo el modelo de McCulloch y Pitt como un ejemplo exitoso de micro-reducción. En particular, un ejemplo de reducción de fenómenos psicológicos (como, en sus palabras, "el aprendizaje, la inteligencia y la percepción") a fenómenos biológicos; esto es, en este caso a actividad celular (Oppenheim y Putnam 1958; Piccinini 2004a).

El progreso de una teoría científica podía estimarse de acuerdo con esto en términos de lo cerca que se encontraba de ser reducible a una teoría de un nivel inferior en la jerarquía y el progreso de la ciencia en su conjunto consistía en su gradual unificación alrededor de la física. La idea general era que si todas las disciplinas científicas pudieran reducirse a una sola y dentro de esa disciplina hubiera una teoría que redujera las demás, podría interpretarse el desarrollo científico como un avance hacia una unificación progresiva de todo el conocimiento de la realidad (Díez y Moulines 1999). La convicción en el progreso del conocimiento científico y su interpretación de la idea de progreso en estos términos reduccionistas era parte de lo que justificaba la creencia positivista en la unidad de la ciencia.

3.3.4. El fracaso del modelo reductivo de las relaciones interteóricas: traducción y leyes puente

El desarrollo del concepto positivista de reducción no se fundó en el estudio de la historia de la ciencia sino en consideraciones abstractas relativas a una fundamentación empirista del conocimiento científico. A pesar de esto, el concepto debía poder aplicarse de alguna manera a casos concretos de la historia de la ciencia, so pena de dejar sin contenido empírico el proyecto positivista. El escollo principal

que se interponía en esta aplicación práctica era la inconmensurabilidad de los vocabularios científicos. La existencia de relaciones deductivas entre teorías –el núcleo del concepto de reducción– exige que estas estén expresadas en un vocabulario común (Bechtel y Hamilton 2007).

Esta dificultad estuvo en la primera plana de la agenda positivista desde el comienzo del movimiento. Carnap y Hempel contemplaron la posibilidad de superarla a través de la construcción de un único lenguaje universal de la ciencia al cual pudieran traducirse todos los enunciados científicos. Este lenguaje universal sería el lenguaje de la física reconstruido como un cálculo deductivo. Los predicados y enunciados constitutivos de cualquier teoría científica debían de acuerdo con esto ser susceptibles de traducción a predicados y enunciados físicos. Un ejemplo interesante a este respecto es la contemplada traducibilidad de predicados y enunciados psicológicos a conceptos y enunciados físicos. Un predicado psicológico como "*x* está excitado" se traduciría a un predicado "físico" del tipo: "el cuerpo de *x* tiene una estructura física caracterizada por un pulso y frecuencia de respiración altas, por respuestas vehementes y factualmente insatisfactorias a preguntas presentadas, por la ocurrencia de movimientos agitados como respuesta a ciertos estímulos, etc." (Walter y Eronen 2014). La traducibilidad de lo psicológico a lo físico presupone en este caso, como resulta fácil advertir, un tipo de conductismo con respecto a lo mental. Aunque Carnap y otros positivistas eran conductistas, el núcleo de la idea de reducción no exige el compromiso con ninguna idea específica sobre la manera en que las teorías de una disciplina habrían de reducirse a las teorías de un nivel inferior y, como se vio, Putnam y Oppenheim usaron en cambio el modelo de McCulloch y Pitt como ejemplo de reducción de lo psicológico. Lo que exige la idea positivista de reducción es algún tipo de conmensurabilidad entre vocabularios que permita establecer relaciones deductivas entre teorías. La estrategia de satisfacer esta exigencia mediante la construcción de un lenguaje universal al que pudieran traducirse todos los enunciados significativos de la ciencia se reveló especialmente problemático y los positivistas terminaron recurriendo a estrategias más modestas.

La exigencia de traducibilidad suponía algún tipo de sinonimia entre predicados de diferentes teorías. El fracaso en el establecimiento de estas equivalencias de significado condujo a algunos a pensar que aun cuando los predicados de diferentes teorías fueran estrictamente intraducibles, podía lograrse un tipo de conmensurabilidad que preservara la posibilidad de relaciones deductivas a través del establecimiento de conexiones menos estrictas entre estos predicados. Ernest Nagel propuso en este contexto conectar los vocabularios de diferentes teorías a través de lo que denominó "leyes puente" (Bechtel y Hamilton 2007). La idea básica de Nagel era que aun cuando los predicados de diferentes teorías tuvieran significados diferentes podían designar las mismas propiedades. Así, la estrategia consistía en tomar los predicados de diferentes teorías y articular enunciados

generales (leyes) que incluyeran estos predicados y expresaran un tipo de equivalencia no semántica entre los mismos. Tales enunciados tomarían la forma de bicondicionales. Suponiendo que B_1 es un predicado de una teoría y B_2 un predicado de una teoría diferente, una ley puente tendría la siguiente forma: "Algo es un B_2 si y solo si es un B_1". Un ejemplo de una ley de este tipo sería: "Un metal tiene una cierta conductividad eléctrica y está a cierta temperatura si y solo si tiene cierta conductividad térmica". Mediante esta ley puente se conectarían predicados de diferentes teorías con diferentes significados ("conductividad eléctrica" y "temperatura", por un lado, y "conductividad térmica") de una manera que preservaría la posibilidad de establecer relaciones deductivas entre ellas. A través de estas leyes puente, pretendía Nagel, podría mostrarse cómo las leyes de una teoría podían deducirse de las leyes de otra teoría más básica, de manera que la segunda cubriría las mismas regularidades empíricas que la primera y reducirla, en el sentido pretendido por los positivistas (Nagel 1961).

El esfuerzo de aplicar este modelo de reducción a casos específicos de la historia de la ciencia, sin embargo, se reveló igual de problemático que el esfuerzo de traducir todos los enunciados de la ciencia a un lenguaje fisicalista universal. Este fracaso promovió pronto un escepticismo general sobre el concepto positivista de reducción como una buena manera de entender las relaciones intertéoricas. Incluso en el caso favorito de Nagel, la pretendida reducción de la termodinámica a la mecánica estadística, el escrutinio detallado puso de manifiesto que el modelo no podía aplicarse efectivamente: conceptos termodinámicos centrales como el de entropía están asociados con una variedad de conceptos diferentes en la mecánica estadística, que no corresponden ni separada ni conjuntamente con los primeros de la manera requerida por el modelo (Sklar 1999). De manera poco sorprendente, el intento de aplicar el modelo a casos más problemáticos como el de la reducción de teorías biológicas a teorías físico-químicas o el de la reducción de teorías psicológicas a teorías biológicas resultó infructuoso.

3.4 Anti-reduccionismo y ciencias "especiales"

El descrédito del modelo nageliano, la versión mejor desarrollada del concepto positivista de reducción, llevó desde finales de los años sesenta a explorar maneras diferentes de interpretar las relaciones intertéoricas. Estas interpretaciones alternativas supusieron también un cuestionamiento de la idea de unidad de la ciencia que subyacía y motivaba la conceptualización positivista de la noción de reducción. David Hull en relación con la biología y Jerry Fodor con la psicología alcanzaron conclusiones equivalentes con respecto a las razones del fracaso del modelo nageliano. De acuerdo con estos autores, el origen del problema residía en la *realizabilidad múltiple* de los propiedades referidas por los predicados de las teorías

no físicas (las teorías de las "ciencias especiales", como empezó a denominarse a las disciplinas que el modelo positivista creía reducibles).

Hull y Fodor se propusieron mostrar que predicados característicos de la biología y la psicología, respectivamente, podían estar relacionados en diferentes ocasiones con diferentes predicados y caer bajo diferentes generalizaciones de ciencias más básicas (Bechtel y Hamilton 2007). Los genes en la genética mendeliana, por ejemplo, se caracterizan en términos de rasgos fenotípicos que codifican. En la genética molecular, en cambio, los genes se caracterizan en términos de su constitución molecular. Hull señaló que diferentes constituciones moleculares pueden generar el mismo rasgo fenotípico, de manera que los genes mendelianos son *múltiplemente realizables*. Este hecho impide la formulación de una ley puente en el sentido de Nagel. Las regularidades empíricas que cubren los enunciados de ambas teorías son diferentes de una forma que bloquea la reducción deductiva de una a la otra. De la misma manera, sostuvo Fodor, las propiedades referidas por los predicados de las teorías psicológicas son *múltiplemente realizables* por diferentes propiedades estudiadas por teorías neurobiológicas (Fodor 1974). La conclusión de Fodor es equivalente a la de Hull: la realizabilidad múltiple bloquea la deducción de las regularidades cubiertas por las teorías psicológicas a partir de regularidades neurobiológicas y por tanto también su reducción.

Fodor sostuvo que esta circunstancia era común a todas las ciencias especiales, por lo que la idea positivista de reducción constituía un artefacto filosófico sin ningún correlato empírico. La realizabilidad múltiple sería una característica de todos o la mayoría de los fenómenos estudiados por estas ciencias (Fodor 1974). Lo que esto significa es que las ciencias especiales estudian sistemas físicos individuados en términos "funcionales" y no en términos "estructurales". Al estudiar un sistema en términos funcionales, el aspecto crucial es la comprensión de lo que el sistema hace, no de *cómo* lo hace. Considérese un ejemplo típico como ilustración. Un corazón es un sistema cuya función es bombear sangre; de manera más precisa, acopiar sangre desoxigenada y bombearla hacia los pulmones, donde es reoxigenada. La estructura física específica de un corazón no resulta en este sentido relevante para individuar el sistema. Un corazón artificial puede efectuar el mismo trabajo igual de bien que un corazón no artificial. Por otra parte, los cocodrilos y los humanos tienen corazones con estructuras diferentes a la de los reptiles, por ejemplo, por no hablar de las diferencias entre la manera en la que opera el corazón de una ballena blanca en comparación con el de una especie cualquiera de pájaro (Bermúdez 2014). Los corazones son en este sentido sistemas múltiplemente realizables, de manera que en su estudio resultaría equivocado enfocarse en los detalles relativos a cómo cada tipo de corazón efectúa su trabajo.

La idea de Fodor era que las propiedades referidas por los predicados de las ciencias especiales son típicamente propiedades *disposicionales*. Estas son

propiedades que designan formas en que los objetos que caen bajo ellas se comportan en ciertas circunstancias. Ejemplos típicos son propiedades como *fragilidad*, *elasticidad* o *rojo*. Un objeto es "frágil", por ejemplo, si en caso de que se aplicara sobre él cierta presión, se quebraría. Una superficie es "roja" si en caso de que incidiera sobre ella luz blanca absorbería tales y tales frecuencias del espectro (Díez y Moulines 1999). Una manera común de especificar la naturaleza de este tipo de propiedades es señalar que designan poderes causales de objetos que se manifiestan de diferentes maneras en un rango específico de circunstancias propiciatorias.

El anti-reduccionismo fodoriano acepta un tipo genérico de fisicalismo de acuerdo con el cual las propiedades disposicionales "descansan" en o son "realizadas" por propiedades físicas. Si un objeto es frágil o rojo lo es en virtud de su composición microfísica. Ahora bien, la propiedad microfísica que realiza una propiedad disposicional no es la misma en todos los casos en que se ejemplifica esta última. En cada caso particular de un objeto frágil, su fragilidad se debe a propiedades microfísicas del objeto en cuestión. En diferentes tipos de objetos, sin embargo, la propiedad realizadora es diferente. Cuando una propiedad es múltiplemente realizable en este sentido no cabe explicar su dependencia de una propiedad microfísica reduciéndola o identificándola con ella (Piccinini y Maley 2014).

Las ciencias especiales, de acuerdo con el anti-reduccionismo desarrollado por Fodor, estudian disposiciones en la naturaleza: *capacidades* específicas de ciertos sistemas que se manifiestan en un rango específico de circunstancias y que pueden ser realizadas por diferentes constituciones micro-físicas (Fodor 1974).

3.4.1. **Autonomía explicativa y desunión de la ciencia**

La omnipresencia de la realizabilidad múltiple bloquea la posibilidad de reducción de los fenómenos estudiados por las ciencias especiales. Pero esta reducibilidad era la condición de posibilidad de la unidad de la ciencia de acuerdo con los positivistas. Oponiéndose al antiguo ideal mecanicista acogido por el positivismo lógico, Fodor extrajo la consecuencia inevitable de sus ideas anti-reduccionistas y afirmó de manera categórica la desunión de la empresa científica (Fodor 1974; Ruphy 2016).

Además de por el fracaso del modelo reductivo que él mismo se ocupó de poner de manifiesto, Fodor se opuso al modelo de unidad de los positivistas por dos características del mismo que estimaba contraintuitivas. En primer lugar, el modelo positivista suponía que la ciencia era una empresa progresiva y que este progreso consistía en la gradual expansión explicativa de la física. Los fenómenos que son estudiados hoy por las ciencias especiales serían eventualmente, en la medida en que estas ciencias progresan, parte de la jurisdicción teórica de la física. Así, el modelo tiene la curiosa consecuencia de que mientras más avancen estas ciencias, más cerca estarían de desaparecer. En segundo lugar, el reduccionismo suponía que los fenómenos estudiados por las ciencias especiales no tenían una realidad por propio

derecho. Los conceptos empleados para hacer referencia a estos fenómenos, como indiqué, debían ser "legitimados" de manera fisicalista como descripciones oblicuas de fenómenos físicos. Estos conceptos, dicho en otros términos, no hacían referencia a poderes causales, por lo que no podían emplearse realmente en la explicación de ningún fenómeno. Las ciencias especiales no podían ser de acuerdo con esto ciencias explicativas que desvelaran el origen causal de ningún fenómeno sino solo herramientas heurísticas en el desarrollo de la ciencia básica (Bechtel y Hamilton 2007).

En contra de estas dos consecuencias contraintuitivas, que ponían en cuestión los poderes epistémicos de todas las ciencias no básicas, Fodor defendió la *autonomía* del nivel funcional con respecto al nivel físico. Los conceptos de las ciencias funcionales hacen referencia a poderes causales de sistemas físicos que, en virtud de la realizabilidad múltiple, son irreducibles a propiedades físicas. Las ciencias especiales siguen una agenda teórica *independiente* de la agenda teórica de la ciencia básica en el sentido de que ofrecen explicaciones distintivas de fenómenos concebidos en un nivel distintivo de descripción. Estas explicaciones y este nivel de descripción no se relacionarían de manera *epistémicamente* relevante con las explicaciones y el nivel de descripción físico (Kaplan 2017).

La jerarquía de niveles propuesta en el modelo de Oppenheim y Putnam es reemplazada en la nueva visión fodoriana de la ciencia por una partición de dos niveles: el nivel funcional y el nivel físico. Mientras que los niveles del modelo de Oppenheim y Putnam eran niveles de organización de la realidad tanto como niveles de organización de la ciencia que estudia esa realidad, los niveles articulados por Fodor son niveles de descripción. Existe un único nivel de realidad que puede describirse en dos niveles diferentes: el funcional y el micro-físico (Elber-Dorozko y Shagrir 2019). En un nivel se estudian las capacidades de un sistema, su perfil funcional, sin hacer alusión a sus realizadores micro-físicos; en el otro nivel se estudia la composición física de un sistema sin hacer referencia a su interpretación funcional.

De acuerdo con este modelo, la ciencia estaría bifurcada por principio en la investigación de lo funcional y la investigación de lo físico, donde ambas investigaciones procederían de manera autónoma una de la otra. La consecuencia de la autonomía es así la desunión de la empresa científica.

Mientras que en las ciencias biológicas el nuevo modelo tuvo un éxito modesto y la cuestión de la reducción siguió siendo discutida más allá de los márgenes de la filosofía positivista de la ciencia, como consecuencia de lo cual emergería un nuevo tipo de mecanicismo (como indicaré en el capítulo final), la nueva empresa interdisciplinaria de la ciencia cognitiva lo acogió con entusiasmo. El "funcionalismo" se convirtió en la filosofía oficial de la "nueva ciencia de la mente" (Gardner 1985; Bechtel et al. 1998).

3.5. Funcionalismo sobre la mente: la realizabilidad múltiple de los estados cognitivos

Las consideraciones reseñadas sobre irreducibilidad, realizabilidad múltiple y autonomía surgieron como resultado del escrutinio de la filosofía positivista de la ciencia y tienen en consecuencia un carácter general: representan ideas generales sobre la naturaleza de la ciencia y sus productos. El concepto de *realizabilidad múltiple*, sin embargo, fue acuñado y adquirió especial notoriedad en el contexto de reflexiones sobre un ámbito circunscrito de la ciencia: la cognición. El responsable de la introducción de este concepto (en el contexto de reflexiones sobre la individuación y la metafísica de los estados mentales) fue un filósofo inicialmente defensor del modelo reduccionista: Hilary Putnam. En el curso de los años sesenta, Putnam abandonó sus convicciones reduccionistas y articuló una doctrina "funcionalista" sobre la naturaleza de la mente fundada en el concepto de realizabilidad múltiple (Putnam 1960; 1967; 2010b). De acuerdo con esta doctrina, la mente y los fenómenos mentales conforman un ámbito de estudio *autónomo* con respecto al ámbito de estudio de la neurobiología. Los fenómenos mentales y las causas de la conducta inteligente, en contra de lo pretendido por McCulloch y Pitt, son irreducibles a fenómenos neuronales. El sentido de autonomía en juego era el explicitado de manera sistemática por Fodor, que adoptó y desarrolló las ideas de Putnam, y se convirtió a su vez en uno de los principales defensores del funcionalismo sobre la mente (Piccinini 2004a).

La defensa del funcionalismo con respecto a un tipo de fenómeno es inseparable de la defensa de la existencia de realizabilidad múltiple en las propiedades y procesos causales que lo identifican. En relación con los fenómenos mentales, esta defensa tomó diferentes formas en la literatura funcionalista temprana. En todos los casos, sin embargo, la estrategia consistió en el uso de analogías entre estados mentales típicos de los seres humanos y otros estados, propiedades y entidades cuya individuación se efectuaba en términos funcionales. Así, por ejemplo, Putnam señaló la analogía entre los estados mentales de los seres humanos y los de otros animales inteligentes; así como los posibles estados mentales de entidades como robots, marcianos y ángeles (Putnam 1960). La atribución e individuación de estados mentales en todos estos casos dependería de factores funcionales y no neurobiológicos. Animales no humanos, robots, marcianos y ángeles, pretendidamente, pueden poseer, igual que los seres humanos, estados mentales como creencias o dolores. Los cerebros de estas entidades no humanas, sin embargo, pueden estar organizados neuronalmente de maneras diferentes, estar constituidos por estructuras micro-físicas diferentes o no existir en absoluto. En consecuencia, los estados mentales pueden ser "realizados" por estados físicos diferentes y no pueden identificarse o reducirse a ellos.

Fodor, por otra parte, sugirió que la naturaleza de los estados mentales de los humanos puede entenderse en analogía con la de entidades como relojes o motores

(Fodor 1968). Relojes y motores pueden estar hechos a partir de componentes físicamente diferentes, de manera que *ser un reloj* o *ser un motor* designa una propiedad individuada funcional y no físicamente (toda vez que estos artefactos se conceptualicen, como es usual, en términos de aquello para lo que son usados). Por último, Fodor sugirió también que los estados mentales podían entenderse en analogía con rasgos biológicos como la posesión de alas, pies o corazones. Insectos, aves y murciélagos tienen alas y pies constituidos micro-físicamente de manera diferente. Esta diferencia, sin embargo, no hace que estas características dejen de ser categorizadas como alas y pies; lo que pone de manifiesto que se trata de características funcionales, múltiplemente realizables.

Mediante todas estas analogías los funcionalistas pretendían otorgar credibilidad a la tesis según la cual los estados mentales son estados múltiplemente realizables en cuya individuación resulta irrelevante recurrir a consideraciones neurobiológicas. La oposición a esta convicción era descartada como una manifestación de "chovinismo" de especie (Block 1978); esto es, una injustificada atención a las contingencias de la biología humana con respecto al ámbito de la inteligencia y de la mente.

3.5.1. Realizabilidad múltiple y máquinas de Turing

La analogía original que motivó el funcionalismo sobre la mente y en la cual residía su justificación fundamental era sin embargo una analogía diferente. Aunque la alusión a estados mentales posibles en otras entidades, a artefactos mecánicos como relojes o motores y a rasgos biológicos hacía conceptualmente verosímil la idea de la realizabilidad múltiple, la analogía que otorgaba credibilidad empírica al funcionalismo sobre la mente era la analogía entre la mente y los dispositivos computacionales inventados por Turing.

De acuerdo con la analogía original de Putnam, los estados mentales son análogos a los estados "internos" de una máquina de Turing. Una máquina de Turing simple, como indiqué en el capítulo anterior, se individúa en términos de una serie de instrucciones "almacenadas" en una tabla que especifica estados de la máquina y transiciones entre esos estados. Turing habló ocasionalmente de los estados en los que se encuentra una máquina en cada momento de la ejecución de sus instrucciones como estados "internos". La analogía de Putnam hace referencia a estos estados.

El aspecto clave en el que se basa la analogía es que una máquina de Turing efectivamente construida es un sistema físico tal que los estados y procesos que definen su comportamiento son estados y procesos funcionales. La máquina recibe *inputs*, los procesa selectivamente de acuerdo con una serie de estados internos y genera *outputs*, y estos *inputs*, *outputs* y estados internos pueden "realizarse" en soportes físicos diferentes. Lo que importa en la individuación de estas máquinas es su "organización funcional"; esto es, en términos generales, el conjunto de estados que definen la manera en que la máquina genera un conjunto definido de *outputs* a

partir de un conjunto específico de *inputs*. La tabla de una máquina simple de Turing es en este sentido una organización funcional (Piccinini 2004a). En una máquina de Turing física, por otra parte, las transiciones entre los estados de la máquina son eventos causales y las relaciones entre cada estado y entre el *input* y el *output* son relaciones causales. Un sistema de este tipo exhibe entonces cómo un proceso es susceptible de ser *descrito* como un proceso causal sin necesidad de hacer referencia a ninguna estructura física específica. Este era el aspecto que interesaba a Putnam y que motivó su analogía.

El diseño de Turing muestra cómo un conjunto de estados funcionales puede ser causalmente eficaz en la producción de un cierto tipo de conductas. La mente, de acuerdo con Putnam, estaría conformada por un conjunto de estados funcionales de este tipo y, en este sentido, podría concebirse como la "organización funcional" del cerebro (Putnam 1967). Para el funcionalismo, la estructura bio-química específica del sistema nervioso es irrelevante para efectos de explicar la cognición. Lo importante es la existencia de un sistema físico capaz de realizar la organización funcional en que consiste la mente. Si este sistema está hecho de queso suizo, llegó a afirmar Putnam, es irrelevante (Putnam 2010a).

3.5.2. Conocimiento de la cognición, modelos funcionales y autonomía

El conocimiento de la mente es de acuerdo con el tipo de funcionalismo descrito similar al conocimiento de la organización funcional de una máquina de Turing. El científico de la mente estudia los organismos inteligentes desde la perspectiva en la que un lógico estudia las instrucciones de una máquina de este tipo, mientras que el científico del cerebro estudia el mismo fenómeno desde la perspectiva del ingeniero que examina el funcionamiento del artefacto físico: "[The] important thing is that descriptions of the functional organization of a system are logically different in kind either from descriptions of its physical–chemical composition" (Putnam 1967, p. 200). Esto no supone, sin embargo, que el estudio de la mente no se ocupe de la determinación de procesos causales y no constituya una disciplina explicativa. Así como un investigador que se encuentre con un sistema físico que recibe *inputs* y produce *outputs* de manera aparentemente no aleatoria puede estudiar su comportamiento sin conocer nada sobre su estructura interna, el investigador de la mente puede estudiar el comportamiento de un organismo sin necesidad de preocuparse de los sistemas físicos que subyacen a la misma. En ambos casos, el trabajo de estos investigadores consiste en buscar regularidades en la conducta y postular de una serie de estados funcionales –esto es, una organización funcional– que permitan *predecir* y explicar esa conducta. Este trabajo puede interpretarse, y ha sido usualmente interpretado en la literatura funcionalista sobre la mente, como un tipo de *ingeniería inversa* (Bermúdez 2014). La postulación de una organización

funcional resultante de un estudio de ingeniería inversa de este tipo puede interpretarse como la formulación de un modelo funcional.

Un modelo funcional en este sentido no es un modelo mecanicista. En contra de la convicción mecanicista, el funcionalismo pretende que en el establecimiento de dependencias causales que determinan y *explican* el comportamiento de sistemas con capacidades "funcionales" es innecesaria la descripción de mecanismos en el nivel físico. Como indiqué en el primer capítulo, el mecanicismo está comprometido en cambio con una heurística conforme a la cual la explicación de las capacidades de un sistema físico requiere tanto la descomposición funcional como la descomposición estructural de los mismos. Una explicación mecánica exhibe por qué una capacidad tiene lugar mostrando cómo tiene lugar y esto último exige vincular las funciones que efectúa un sistema con los componentes físicos que lo constituyen.

Una *consecuencia* del funcionalismo es la sanción de una aproximación "descendente" (*top-down approach*) al estudio de la cognición (Bermúdez 2014). De acuerdo con esta perspectiva, el estudio de la cognición procede a partir de la observación de la conducta, la determinación de regularidades funcionales y la postulación de modelos funcionales que den cuenta de esas regularidades. La corrección de los modelos cognitivos resultantes de este trabajo es una cuestión empírica. Si un modelo es correcto y por tanto explicativo de algún tipo de capacidad es algo que ha de evaluarse de acuerdo con procedimientos de corroboración usuales en el trabajo científico. Esto es, principalmente, mediante pruebas experimentales de las hipótesis implicadas por el modelo y mediante su comparación con modelos alternativos en términos de su grado de satisfacción de valores teóricos usuales como respaldo evidencial, consistencia interna, sistematicidad, etc. De manera crucial, sin embargo, la corrección de un modelo funcional no depende de factores relativos a su implementación física. Aunque algún hallazgo neurobiológico relativo al funcionamiento del cerebro puede emplearse para respaldar un modelo cognitivo, al mostrar cómo este podría implementarse en el cerebro y en ese sentido darle credibilidad adicional, ningún hallazgo de este tipo puede descartar un modelo cognitivo (Weiskopf 2017; 2019). Para los funcionalistas, la mente es el "software" del cerebro, en un sentido que explicitaré en la próxima sección. El mismo software, tal es el punto crucial, puede ser implementado por diferentes tipos de hardware; incluso por tipos de hardware desconocidos. Este es el sentido en el que el ámbito de estudio de la cognición es *autónomo*.

De esta manera, partiendo de la misma herramienta, Putnam llegó a una conclusión exactamente opuesta a la de McCulloch y Pitt. Mientras que estos últimos usaron el formalismo de Turing para explicar la cognición en términos de mecanismos neurobiológicos, el funcionalismo se sirve del mismo para mostrar la irreducibilidad y autonomía de la cognición respecto del cerebro.

3.5.3. Funcionalismo y computacionalismo

El funcionalismo sobre la mente descrito hasta ahora no representa ningún tipo de computacionalismo. Putnam no usa el formalismo de Turing para proponer ninguna hipótesis explicativa *específica* sobre las causas de la conducta inteligente sino solo para sostener que cualquier hipótesis en este sentido tiene que conllevar la postulación de estados y organizaciones *funcionales*. No todas las organizaciones funcionales, sin embargo, son organizaciones computacionales. Volviendo a la analogía de Fodor, es posible explicar el funcionamiento de un motor o de un corazón a través de la descripción de una organización funcional. Pero mientras que la organización funcional de una máquina de Turing es una organización computacional, no cabe decir lo mismo en el caso del motor. Una máquina de Turing física en buen estado efectúa computaciones y las relaciones entre sus *inputs*, estados internos y *outputs* son relaciones computacionales. Afirmar que un motor o un corazón efectúan computaciones y que las relaciones entre sus componentes son relaciones computacionales, en cambio, supone una extensión problemática del concepto de *computación*. El punto importante es señalar que el funcionalismo sobre la mente es una tesis metafísica sobre la individuación de los estados de ciertos sistemas que no implica ninguna forma de computacionalismo. El computacionalismo no es en primer lugar una tesis metafísica sino una tesis empírica sobre el tipo de organización funcional de ciertos sistemas. Es posible ser funcionalista sin ser computacionalista, a pesar de la confusión de ambas tesis que empezó a proliferar desde los años setenta en la reflexión sobre la cognición (Piccinini 2010).

3.6. Computacionalismo funcionalista

A pesar de su independencia conceptual, el funcionalismo y el computacionalismo no son doctrinas estrictamente incompatibles. Al agregar a la idea funcionalista según la cual la mente es la organización funcional del cerebro la idea de que esta organización es computacional resulta lo que puede denominarse "computacionalismo funcionalista". El desarrollo del funcionalismo en una hipótesis empírica sobre el funcionamiento de la mente en esta dirección fue principalmente obra de Jerry Fodor (Piccinini 2004a).

El componente *funcionalista* del computacionalismo desarrollado por Fodor reside en la conceptualización de la mente como la organización funcional de un sistema físico; en el caso de los seres humanos, el cerebro. El estudio de la mente es de acuerdo con esto el estudio de estados funcionales y de su organización en un sistema físico. Este estudio tomaba para Fodor la forma de un "análisis funcional" (Fodor 1968a). Un análisis funcional no es nada distinto a lo que en la heurística mecanicista descrita en el primer capítulo se denomina una "descomposición funcional de un sistema". El objetivo de un análisis de este tipo es identificar las

funciones o capacidades características de un sistema y postular modelos en los cuales estas capacidades resultan o se generan a partir de la interacción de estados funcionales del mismo. Cada modelo de este tipo tiene diferentes "realizaciones mecánicas" posibles, por lo que es irreducible a estas.

En el caso de sistemas físicos como motores, relojes o estómagos, los modelos resultantes de un análisis funcional no son modelos computacionales. En el caso de los sistemas cognitivos, sin embargo, de acuerdo con Fodor, los modelos resultantes de su análisis funcional *deben* ser modelos computacionales. Para Fodor, un modelo funcional de un sistema con capacidades cognitivas *habría de* tomar la forma de una lista de instrucciones de computación en el sentido de una tabla de una máquina de Turing: "the paradigmatic psychological theory is a list of instructions for producing behavior" (Fodor 1968b, p. 630).

Fodor fundió de esta manera dos conceptos diferentes implicados en la formulación de modelos funcionales: el concepto de análisis funcional y el concepto de "análisis de tareas" (*task analysis*) (Piccinini 2010). Un análisis funcional, de acuerdo con lo dicho, modela un sistema en términos de un conjunto de componentes individuados funcionalmente y de interacciones entre estos componentes. Estos componentes y sus interacciones se representan como los factores causalmente responsables de alguna de las capacidades de un sistema. La capacidad de un motor para generar energía mecánica, por ejemplo, puede explicarse mediante un modelo en el que se identifiquen componentes como válvulas, cilindros y pistones y en el que se detalle cómo estos componentes interactúan. Un análisis de tareas, en cambio, modela un sistema en términos de la ejecución de tipos finitos de secuencias de operaciones, las cuales pueden analizarse a su vez en términos de subrutinas de ejecución de secuencias de operaciones pero que no necesitan atribuirse a componentes del sistema y a sus funciones. La explicación de una capacidad de un sistema en términos del rol causal de una serie de *instrucciones* individuadas por un análisis de este tipo es el tipo de explicación de la cognición consagrada por el funcionalismo.

La fusión fodoriana de estos dos conceptos resultaría enormemente influyente y terminaría convirtiéndose en un elemento nuclear del programa teórico de la ciencia cognitiva clásica. De acuerdo con el computacionalismo funcionalista de Fodor, una teoría cognitiva se entendería a partir de entonces como un análisis funcional, el cual consistiría a su vez en la formulación de una secuencia de operaciones de procesamiento computacional (Piccinini 2004a). La justificación de esta fusión entre lo funcional y lo computacional, sin embargo, como indicaré en el próximo capítulo, no carece de serios problemas.

3.6.1. **La mente como el "software del cerebro"**

McCulloch y Pitt emplearon en su modelo general de la cognición mecanismos con poderes computacionales modestos. Sus redes neuronales eran capaces de computar funciones lógicas básicas de una manera que no requería la ejecución de instrucciones complejas como en el caso de algunas máquinas de Turing. En términos actuales, las redes de McCulloch y Pitt son computacionalmente equivalentes a autómatas finitos. Los autómatas finitos son mecanismos de computación básicos cuyo comportamiento está determinado por un conjunto limitado de funciones de transición fijas y no por el "seguimiento" de una serie compleja de instrucciones almacenadas en un soporte externo (Patterson y Hennesy 2014). Los autómatas finitos, así, son dispositivos no-programables. McCulloch y Pitt pensaban que la conexión adecuada de diferentes dispositivos de este tipo tendría el poder computacional suficiente para modelar causalmente todo lo que el cerebro humano podía computar (Piccinini 2004c). En su modelo, sin embargo, no tenía lugar la idea de un mecanismo lógico programable. Esta idea, en cambio, es central para el computacionalismo fodoriano.

Al analizar funcionalmente un sistema para efectos de explicar una capacidad cognitiva, el teórico de la cognición, de acuerdo con Fodor, habría de proponer una lista de instrucciones que el sistema ejecuta en el mismo sentido en el que un computador digital programable ejecuta una lista de instrucciones. Dado un *input* determinado, un artefacto de este tipo efectúa una serie de operaciones automáticas de acuerdo con un conjunto definido de instrucciones y genera un *output*. De la misma manera en que la conducta flexible y sistemática de un computador se *explica* apelando a una serie de operaciones efectuadas por componentes internos del mismo acuerdo con una serie definida de instrucciones almacenadas en otros componentes internos, la explicación de la conducta de un organismo inteligente se explicaría postulando operaciones efectuadas por componentes internos de acuerdo con patrones causales definidos que pueden *describirse* como la ejecución de programas computacionales. Cada evento de manifestación de una capacidad cognitiva podría entenderse así como el resultado de un proceso causal de ejecución de instrucciones por componentes internos de un sistema (Piccinini 2010).

La analogía original de Putnam entre mente y máquina se extiende así en las manos de Fodor. Mientras que en el caso de Putnam el aspecto importante de la comparación era solo el carácter funcional de los estados que generan causalmente el comportamiento en un sistema, Fodor explota otra dimensión de la analogía. Tanto las mentes como los computadores, de acuerdo con este, manipulan estructuras combinacionales complejas de maneras sistemáticas. La descripción de los *inputs* y *outputs* propios de ambos tipos de sistemas, así como de las relaciones entre ellos, exigiría por tanto el empleo de reglas recursivas. Estos *inputs* y *outputs* exhiben la estructura y sistematicidad de un lenguaje, en virtud de lo cual pueden combinarse en

configuraciones de una complejidad potencialmente ilimitada. Mientras en Putnam la analogía es teóricamente limitada –puesto que, como he indicado, el comportamiento de diferentes sistemas físicos puede modelarse mediante la postulación de organizaciones funcionales–, la interpretación de Fodor le da fuerza teórica: la mayoría de los sistemas físicos conocidos no manipulan estructuras combinacionales de una complejidad arbitraria de acuerdo con reglas recursivas (Piccinini 2010). En la medida en que tanto los cerebros como los mecanismos computacionales programables tienen esta capacidad, resultaría justificado pensar que la explicación de su comportamiento debe apelar al mismo tipo de principios y estrategias (Fodor 1975). La mente es, de acuerdo con esta forma de computacionalismo funcionalista, como devino popular afirmar, "el software del cerebro" (Block 1995).

3.6.2. El computacionalismo funcionalista: una hipótesis no mecanicista

En contraste con el computacionalismo de McCulloch y Pitt, y en contravía de la tradición teórica en la que surgió la analogía entre la mente y la computación, el computacionalismo funcionalista no es una hipótesis mecanicista. Los "programas" que conformarían el software del cerebro, al igual que los programas que conforman el software de los computadores digitales, podrían codificarse, almacenarse y ejecutarse a través de soportes físicos distintos. De esta manera, ni los programas ni los eventos de su ejecución podrían identificarse ni reducirse a estados y procesos físicos específicos. Un programa computacional es múltiplemente realizable: sus condiciones de individuación son condiciones de individuación funcionales. Supuesto esto, el uso de los mecanismos abstractos ideados por Turing en modelos de capacidades cognitivas no tiene la función de revelar *cómo* –esto es, mediante qué mecanismos descritos en el nivel físico– el cerebro realiza estas capacidades. McCulloch y Pitt desarrollaron en cambio su modelo como una herramienta para desentrañar los mecanismos neurológicos que subyacen y constituyen las capacidades cognitivas y, en consecuencia, su corrección, como ambos reconocían, dependía del acopio de evidencia neurobiológica. De acuerdo con estos autores, en consecuencia, el estudio de la cognición no podría conceptualizarse como un ámbito de estudio autónomo.

La ciencia cognitiva se desarrolló institucionalmente como un programa de estudio interdisciplinario a partir de los años sesenta sobre la base de la interpretación funcionalista del computacionalismo descrita en la parte final de este capítulo (Gardner 1985). El componente integrador fundamental de este programa era una concepción de la cognición como procesamiento computacional de información y de la mente como el software del cerebro. Alrededor de estos presupuestos, el trabajo realizado en las diferentes disciplinas que conformaban la "nueva ciencia de la mente" convergería pretendidamente en un conjunto unificado de hipótesis y modelos

explicativos de las capacidades cognitivas. En el siguiente capítulo describiré con mayor detalle la forma que tomó este programa de investigación y examinaré las razones por las que el computacionalismo funcionalista resultó en último término un fundamento endeble para una ciencia de la cognición.

CAPÍTULO 4

4.1. Introducción

El surgimiento a mediados del siglo pasado de la iniciativa institucional denominada "ciencia cognitiva" representa uno de los últimos episodios en la historia de los esfuerzos por darle estatus científico al estudio de las causas de la conducta inteligente. El percibido potencial teórico de la analogía entre la mente y los computadores convenció a un grupo de investigadores de que las circunstancias eran propicias para el desarrollo de una "nueva ciencia". Estas pretensiones tomaron forma en Estados Unidos a finales de los años cincuenta. A partir de ese momento empezarían a abrirse programas de posgrado y a desarrollarse proyectos de investigación encausados y justificados bajo el nombre de la nueva ciencia (Bechtel et. al 1998; Núñez et. al. 2019). La idea fundamental de los investigadores que se agruparon bajo esta iniciativa institucional era que la cognición podía entenderse en términos de estructuras en la mente y de procedimientos computacionales que operan sobre estas estructuras (Gardner 1985).

El marco teórico de la ciencia cognitiva "clásica" es multidisciplinar y entraña una división estricta del trabajo. De acuerdo con la bifurcación funcionalista del trabajo científico reseñada en el capítulo anterior, la nueva ciencia se dividiría en dos empresas idealmente complementarias pero autónomas. Por un lado, estaría la psicología, la inteligencia artificial, la lingüística y la filosofía. Estas disciplinas conformarían el estudio funcional de la cognición y su objetivo sería la explicación cognitiva de la conducta inteligente a través de la formulación de modelos funcionales entendidos como programas computacionales. Por otro lado, estarían las neurociencias, que se ocuparían del nivel neuronal, mecánico o de "implementación". El objetivo de estas disciplinas sería determinar cómo procesos y estados en el cerebro individuados neurofisiológica y no cognitivamente "realizan" los modelos postulados por las explicaciones cognitivas. Las explicaciones potencialmente desarrolladas en ambos niveles serían tipos de explicaciones distintas y *autónomas*. Esta división del trabajo no dejaba lugar a un estudio *cognitivo* de los mecanismos neuronales del tipo desarrollado por McCulloch y Pitt. De acuerdo con este esquema de trabajo, la idea de una neurociencia cognitiva constituía una especie de oxímoron (Boone y Piccinini 2015).

Las dos disciplinas centrales de la ciencia cognitiva clásica eran la psicología y la inteligencia artificial. Bajo el entendido de que la mente es el software del cerebro y de que la noción de computación designa un fenómeno semántico (una tesis que discutiré en la sección 5 de este capítulo), la primera pretendía explicar todas las capacidades cognitivas en términos de ejecuciones de programas computacionales, mientras que la segunda buscaba emplear programas de este tipo para diseñar y

construir agentes artificiales inteligentes que exhibieran todas las capacidades típicas de la mente humana (Bringsjord y Govindarajulu 2018; *Proudfoot y Copeland 2012).*

La inclusión de la inteligencia artificial como parte fundamental de la ciencia cognitiva pone de manifiesto la modificación sustancial que el funcionalismo efectuó en el computacionalismo. En su interpretación funcionalista el computacionalismo pasó de ser una herramienta para investigar la manera en que la cognición podía ocurrir en el cerebro a una plataforma para obviar esta investigación y estudiar la cognición como un fenómeno abstracto. En tanto hipótesis mecanicista, el computacionalismo representa una herramienta para integrar el estudio de la cognición en la empresa de la ciencia natural. La ciencia de la cognición fundada en esta hipótesis sería, en línea con la historia reseñada en los dos primeros capítulos, el estudio mecanicista de un fenómeno biológico y en consecuencia un segmento de la biología. En su encarnación funcionalista, en cambio, el computacionalismo es una herramienta para aislar el estudio de la cognición de las ciencias naturales. La ciencia cognitiva clásica es una ciencia funcional, metodológica y metafísicamente independiente del conjunto de las ciencias naturales (Block 1995).

La hegemonía y el entusiasmo generado por este programa de investigación funcionalista duraron relativamente poco. Aunque desde el comienzo hubo voces disidentes, a partir de los años ochenta empezó a hablarse de la disolución del programa clásico y emergieron programas alternativos que cuestionaban los fundamentos sobre los que se había construido la ciencia cognitiva en los años cincuenta y sesenta (Bechtel et. al 1998, Marrafa y Paternoster 2012). Algunos cuestionaron el computacionalismo fundacional de esta empresa y propusieron marcos teóricos diferentes para estudiar la cognición. Una de las alternativas que empezó a contemplarse fue el empleo de formalismos matemáticos diferentes, como el de la teoría de sistemas dinámicos (Van Gelder 1995; Faries y Chemero 2019). El progreso en el conocimiento del cerebro y el sistema nervioso, sin embargo, llevó al resurgimiento del estudio de la cognición desde una perspectiva neuronal-computacional del tipo desarrollada por McCulloch y Pitt. El trabajo efectuado desde esta perspectiva a partir de los años ochenta –agrupado bajo el nombre de "conexionismo"– cuestionaba el esquema de división del trabajo consagrado por la ciencia cognitiva clásica y fue uno de los elementos protagónicos en las discusiones fundacionales que empezaron a tener lugar en el momento (Akagi 2017). En este capítulo trataré de mostrar por qué la disolución de la ciencia cognitiva clásica no representó en evento sorprendente sino un resultado esperable del carácter endeble de los presupuestos en los que estaba fundada.

En la sección 2 expongo el tipo de descripciones computacionales de las capacidades cognitivas sancionadas por el funcionalismo de la ciencia cognitiva clásica: los llamados "modelos cognitivos clásicos". Al final de la sección planteo las

dificultades a las que se enfrenta la justificación del carácter *explicativo* de estos modelos. En el resto del capítulo me ocupo de los intentos de resolver estas dificultades dentro del marco de referencia del funcionalismo. En la sección 3 discuto el rol que el uso de la tesis Church-Turing tuvo en este sentido en la literatura funcionalista. En la sección 4 describo la manera en que el empleo equívoco del concepto de *mecanismo* caracteriza la interpretación funcionalista del computacionalismo y la comprensión de la manera en que los modelos clásicos pueden ser explicativos de la cognición. En la sección 5 presento y discuto críticamente la concepción semántica de la computación y la idea según la cual el computacionalismo conlleva un tipo de representacionalismo sobre la mente. En la sección 6 describo el conexionismo surgido en los años ochenta como un tipo de computacionalismo y como un programa de investigación de la cognición que, a pesar de las pretensiones de sus promotores, supuso en último término una variedad del funcionalismo. En la sección 7, por último, sostengo que el funcionalismo conduce a un tipo de nihilismo computacional que socava el carácter *explicativo* del uso de modelos computacionales en la investigación de la cognición.

4.2. Modelos cognitivos y explicación computacional en la ciencia cognitiva

Una manera de delimitar el ámbito de la ciencia cognitiva clásica es a través de la explicitación de las características distintivas de los modelos que sus presupuestos teóricos sancionan. Los modelos cognitivos "clásicos" son típicamente modelos que caracterizan en términos computacionales la manera en que una capacidad cognitiva específica se produce o ejerce. El objetivo de estos modelos es explicar estas capacidades a través de la descripción de procesos computacionales que pretendidamente subyacen al ejercicio de las mismas. Cada capacidad cognitiva se describe como la producción mediante un proceso computacional de un tipo definido de *outputs* a partir de un tipo definido de *inputs*. Un modelo de este tipo se concibe de manera más exacta como la especificación de una secuencia de instrucciones de computación; esto es, como un programa computacional. Los modelos cognitivos clásicos representan entonces las capacidades cognitivas como tipos de manipulaciones algorítmicamente especificables de vehículos computacionales denominados, en consonancia con la teoría matemática de la computación, como "símbolos" (Samuels 2019).

La idea comúnmente presupuesta en la formulación de los modelos clásicos es que los símbolos son vehículos de información que conforman sistemas representacionales similares a un lenguaje lógico. Estos sistemas están definidos en términos de una serie de reglas sintácticas que definen de manera recursiva cuáles combinaciones de símbolos constituyen expresiones complejas correctas y en términos de una serie de reglas semánticas que asignan significados a las expresiones

así formadas. Dado que los símbolos son los vehículos básicos de los procesos cognitivos, se dice que el medio de la cognición es un "lenguaje" del pensamiento (Fodor 1975; Piccinini 2012). Aquello en lo que consiste la cognición, la característica esencial de todas las "operaciones" de la mente, es el procesamiento algorítmico de estos vehículos de información.

El procesamiento en cuestión es *algorítmico* en la medida en que la manipulación de los símbolos y de las ristras de símbolos obedece o responde solo a las propiedades formales y no a las propiedades semánticas de estos. Dada una serie de símbolos, un algoritmo genera de manera sistemática a través de una secuencia de pasos estrictamente definidos de manipulación de estos símbolos otra serie de símbolos. Esta manipulación algorítmica, sin embargo, "preserva" o "respeta" las propiedades semánticas de los símbolos, de manera que un evento de procesamiento de este tipo puede interpretarse en su conjunto como el "mapeo" de un conjunto de vehículos de información en otro conjunto de vehículos de información. Todo aquello de lo que la mente es capaz puede describirse y explicarse de acuerdo con esta concepción clásica de la cognición como el procesamiento algorítmico de una serie de símbolos (Fodor 1975; Boden 2006).

El lenguaje, por ejemplo, constituye un conjunto de capacidades cognitivas, cada una de las cuales es susceptible de descripción en términos de un programa computacional. La capacidad de recibir estímulos físicos bajo la forma de patrones de intensidad de ondas lumínicas o sonoras y convertirlos en mensajes lingüísticos con contenidos proposicionales es una capacidad de la mente que habría de poder describirse como el procesamiento secuencial y sistemático de ciertos vehículos de información. Al proponer un modelo de estas secuencias específicas de procesamiento, el teórico cognitivo clásico delimita una de las capacidades distintivas de la mente y busca explicitar el conjunto de instrucciones computacionales en virtud de las cuales esta capacidad se ejerce. No solo el lenguaje, sino la memoria, el razonamiento práctico, la atención y cualquier capacidad cognitiva podría describirse y explicarse mediante la formulación de modelos computacionales de este tipo.

4.2.1. Arquitecturas cognitivas, inteligencia artificial y autonomía explicativa

Idealmente, los modelos computacionales de cada una de las capacidades de la mente habrían de poder integrarse en un único modelo unificado. El cumplimiento de este propósito es el objetivo de los llamados modelos de "arquitecturas cognitivas" (Newell 1990; Bermúdez 2014). Mientras que los modelos "clásicos" típicos toman como sus *objetos* aspectos restringidos de la cognición, las arquitecturas cognitivas propuestas dentro del paradigma clásico buscan proporcionar especificaciones detalladas y comprensivas de cómo la cognición opera a través de un amplio rango de tareas y dominios. Es en este sentido que entrañan modelos "unificados". El objetivo central detrás de la articulación de este tipo de modelos es la determinación del

conjunto nuclear de estructuras, operaciones computacionales y recursos básicos de los que depende la cognición (Samuels 2019).

La analogía que motiva la construcción de este tipo de modelos es la analogía entre la mente y los computadores digitales. Un computador de este tipo tiene típicamente cuatro tipo de unidades funcionales básicas: dispositivos de entrada, unidades de memoria, unidades de procesamiento y dispositivos de salida (Piccinini 2008a). El adecuado ensamblaje de estas unidades conforma la *arquitectura* de un computador. Esta arquitectura junto con los programas –el software– que un artefacto de este tipo ejecuta son los elementos de los que depende la comprensión de cada una de sus capacidades. La arquitectura define la manera en la que el sistema recibe estímulos, los transforma en *inputs* adecuados para su manipulación mecánica, los procesa de acuerdo con una serie de instrucciones y genera *outputs* adecuados. De la misma manera, se presumía, si pudieran determinarse las unidades funcionales básicas de la mente, este conocimiento podría conjugarse con el conocimiento aportado por los modelos de capacidades cognitivas circunscritas para articular un modelo unificado de todas las capacidades de la mente.

En el trabajo efectuado sobre las arquitecturas cognitivas se ponen de manifiesto los aspectos distintivos de la ciencia cognitiva clásica. Si la cognición es procesamiento computacional de información, toda instancia de conducta inteligente debe poder explicarse como el resultado de procesos este tipo. Por otra parte, si estos procesos, como enfatizaban los funcionalistas, son *múltiplemente realizables*, entonces debe ser por principio posible construir un artefacto artificial que "implemente" físicamente los procesos computacionales abstractos relevantes y de esta manera *replique* o simule todas las posibles capacidades de la mente humana. Un artefacto de este tipo es un "agente artificial inteligente". El diseño y construcción de agentes artificiales inteligentes, como indiqué en la sección introductoria, es el objetivo central de una de las disciplinas nucleares de la ciencia cognitiva clásica: la inteligencia artificial (Bringsjord y Govindarajulu 2018). Al formular una arquitectura cognitiva se daba por sentado que la misma constituía un punto de referencia para el diseño de un agente artificial inteligente, cuando no una versión del diseño mismo. Una arquitectura cognitiva puede entenderse en este sentido como la especificación de un sistema abstracto organizado de subsistemas que realizan diferentes tareas de procesamiento. En su conjunto, el sistema toma como *inputs* estímulos ambientales filtrados como representaciones o vehículos adecuados de computación y genera como *outputs* modificaciones en estados internos del sistema o eventos de conducta (Bermúdez 2014).

Los detalles neurobiológicos que subyacen a las capacidades cognitivas en el caso de organismos como los seres humanos son irrelevantes para efectos de diseñar un agente artificial, así como, en general, para estudiar la cognición desde el punto de vista de la inteligencia artificial. Es en este sentido que Daniel Dennett sostiene que el

programa teórico de la inteligencia artificial puede concebirse como "la investigación más fundamental sobre la posibilidad de la inteligencia o la cognición" (Dennett 1981, p. 119). Para Dennett, este proyecto representa el intento de explicar la cognición no a través del estudio de los mecanismos neurobiológicos que subyacen a la misma en los organismos vivos –en particular, los seres humanos– sino a través del diseño e implementación de algoritmos abstractos que capturen las tareas que efectúa la mente y que constituyen sus capacidades (Dennett, 1981; Bringsjord y Govindarajulu 2018).

Esta perspectiva sobre la explicación de la cognición conlleva, como se indicó en el capítulo anterior, una aproximación descendente (*top down*) a su estudio científico. La determinación de los algoritmos o programas computacionales que delimitan el ámbito de la cognición se hace al margen de cualquier consideración relativa a la manera en que funciona el cerebro. Una vez que estos algoritmos se han diseñado y se han puesto a prueba en la construcción de artefactos artificiales que simulen convincentemente las capacidades humanas, podrían entonces servir de guía para estudiar y rastrear los mecanismos mediante los cuales el cerebro realiza las mismas capacidades. Esta segunda tarea, sin embargo, no haría parte de la ciencia de la cognición sino de las disciplinas concernidas con los sistemas que "implementan" la cognición –en el caso de los seres humanos, disciplinas biológicas (Block 1995). En la división del trabajo consagrada por el programa de la ciencia cognitiva clásica, la investigación de la cognición se agota en la indagación de las maneras en que las capacidades cognitivas pueden simularse computacionalmente. La inteligencia artificial y el proyecto global de la ciencia cognitiva del que hace parte conciben así la mente como un tipo de sistema computacional abstracto y el estudio de sus capacidades como una empresa *autónoma* con respecto a la biología y al resto de las ciencias naturales.

4.2.3. El problema de la justificación del computacionalismo funcionalista

La tesis central del computacionalismo entendido en clave funcionalista es que la mente es la "organización funcional" del cerebro –o de un sistema físico capaz de "realizar" las capacidades cognitivas. La expresión "mente" no es entendida aquí como el nombre de algún tipo de entidad sino como una expresión mediante la cual hacer referencia de manera abreviada a un conjunto de capacidades de procesamiento algorítmico de información. El "soporte" físico, el sistema mediante cuya estructura se "realizan" estas capacidades en el caso de organismos como los seres humanos es el cerebro. En este sentido, el cerebro es concebido como un tipo de sistema computacional programable y la mente es su "software": el conjunto de programas almacenados en el mismo que definen y explican su comportamiento. El computacionalismo así entendido es el presupuesto teórico fundamental de la ciencia cognitiva clásica (Bechtel et al 1998).

Una consecuencia crucial de lo dicho antes es la siguiente: en la medida en que la ciencia cognitiva pretende ser una disciplina *explicativa* de la cognición, su viabilidad teórica depende de que las capacidades cognitivas de organismos inteligentes como los seres humanos sean susceptibles de explicación a través del mismo tipo de factores que explican las capacidades exhibidas por sistemas computacionales concretos; en particular, por artefactos computacionales como los computadores programables. Que las capacidades de estos artefactos pueden explicarse a través de modelos en los que se recurre a los formalismos abstractos desarrollados por la teoría matemática de la computación es algo que puede darse por sentado en virtud del hecho de que conocemos cómo están compuestos estos artefactos (Piccinini 2008a). Sabemos, en particular, que estos artefactos fueron diseñados y construidos con base en esos formalismos y que los procesos físicos que conforman su funcionamiento son procesos que implementan los procesos abstractos descritos por la teoría. Que un computador es un sistema físico complejo que efectúa computaciones y que sus componentes son a su vez mecanismos que realizan computaciones es algo sobre lo que no cabe discutir. Que el cerebro sea un sistema físico similar, sin embargo, es una tesis que requiere de justificación, como resultaba claro en la tradición mecanicista en la que surgió inicialmente el computacionalismo. A diferencia de lo que ocurre con los mecanismos de computación física diseñados por los seres humanos, no está tan claro cómo funciona el cerebro; esto es, en virtud de qué tipo de procesos o mecanismos realiza sus tareas características. Dada esta ignorancia no es posible tampoco determinar con certeza qué tipo de modelos representen mejor esos mecanismos y permitan en consecuencia explicar las capacidades del cerebro.

El computacionalismo constituye una apuesta teórica en este sentido. De acuerdo con esta apuesta, la relación entre las capacidades de la mente y los factores físicos que las explican es el mismo *tipo* de relación (o bien una relación homologable) a la existente entre las capacidades de un artefacto computacional concreto y los factores que explican sus capacidades. Pero, ¿qué tipo de evidencia cuenta como evidencia favorable a esta apuesta?

El modelo computacional de McCulloch y Pitt representaba una hipótesis mecanística, en el sentido de que lo que pretendían representar los mecanismos abstractos descritos en el modelo, de manera correcta o incorrecta, eran los mecanismos físicos concretos constitutivos del funcionamiento del cerebro. De acuerdo con esta versión del computacionalismo, por tanto, si el cerebro es o no un sistema computacional es algo que puede determinarse de manera directa estudiando los procesos mediante los que realiza sus tareas y examinando si estos son procesos de computación. El funcionalismo bloquea sin embargo esta interpretación de la apuesta computacionalista. Los procesos descritos en los modelos "clásicos" no son representaciones de procesos y mecanismos en el nivel físico de descripción. En el

marco del funcionalismo, si el cerebro es un sistema computacional, en el sentido de que la *explicación* de sus capacidades puede efectuarse a través de modelos computacionales, es algo que debe establecerse por medios indirectos. La exposición de estos diferentes medios indirectos será el objeto de las siguientes sesiones.

4.3. **La tesis Church-Turing y el computacionalismo**

La justificación del supuesto según el cual el cerebro –o cualquier sistema cognitivo físico– es un sistema computacional estuvo vinculada de manera más o menos directa y más o menos explícita en la literatura funcionalista con una interpretación de la tesis Church-Turing mencionada en el primer capítulo. Esta tesis afirma que el conjunto de funciones matemáticas computables mediante procedimientos efectivos o algorítmicos es idéntico o tiene la misma extensión que el conjunto de funciones computables mediante una máquina de Turing. Planteado en otros términos, la tesis Church-Turing sostiene que el conjunto de todas las "tareas de procesamiento algorítmico" posible está delimitado por el conjunto de tareas de procesamiento algorítmico susceptibles de ejecución por una máquina de Turing (Copeland, 2017). Así, *si* todas las capacidades de la mente se efectúan a través de la ejecución de tareas de este tipo, *si* la mente consiste en el procesamiento algorítmico de vehículos de computación, entonces dos conclusiones parecen seguirse: (a) un artefacto físico con un poder computacional equivalente al de una máquina de Turing (adecuadamente programado) debe poder replicar todas las capacidades cognitivas; (b) todas las capacidades cognitivas deben poder describirse y explicarse en términos de las propiedades computacionales de una máquina universal de Turing. A pesar de su importancia decisiva para la fundamentación de las pretensiones teóricas de la inteligencia artificial y de la ciencia cognitiva clásica en general, el uso de la tesis Church-Turing en este contexto solo empezó a ser objeto de un escrutinio crítico detallado recientemente (Copeland 2000).

En sentido estricto, la tesis Church-Turing es una afirmación matemática relativa a qué conjunto de funciones matemáticas pueden resolverse mediante algoritmos, cuya corrección se acepta con base en consideraciones que cabría denominar "a priori". En particular, la tesis se ha aceptado como verdadera en la comunidad matemática sobre la base de que no se han encontrado contraejemplos, de que diferentes intentos de formalizar la noción de *computación* han resultado en formalismos computacionalmente equivalentes y de que la idea de una máquina de Turing captura bien la noción intuitiva de computación empleada en el trabajo matemático (Piccinini 2007b; Sieg 2009). El uso de la tesis en el ámbito del estudio de sistemas y procesos físicos se funda en una extensión de la misma que conviene apreciar con claridad.

4.3.1. La tesis Church-Turing física

Dado que los mecanismos abstractos ideados por Turing inspiraron la construcción de artefactos computacionales físicos que los "implementaban", se dio por sentado que la tesis Church-Turing podía *interpretarse* en términos físicos. Lo que podría denominarse la "tesis Church-Turing física" se refiere no a funciones matemáticas abstractas sino a "funciones" cuyos valores son generados por procesos constitutivos de sistemas físicos. De acuerdo con esta tesis, todas las funciones *físicamente* computables son computables por una máquina de Turing. Si una función es físicamente computable, esto es, entonces forzosamente cae bajo el conjunto de funciones computables por una máquina de Turing. Dependiendo de qué se entienda como la determinación de un valor por parte de un sistema físico pueden discernirse al menos dos versiones de la tesis. De acuerdo con una versión radical de la tesis, cualquier proceso físico –cualquier cosa que un sistema físico pueda "hacer"– puede describirse en términos de la generación de *outputs* a partir de *inputs* y por tanto puede modelarse como una función matemática. Cualquier proceso físico descrito en estos términos, de acuerdo con esta versión radical, es computable por una máquina de Turing. La versión modesta de la tesis concierne no a cualquier proceso realizable por medios físicos sino a los procesos constitutivos de ciertos sistemas especiales (Piccinini 2007b).

4.3.1.1. La versión radical de la tesis: metafísica y simulación computacional de sistemas físicos

La versión radical de la tesis Church-Turing física sostiene que todo proceso causal *es* un proceso computacional y todo sistema físico es por tanto un sistema computacional. Un corolario inmediato de esto es que todos los procesos neuronales –un subconjunto de los procesos causales– son procesos computacionales y que el funcionamiento del cerebro puede explicarse mediante la descripción de mecanismos computacionales como las máquinas de Turing. El resultado es un tipo de *pancomputacionalismo*. Este resultado puede leerse de dos maneras: una que cabría llamar "metafísica" y otra "metodológica".

De acuerdo con la lectura más literal, lo que afirman los partidarios de esta versión de la tesis es que la realidad física está en último término constituida metafísicamente por mecanismos computacionales deterministas (Piccinini 2011). Una de las condiciones definitorias de una máquina de Turing es la llamada "condición de determinación". De acuerdo con esta condición, el comportamiento de una máquina *m* en cualquier momento dado *t* está completamente determinado por la configuración o el estado de máquina en el que *m* se encuentra y el símbolo que está escaneando en *t*. La transición entre una acción de *m* y la siguiente está determinada por una única lista de instrucciones que establecen, junto con los *inputs* de *m*, una secuencia definida de acciones (De Mol 2018). Aunque el formalismo de

Turing admite mecanismos cuyo comportamiento está regido por procesos probabilísticos, de manera que lo que una máquina hace en un tiempo *t*+1 está solo parcialmente determinado por el estado en el que estaba y la acción que ejecutó en un tiempo *t*, ninguna máquina de Turing puede comportarse de manera genuinamente aleatoria. Lo que esto significa es que una secuencia genuinamente aleatoria no es computable por una máquina de Turing (Piccinini 2011). En consecuencia, si todo proceso físico es computable por una máquina de Turing, no existen procesos aleatorios ni en general sistemas no-deterministas en la naturaleza. Dado que la existencia de procesos de este tipo es una cuestión abierta, la corrección de la versión radical de la tesis Church-Turing física es también una cuestión abierta.

Ahora bien, aun suponiendo que el universo estuviera constituido solo por sistemas deterministas, la tesis no carece de problemas. El aparato matemático empleado de manera más exitosa para describir sistemas físicos en general usa funciones matemáticas de variables reales *continuas*, cuyo dominio y rango incluye conjuntos no numerables de valores. Las funciones computables por una máquina de Turing, en cambio, son funciones de variables *discretas*, cuyo dominio y rango incluyen conjuntos numerables de valores (Sieg 2009). En virtud de esto, las funciones de variables reales empleadas en la modelación más exitosa empíricamente del comportamiento de sistemas físicos no pueden ser correlacionadas o mapeadas directamente en funciones computables por máquinas de Turing (Piccinini 2007b). En consecuencia, no existen tampoco buenas razones para creer que los procesos que constituyen el funcionamiento de los sistemas físicos sean procesos computacionales deterministas y puedan simularse mediante una máquina de Turing.

La versión radical de la tesis Church-Turing física, sin embargo, puede leerse en otra clave. De acuerdo con una interpretación "metodológica", lo que esta versión de la tesis afirma es que todos los sistemas físicos pueden modelarse de manera empíricamente provechosa como mecanismos computacionales de *inputs* y *outputs*. La idea en juego es que la evolución dinámica de un sistema físico cualquiera puede estudiarse empleando formalismos computacionales que permitan predecir el estado del sistema en intervalos de tiempo definidos.

La manera más simple de determinar la evolución dinámica de un sistema físico *S* es detallando representaciones de los estados de *S* en momentos subsecuentes de tiempo obtenidas midiendo variables relevantes de *S* en esos momentos. Mediante una determinación de este tipo podría construirse una tabla de consulta (*lookup table*) de la evolución dinámica de *S*. Para establecer el estado de un sistema en un momento *t* bastaría entonces con recuperar un valor de una tabla de este tipo. De manera más común en las ciencias físicas, sin embargo, la evolución dinámica de un sistema se detalla mediante una descripción matemática que especifica cómo las variables de *S* varían como una función del estado de *S* en diferentes momentos de tiempo; típicamente, a través de un sistema de ecuaciones diferenciales. Dado un

sistema de ecuaciones de este tipo, una "solución analítica" es una formula tal que, dada cualquier condición inicial del sistema y cualquier tiempo subsecuente *t*, la fórmula permite establecer (esto es, computar) el estado del sistema en *t*. Si un sistema de ecuaciones tiene una solución analítica y si la solución es conocida, puede computarse de manera efectiva un valor que represente el estado de *S* en cualquier momento *t* (Piccinini 2007a).

La mayoría de sistemas de ecuaciones diferenciales, sin embargo, como es sabido, no pueden resolverse analíticamente. Es en este punto en que el empleo de procedimientos computacionales en la modelación de sistemas físicos resulta más útil. Los llamados "métodos numéricos" son procedimientos algorítmicos mediante los cuales se obtienen soluciones *aproximadas* de sistemas de ecuaciones. Un procedimiento de este tipo consiste en una lista de instrucciones que detalla secuencias de operaciones aritméticas y lógicas que manipulan valores discretos de entrada y generan valores discretos de salida. Un método numérico puede concebirse en este sentido como un programa de computación simple. Con base en estos métodos se han construido programas de computación complejos empleados para modelar diferentes tipos de sistemas físicos (Piccinini 2007b).

Un *modelo computacional* en este sentido es un programa que explota métodos numéricos adecuados para computar representaciones de estados subsecuentes de un sistema físico sobre la base de ecuaciones que representan su evolución dinámica y de datos que representan condiciones iniciales del mismo. Un modelo computacional, dicho en otros términos, es en este sentido un formalismo empleado para estudiar y describir de manera aproximada la evolución dinámica de sistemas físicos (Illari y Russo 2014).

Al emplear un modelo de este tipo para estos efectos el resultado es una "simulación computacional" de un sistema de interés científico. Las simulaciones computacionales han sido usadas con éxito en una gran variedad de disciplinas, como en la física, la biología y en diferentes ciencias sociales (Winsberg 2015). Estas simulaciones son de cierta manera sustitutos de experimentos cuya realización es imposible o técnicamente muy difícil. Su objetivo es *imitar* computacionalmente el comportamiento del sistema en diferentes circunstancias, de manera que los datos aportados por este proceso arrojen información valiosa sobre las propiedades dinámicas del sistema simulado. *Si* los parámetros de la simulación son representaciones físicamente realistas del sistema real, los resultados de la simulación *pueden* ser iluminadores sobre el comportamiento del sistema (Illari y Russo 2014). La popularidad de esta estrategia de modelación en diferentes sectores de la empresa científica puede ser la razón que motiva la afirmación de la versión radical de la tesis Church-Turing física. Esta versión de la tesis, de hecho, se presenta a menudo como la afirmación según la cual *todo* puede ser "simulado" por un mecanismo computacional. De manera más precisa, la idea en juego es que si un

sistema físico *S* puede describirse matemáticamente mediante un sistema de ecuaciones, entonces su comportamiento –su evolución dinámica– puede simularse mediante un modelo computacional.

Como una consecuencia crucial de lo anterior, las máquinas ideadas por Turing terminan entendiéndose no como mecanismos especiales para efectuar ciertas tareas circunscritas sino como un modelo general de *todo* lo que pueda describirse con precisión matemática. Putnam, por ejemplo, sostuvo en uno de los artículos fundacionales del funcionalismo sobre la mente que "todo es un autómata probabilístico bajo alguna descripción" (Putnam 1975b, p. 435). Un "autómata probabilístico" es un tipo de máquina de Turing. Al hacer esta afirmación, Putnam mantenía, en línea con lo expuesto antes, que todo puede ser modelado como un mecanismo computacional en cierto grado de aproximación.

Es fundamental apreciar lo que conlleva que las descripciones aportadas por un modelo computacional empleado en una simulación sean solo aproximadas y el tipo de usos que reciben estas descripciones en la práctica científica.

Si un modelo es más o menos aproximado depende de factores relativos a qué tanto se conocen las propiedades dinámicas del sistema modelado, qué tanto de ese conocimiento puede integrarse en el modelo sin comprometer su poder computacional, qué tan precisos sean los métodos numéricos que incluye, entre otros. El grado de aproximación admisible depende de factores pragmáticos, relativos a los objetivos de los investigadores que construyen el modelo. El punto crucial es que un modelo computacional puede ser un "buen modelo" aun cuando existan razones concluyentes para pensar que no es un modelo muy aproximado, esto es, que no "refleja" de manera precisa la naturaleza de un sistema. Un buen modelo es un modelo que le permite a un investigador o un grupo de investigadores satisfacer sus objetivos al formular el modelo. Un mismo modelo, por otra parte, puede ser suficientemente aproximado para ciertos propósitos pero no para otros (Winsberg 2015). Un objetivo típico en este caso es la *predicción*. En este sentido, un modelo puede generar predicciones útiles del comportamiento de un sistema aun cuando se sepa que no refleja de manera precisa los factores que *explican* este comportamiento. Lo que esto significa es que del éxito de un modelo computacional para simular un sistema físico no implica de manera directa nada sobre los procesos o mecanismos físicos que explican las regularidades o comportamientos del sistema. En general, la importancia teórica de una simulación computacional no reside en que entrañe alguna tesis sobre la naturaleza de los sistemas físicos y la explicación de sus regularidades (Proudfoot y Copeland 2012).

En consecuencia, aun cuando pueda decirse que todos los procesos naturales son procesos computacionales en el sentido de que pueden simularse con cierto grado de aproximación mediante un programa computacional, no cabe decir que todos los procesos naturales son computacionales en el sentido de que puedan *explicarse* en

términos de procesos de computación física. Pero esto último era justamente lo que afirmaba la versión radical de la tesis Church-Turing física.

En resumen, ni bajo una lectura metafísica ni bajo una lectura metodológica existen buenas razones para suscribir la versión radical de la tesis Church-Turing física y en consecuencia tampoco para pensar que esta versión de la tesis justifica la creencia en que mecanismos computacionales abstractos revelan la naturaleza de la cognición.

4.3.1.2. La versión modesta de la tesis, la falacia Church-Turing y la noción de *mecanismo*

La versión modesta de la tesis Church-Turing física hace referencia solo a un subconjunto de los sistemas físicos. Mientras que algunos sistemas (como los computadores electrónicos y, tal vez, los cerebros) son tales que su funcionamiento consiste en la efectuación de computaciones, otros sistemas (como los huracanes, los estómagos o los motores) no exhiben a primera vista esta característica. Solo los procesos constitutivos del comportamiento del primer tipo de sistemas son de interés en este caso. Esta versión de la tesis hace referencia por tanto no al universo en su conjunto sino solo a ciertos sistemas físicos pequeños confinados dentro de regiones espacio-temporales relativamente pequeñas. Lo que afirma puede plantearse en términos condicionales: si un sistema físico efectúa computaciones, entonces todas las posibles funciones que computa son funciones computables por una máquina de Turing (Piccinini 2007b).

La verdad de esta tesis depende de la imposibilidad física de la existencia de sistemas que computen funciones no computables por una máquina de Turing; esto es, que determinen por medios físicos valores de funciones no computables por una máquina de Turing. Un sistema de este tipo es lo que se ha denominado un "hipercomputador" (Proudfoot y Copeland 2012). Un hipercomputador es un sistema físico posible que realiza tareas de procesamiento algorítmico que exceden las capacidades de una máquina universal de Turing. Un ejemplo de un hipercomputador sería una máquina de Turing "de aceleración infinita". Una máquina de este tipo efectúa cada operación computacional en la mitad del tiempo empleado en la efectuación de la operación previa y en consecuencia completa un número infinito de operaciones (una "súper tarea de procesamiento") solo en el doble del tiempo que le toma completar la primera de esas tareas. En virtud de esta súper capacidad, una máquina así puede computar funciones no computables por una máquina de Turing usual. Todo lo que la versión modesta de la tesis Church-Turing física afirma es que estas máquinas son físicamente imposibles. Aun cuando la posibilidad de estas máquinas ha sido objeto de discusión en las últimas décadas y la cuestión permanece hasta cierto punto abierta, existen buenas razones para pensar que por ahora los hipercomputadores son solo posibilidades nocionales o teóricas y que, en

consecuencia, la versión modesta de la tesis Church-Turing física es correcta (Piccinini 2011). Para efectos de las cuestiones discutidas en este documento, sin embargo, esta conclusión es de limitada relevancia.

Si la versión radical de la tesis Church-Turing física fuera correcta, esto tendría consecuencias inmediatas con respecto a la naturaleza de la cognición y su estudio. Si todos los procesos naturales son computacionales entonces los procesos neuronales y en general los procesos cognitivos serían computacionales y podrían explicarse apelando al mismo tipo de factores que explican los procesos efectuados por una máquina de Turing. La versión modesta de la tesis establece una conclusión solo con respecto a cierto tipo de sistemas: sistemas tales que su comportamiento consiste en efectuar computaciones. De la verdad de esta versión de la tesis no se sigue nada con respecto a si el cerebro es un sistema computacional o si los procesos cognitivos son computacionales. Suponer lo contrario es incurrir en lo que Jack Copeland ha llamado la "falacia Church-Turing" (Copeland 2000). Todo lo que puede concluirse de la tesis es que *si* el cerebro es un sistema tal que su comportamiento consiste en efectuar computaciones, entonces las funciones que computa son computables por una máquina de Turing. La verdad del antecedente de este condicional debe establecerse por medios independientes. Así, en la medida en que (una extensión física de) la tesis Church-Turing pretendía emplearse como una estrategia para establecer que el cerebro debía ser un sistema computacional, puede concluirse que la estrategia en cuestión es infundada. O bien la estrategia confunde descripción con explicación o bien incurre en la falacia mencionada.

Un ejemplo diciente de la comisión de esta falacia puede reconocerse en una influyente defensa del computacionalismo funcionalista articulada por Daniel Dennett. De acuerdo con Dennett, la corrección de la tesis Church-Turing conlleva una restricción metodológica inevitable con respecto a las teorías cognitivas en el sentido de que estas teorías *deben* formularse como algorítmicos o procedimientos efectivos. La consecuencia de no ceñirse a esta restricción sería pretendidamente que la cognición no es susceptible de tratamiento científico. El razonamiento de Dennett es el siguiente. Cualquier teoría que postule operaciones no efectivas –no computables por una máquina de Turing– para explicar una capacidad cognitiva recurre a la postulación de mecanismos no explicables. Un mecanismo de este tipo es, en términos de Dennett, un "homúnculo" no justificado (*undischarged homunculus*); esto es, un proceso inteligente no explicado. Si un proceso no explicable se postula como el explanans de un proceso diferente, entonces la explicación en cuestión carece en último término de sustento. El primer proceso podría tratar de explicarse mediante la postulación de otro "homúnculo", con el resultado de producir un potencial regreso al infinito de homúnculos justificados por otros homúnculos no justificados. La única manera de escapar a este regreso vicioso, de acuerdo con Dennett, es recurrir en la explicación de una capacidad cognitiva únicamente a procedimientos efectivos:

procesos mecánicos modelables como el procesamiento algorítmico de valores adecuados de salida a partir de valores adecuados de entrada (Dennett, 1981).

El presupuesto en el que se funda el razonamiento de Dennett es la idea según la cual las máquinas de Turing constituyen un *modelo general de lo que es un mecanismo explicativo*. Dicho en otros términos, si el comportamiento de un sistema puede explicarse mediante la postulación de un mecanismo, entonces ese mecanismo ha de ser un mecanismo computacional del tipo de una máquina de Turing. Este presupuesto no es nada distinto a la tesis Church-Turing física. Como se vio, sin embargo, ni la versión radical ni la versión moderada de la tesis aportan una justificación de la tesis que interesa a Dennett: que los mecanismos explicativos de la cognición son realizaciones físicas de procedimientos efectivos.

4.4. **Modelos funcionales, mecanismos y computación**

En el razonamiento de Dennett descrito en la sección anterior y en general en la estrategia de justificar el carácter computacional de los procesos cognitivos mediante el uso de la tesis Church-Turing tiene lugar un uso equívoco del concepto de *mecanismo* (Copeland 2000). La misma noción se usa para hacer referencia a dos conceptos distintos. En un sentido que podría denominarse "amplio", un mecanismo es un sistema o una estructura física de componentes organizados que genera o produce un fenómeno. Esta es la concepción de mecanismo articulada por la filosofía mecanicista. En un sentido "restringido", por otra parte, un mecanismo es un proceso que determina un valor de una función computable por una máquina de Turing (o bien, una realización física de este proceso). Este tipo de mecanismos conforma un subconjunto del conjunto de los mecanismos en sentido amplio.

En ambos sentidos, un mecanismo puede ser descrito en un modelo para efectos de explicar un fenómeno o una capacidad dada de un sistema. No todos los mecanismos descritos con este propósito, tal es el punto central, son por tanto mecanismos computacionales. Un mecanismo puede emplearse por ejemplo para explicar cómo el sistema digestivo de algún organismo digiere cierto tipo de alimentos o para explicar cómo un tipo de motor genera energía mecánica. ¿Por qué pensar que lo que hace el cerebro *debe* explicarse mediante la postulación de mecanismos computacionales y no de mecanismos en sentido amplio? La idea general que subyace a la restricción metodológica defendida por Dennett puede aceptarse sin adoptar un compromiso con la tesis de que los mecanismos explicativos de la cognición son mecanismos computacionales. La idea general consiste en que si en la explicación de la capacidad de un sistema cognitivo para realizar una tarea dada se postula como explanans un mecanismo, este último puede representar también un explanandum cuya explicación puede tener lugar en términos mecánicos. Este proceso iterativo puede repetirse hasta que en algún punto se llegue a un mecanismo

físico primitivo. En la medida en que los mecanismos postulados no sean computacionales, sin embargo, la tesis Church-Turing es irrelevante (Piccinini 2007b).

La equivocidad descrita con respecto al concepto de *mecanismo* se extendió en la literatura funcionalista a partir de su inclusión en la formulación original del computacionalismo funcionalista por parte de Fodor, descrita en el capítulo anterior. De acuerdo con Fodor, como se recordará, el componente central de una teoría cognitiva caracteriza un sistema físico a partir de un "análisis funcional". Un análisis funcional identifica los componentes funcionalmente relevantes de un sistema y la manera en que estos componentes interactúan para generar las capacidades del sistema (Fodor 1968a; Cummins 1983). Al desarrollar un análisis de este tipo, el teórico de la cognición propone un modelo funcional de cómo, dadas ciertas circunstancias propiciatorias (*inputs*), un sistema genera cierto tipo de actividades (*outputs*).

Un modelo funcional entendido de esta manera puede pero no tiene que formularse como un modelo computacional; esto es, como un conjunto de instrucciones que detallan transiciones entre estados en virtud de procesos de manipulación algorítmica de ciertos vehículos físicos. A pesar de esta falta de implicación –de la independencia conceptual entre un modelo funcional y un modelo computacional de un sistema– Fodor propuso que el análisis funcional de una capacidad cognitiva consiste en la formulación de un modelo computacional. De acuerdo con esto, un modelo funcional de un sistema con capacidades cognitivas toma la forma de una lista de instrucciones de computación en el sentido de una tabla de una máquina de Turing (Fodor, 1968). De esta manera, como indiqué en el capítulo anterior, Fodor fundió dos conceptos diferentes implicados en la formulación de modelos funcionales de las capacidades de un sistema físico: el de análisis funcional y el de análisis de tareas (*task analysis*). Un análisis de este segundo tipo modela un sistema en términos de la ejecución de tipos finitos de secuencias de operaciones, las cuales pueden analizarse a su vez en términos de subrutinas de ejecución de secuencias de operaciones pero que no necesitan ser atribuidas a componentes del sistema y a sus funciones. Dado que una secuencia así es un tipo de mecanismo computacional abstracto, a partir de Fodor se entendería también que las explicaciones de una capacidad cognitiva serían explicaciones en las que se recurre a la descripción de mecanismos.

Aunque Fodor era consciente de que los modelos computacionales eran solo un subconjunto de los modelos funcionales, pensaba que para efectos de la explicación de las capacidades cognitivas de un sistema la diferencia era irrelevante. Todos los modelos funcionales explicativos de la cognición *habrían* de ser modelos computacionales. La justificación de esta convicción residía en el caso de Fodor en

una particular comprensión de la naturaleza de la computación y en su compromiso con un tipo de representacionalismo sobre la mente.

4.5. La concepción semántica de la computación

Bajo el supuesto de que existiera una propiedad esencial en común entre los procesos cognitivos y los procesos computacionales que no fuera una propiedad característica de los procesos físicos en general, estaría justificado describir el funcionamiento de los sistemas cognitivos recurriendo a las mismas herramientas empleadas para describir el funcionamiento de los sistemas computacionales. De acuerdo con Fodor, tal propiedad común existe: todo proceso computacional, al igual que todo proceso cognitivo, consiste en la manipulación de "representaciones". Nada que no sea una manipulación de representaciones, de acuerdo con esto, es un proceso computacional (Fodor 1968b; 1975). Una representación es en este contexto un vehículo físico que "porta" y acarrea cierto tipo de información. En particular, las representaciones son vehículos de información *semántica*. Así, la justificación del componente *computacional* del computacionalismo funcionalista depende en último término de la corrección de lo que se ha denominado recientemente la "concepción semántica de la computación" (Piccinini 2008b; Shagrir 2006).

Puesto que el computacionalismo representa una hipótesis de acuerdo con la cual la dimensión cognitiva del cerebro –o bien de un sistema físico capaz de "realizar" la mente– consiste en la efectuación de computaciones y la concepción semántica de la computación afirma que los procesos de computación se individúan semánticamente –en el sentido de que toda computación consiste en el procesamiento mecánico de vehículos de información semántica–, la conjunción de ambas cosas implica que el computacionalismo entraña un modelo representacional de la mente. Este es el núcleo del computacionalismo funcionalista de Fodor. Fodor pensaba que la razón fundamental para creer que el cerebro es un sistema computacional es que la noción de *computación* constituye la única manera conocida de concebir fenómenos semánticos en términos físicos (Fodor 1975). ¿Existen sin embargo buenas razones para creer en la premisa esencial de estos razonamientos, a saber, que la computación consiste esencialmente en la manipulación de representaciones? Para determinarlo conviene hacer algunas aclaraciones relativas al concepto de *información*.

4.5.1. La noción de *información*

No toda información es información semántica y en consecuencia no todos los vehículos portadores de información acarrean información de este tipo. La teoría matemática de la información desarrollada inicialmente por Claude Shannon en los años cuarenta, para empezar, estudia la información en términos no semánticos.

De acuerdo con Shannon, la información es una medida de la incertidumbre con respecto al resultado de un "proceso estocástico" (Shannon 1948; Piccinini y Scarantino 2010). Un proceso de este tipo puede entenderse como una sucesión de variables aleatorias que evolucionan en función del tiempo, cada una de las cuales tiene asociada una distribución de probabilidad. Un ejemplo simple de un proceso de este tipo es la selección de palabras en la proferencia de un discurso improvisado. Después de que una palabra ha sido proferida, existe incertidumbre con respecto a cuál será la siguiente palabra elegida. La selección y ocurrencia de cada posible palabra en un momento dado del discurso tiene asociada cierta probabilidad. La idea central del concepto de información de Shannon es que la ocurrencia de una palabra *x* genera "más información" que la ocurrencia de una palabra *y* en la medida en que sea menos esperable a la luz de la distribución de probabilidad anterior a su ocurrencia. La proferencia de una palabra es en este sentido un evento físico que porta información. Shannon denominó a los portadores de información "mensajes". En términos generales la información generada por la selección u ocurrencia de un "mensaje" en un proceso estocástico es una función de qué tantos mensajes pueden ocurrir en vez del mensaje seleccionado y de la probabilidad de su ocurrencia. La noción de *mensaje* empleada por Shannon es una noción técnica que conviene no asociar con la noción no técnica, en cuya individuación intervienen factores semánticos. Un mensaje en el sentido de Shannon es un evento individuado en términos no-semánticos. Una manera simple de advertir esto es que la selección en el proceso descrito antes de una expresión carente de significado como "nafertodesx" generaría más información que la selección de una expresión significativa, en virtud de ser menos esperable (Piccinini y Scarantino 2010).

La información en sentido semántico, en cambio, tiene que ver con aquello a lo que un mensaje hace referencia o designa. Para determinar el significado semántico de un mensaje no es suficiente saber cuáles otros mensajes pudieron haber sido seleccionados en vez del mensaje en cuestión y con qué probabilidades. Mensajes equiprobables diferentes portan la misma información en el sentido no semántico de Shannon aunque signifiquen cosas diferentes. La noción semántica de información, sin embargo, tiene que ver también con la reducción de la incertidumbre; en este caso, la incertidumbre con respecto a cuál dentro de un conjunto de estados de cosas posibles es el caso (Piccinini y Shagrir 2014).

Existen dos nociones relevantemente diferentes de información semántica. Siguiendo una distinción introducida por Paul Grice, estas nociones se designan como "información natural" e "información no-natural". La información natural tiene que ver con correlaciones entre tipos de eventos. En este sentido, la ocurrencia de *humo* acarrea información sobre la ocurrencia de *incendio* y la ocurrencia de *manchas* acarrea información sobre la ocurrencia de *sarampión*. La relación entre estos tipos de eventos o variables es un "vínculo informacional". Los vínculos informacionales

que fundan los casos de información natural son *correlaciones fiables*. La ocurrencia de un tipo de evento *A* acarrea información natural sobre la ocurrencia de un tipo de evento *B* solo en el caso de que *A* se correlacione de manera fiable con *B*. Una correlación fiable es una correlación de la que un agente puede fiarse para hacer inferencias en cierto rango de circunstancias futuras y contrafácticas. En virtud de una correlación de este tipo, esto es, un agente puede inferir la ocurrencia de un incendio a partir de su percepción de la ocurrencia de humo (Piccinini y Scarantino 2010).

Los portadores de información natural –eventos o estados físicos– "significan" ciertos estados de cosas por tanto en virtud de estar físicamente conectados a estos. En ausencia de la conexión física apropiada, ninguna pieza de información natural es transmitida. Los portadores de información no natural, en cambio, no necesitan estar físicamente conectados de manera directa a aquello que "significan" o a lo que hacen referencia. Los portadores de este tipo de información son lo que los filósofos de la mente llaman "representaciones". Lo característico de las representaciones es que tienen "condiciones de satisfacción" y en virtud de esta propiedad pueden acarrear información correcta o incorrectamente. La condición de satisfacción de una representación –su "contenido"– es el estado de cosas posible y tal vez inexistente tal que su ocurrencia o no define la verdad o falsedad (o en términos generales la corrección o incorrección) de la misma (Egan 2019).

Las representaciones son estructuras físicas que pueden intervenir en procesos que definen la conducta de ciertos sistemas. Las representaciones, dicho de manera más precisa, pueden determinar causalmente la generación de los *outputs* de ciertos sistemas (Crane 2016). Considérese el ejemplo de un agente artificial diseñado para interactuar con su entorno de diferentes maneras. Un agente artificial puede estar diseñado, por ejemplo, para rastrear moscas y emitir el sonido "hay moscas en el ambiente". Que esté así diseñado significa que algunas de las estructuras físicas que lo constituyen tienen la función de reaccionar ante ciertos estímulos produciendo cierta conducta. Estas estructuras pueden tener esta función aun cuando no la hayan adquirido en virtud de su correlación con la presencia de moscas y aun si en alguna circunstancia realizan la función de manera incorrecta; esto es, si la proferencia del agente de "hay moscas en el ambiente" en un momento *t* es falsa. Así, estas estructuras cuentan como representaciones.

La noción general de información, de acuerdo con lo dicho, sirve para designar ciertas propiedades de estructuras físicas. Una estructura física es un vehículo de información, semántica o no, dadas ciertas condiciones. Una representación, en particular, es una estructura física que vehicula información en la medida en que tiene ciertas condiciones de satisfacción y es causalmente eficaz en la generación de los *outputs* de un sistema. La idea de Fodor es que *todos* los procesos de computación física están definidos sobre estructuras de este tipo. Los vehículos de un proceso

cualquiera de computación –sus *inputs* y *outputs*– son representaciones. ¿Qué razones hay para pensar esto?

4.5.2. Vehículos de computación

Como indiqué en los dos capítulos previos, Turing empleó la noción de "símbolo" para designar los vehículos manipulados en un proceso computacional. Una máquina de Turing estándar es un mecanismo abstracto que tiene dos componentes principales: una cinta potencialmente infinita en la cual están impresas ristras de símbolos y un dispositivo activo cuya función es moverse a lo largo de la cinta y escribir y borrar símbolos. Las instrucciones que definen el comportamiento del mecanismo en cada momento son también símbolos que el dispositivo activo tiene por función detectar y manipular de ciertas maneras definidas. Las ristras de símbolos que constituyen los *inputs*, los *outputs* y las instrucciones de una máquina de Turing *pueden* recibir interpretaciones semánticas, en el sentido de que pueden entenderse como codificaciones de números, fórmulas lógicas o listas de instrucciones computacionales. Al ser interpretadas de esta manera, el proceso efectuado por una máquina de Turing puede entenderse como la manipulación de vehículos con contenido semántico; esto es, como la manipulación de representaciones (Piccinini 2004b).

Este hecho parecería justificar la idea de que un proceso computacional es forzosamente un proceso de manipulación de representaciones. Turing, por otra parte, describió ocasionalmente en términos antropomórficos la conducta de los mecanismos computacionales que ideó. De acuerdo con estas descripciones, una máquina de Turing "escanea" y "ve" los símbolos impresos en su cinta, tiene "memoria", "estados mentales internos", etc. El empleo de este vocabulario mentalista, sin embargo, se dio siempre en el contexto de una explicación informal del comportamiento de sus máquinas y Turing entrecomilló siempre estos conceptos, con el presumible propósito de subrayar su carácter metafórico (Proudfoot y Copeland 2019; Piccinini 2008b). La cuestión clave, en cualquier caso, es la de la *individuación* de los vehículos y los estados y procesos constitutivos de un mecanismo computacional, abstracto o concreto. A este respecto, la posición defendida por el computacionalismo funcionalista de Fodor es que la individuación de un evento computacional se da siempre en términos semánticos. De acuerdo con lo que se ha denominado la "concepción semántica de la computación", los estados y procesos computacionales se individúan y taxonomizan en virtud de sus propiedades semánticas. Una computación es un proceso algorítmico definido sobre ristras semánticamente interpretadas de símbolos (Shagrir 2006).

Desde una perspectiva matemática estricta, sin embargo, la atribución de interpretaciones a los vehículos de un proceso computacional abstracto es una operación opcional, en todo caso irrelevante para entender la manipulación

algorítmica de los mismos. La individuación de un proceso computacional, en el sentido técnico articulado por Turing, se da en términos no semánticos. Una máquina de Turing se individúa en términos de una descripción formal de una secuencia de operaciones. Una descripción formal de este tipo especifica cuáles *inputs* entran en el mecanismo, cómo estos *inputs* afectan los estados internos del mecanismo y cuáles *outputs* se generan bajo ciertas circunstancias definidas. Aunque los vehículos computados se denominan "símbolos", el concepto de *símbolo* en juego aquí, al igual que el concepto de *mensaje* empleado por Shannon, es un concepto técnico en cuya comprensión son irrelevantes las connotaciones semánticas que la expresión exhibe en su sentido no-técnico. Un símbolo es un tipo de marca individuada por su forma geométrica. Estas marcas y su concatenación en ristras de marcas pueden o no recibir una interpretación semántica y en caso de hacerlo pueden ser interpretadas de diferentes maneras dependiendo de lo que quiera demostrarse en diferentes momentos (Piccinini 2008b). Los símbolos así definidos conforman un conjunto finito de tipos de marcas, denominado habitualmente un "alfabeto". La identidad de una máquina de Turing específica, tal es el punto central, no depende de cómo sean interpretados los símbolos. Es perfectamente posible diseñar una máquina que manipule símbolos que carecen de cualquier tipo de contenido. En la descripción del comportamiento de un mecanismo así es suficiente especificar cómo el dispositivo activo del mecanismo reacciona a la presencia de ciertas marcas en la cinta cuando se encuentra en cierto estado "interno". La interpretación semántica de los vehículos y estados de un proceso de computación es a lo sumo una manera de *glosar* ese proceso con los propósitos de entender la "tarea" que lleva a cabo y la razón por la que fue diseñado o descubierto (Egan 2019).

Al seguir una instrucción, un individuo "comprende" lo que la instrucción lo instruye a hacer, esto es, su significado. Por analogía, resulta tentador afirmar que un mecanismo computacional "comprende" las instrucciones que sigue, esto es, que responde a las propiedades semánticas de estas ristras de símbolos (Fodor 1968b). Esta manera de hablar puede resultar útil para glosar informalmente lo que el mecanismo hace pero es una manera incorrecta de describir su comportamiento si se toma al pie de la letra. Lo que el mecanismo "hace" es responder a las propiedades esenciales que individúan los símbolos; estas propiedades, como se dijo, son propiedades geométricas, no semánticas.

En consecuencia, identificar en términos generales las nociones de *computación* y *procesamiento de información* supone un error. Los mecanismos abstractos ideados por Turing y estudiados por la teoría matemática de la computación no pueden describirse como mecanismos de procesamiento de información en ninguno de los sentidos descritos antes; ergo, tampoco en el sentido semántico relevante para los propósitos del computacionalismo funcionalista. ¿Qué hay de los mecanismos de computación concretos?

Un mecanismo de computación concreto es un sistema físico que efectúa computaciones. Algunos artefactos computacionales concretos procesan computacionalmente información semántica de manera relevante para la explicación de su comportamiento. Algunos automóviles, por ejemplo, tienen incorporados artefactos computacionales que reciben y responden a *inputs* relativos a los estados del automóvil; esto es, procesan información relativa a estados del mecanismo físico del que hacen parte. Estos artefactos "usan" variables físicas del sistema al que están incorporados para regular actividades como la inyección de combustible, la velocidad y el tiempo de encendido. Las tareas de computación efectuadas por el artefacto se realizan por tanto sobre vehículos que acarrean información (Piccinini y Scarantino 2010). La información en cuestión es información natural. Los vehículos físicos en cuestión, esto es, acarrean información en virtud de su correlación física fiable con tipos de eventos relativos a los estados del automóvil. Algunos procesos de computación concretos, así, son procesos de manipulación de vehículos de información natural.

No *todos* los procesos de computación concretos, sin embargo, están definidos sobre vehículos físicos de este tipo. Algunos procesos están definidos sobre vehículos físicos que no acarrean información sobre el entorno del sistema del que hacen parte. Por otra parte, aun en los casos en los que los vehículos de una computación física sean vehículos de información natural, la explicación de las tareas efectuadas por el mecanismo computacional concreto se hace en términos no-semánticos; esto es, en términos de los vehículos procesados *qua* vehículos físicos y no *qua* vehículos de información (Piccinini y Scarantino 2010). Dos conclusiones pueden por tanto establecerse. En primer lugar, no basta con determinar que un proceso computacional es relevante para explicar el comportamiento de un sistema físico para concluir que ese proceso consiste en la manipulación de vehículos de información semántica. En segundo lugar, del hecho de que un sistema procese información semántica no se sigue que la explicación de su comportamiento –de los mecanismos en virtud de los cuales genera ciertos *outputs*– deba recurrir a las estructuras físicas que portan esta información *qua* vehículos de información.

El resultado de las consideraciones presentadas en las secciones previas es desfavorable a las pretensiones de Fodor. La noción matemática de computación no es una noción semántica y los procesos de computación física no son procesos causales de manipulación de representaciones.

4.5.3. Representacionalismo y computacionalismo

El punto de partida del computacionalismo fodoriano es una concepción representacionalista de la mente, de acuerdo con la cual los estados y procesos físicos que explican la conducta inteligente son representaciones: vehículos físicos con contenido semántico. La función cognitiva del cerebro es la representación del

entorno, de manera que la explicación de la conducta inteligente debe recurrir a procesos y estados dotados de contenido representacional. Los fenómenos cognitivos se interpretan como fenómenos relativos a la recepción, procesamiento y uso de contenido representacional (Egan 2012). Esta concepción no es en sí misma una hipótesis científica sino parte de una comprensión pre-teórica del ámbito de lo mental. La conversión del representacionalismo en una hipótesis científica depende de la posibilidad de modelar en términos mecánicos los procesos y estados físicos que subyacen a la producción de la conducta y a los que se atribuye contenido. Esta modelación mecánica es lo que en opinión de representacionalistas como Fodor hacen posible los formalismos desarrollados por la teoría matemática de la computación (Fodor 1975; Pinker 1997).

Los mecanismos computacionales abstractos ideados por Turing exhiben cómo un proceso mecánico puede "respetar" relaciones semánticas; cómo transiciones mecánicas entre estados de un sistema físico pueden capturar o "realizar" relaciones semánticas entre contenidos representacionales. Si los vehículos físicos sobre los que se definen los procesos de cualquier sistema computacional son representaciones y si el cerebro es un sistema computacional, entonces la cognición puede explicarse en virtud de mecanismos computacionales concretos y el representacionalismo entraña un modelo empírico sobre los procesos que explican la conducta inteligente. Pero los vehículos de los mecanismos computacionales abstractos, como se vio, no son entidades individuadas en términos semánticos y no existen tampoco razones para pensar que los vehículos sobre los que se definen los procesos computacionales concretos sean representaciones.

El sistema nervioso es presumiblemente un sistema físico que procesa información natural en el mismo sentido en el que un computador integrado a un automóvil. Parte de lo que hace este sistema puede *describirse* en términos de correlaciones causales entre impulsos nerviosos y variables en el entorno externo del organismo; correlaciones además causalmente influyente en los procesos mediante los cuales el sistema genera sus *outputs*. De este hecho, sin embargo, como mostré, no puede inferirse nada con respecto a los tipos de mecanismos que realizan y *explican* estos procesos.

La concepción semántica de la computación era la premisa clave con respaldo en la cual Fodor concluía el carácter computacional de los procesos explicativos de la cognición y la identificación entre modelos funcionales y modelos computacionales de la mente. Pero esta concepción de la computación carece de respaldo teórico y empírico, por lo que la conclusión de Fodor y su computacionalismo funcionalista en conjunto carece de justificación. A pesar de todo lo dicho por Fodor y los funcionalistas, la tesis de que el cerebro –o cualquier sistema físico que realice la cognición– es un sistema computacional en el sentido de que los mecanismos que

explican su conducta son mecanismos computacionales es una tesis que queda sin respaldo dentro de los márgenes del funcionalismo.

En las últimas dos décadas al menos se han presentado diferentes argumentos contra el computacionalismo sobre la base de un rechazo del representacionalismo como una hipótesis empírica (Van Gelder 1995; Brooks 1997). De acuerdo con un sector de los partidarios de una concepción "situada" de la cognición, las capacidades cognitivas habrían de explicarse sin la necesidad de postular estados mentales con contenido representacional ni procesos de computación sobre vehículos de información semántica. Sobre la base de este rechazo se ha defendido la necesidad de abandonar la hipótesis computacional y abrazar en cambio una forma de "dinamicismo". Para los defensores de lo que podría llamarse la "hipótesis dinamicista", la cognición habría de explicarse en cambio en términos de variables dinámicas que caracterizan diferentes tipos de interacciones entre agentes y los entornos en los que actúan. Estas variables se relacionarían de conformidad con formalismos matemáticos provenientes no de la teoría de la computación sino de la teoría de sistemas dinámicos (Faries y Chemero 2019). Más allá de los rendimientos teóricos que esta hipótesis alternativa pueda ofrecer, la crítica anti-representacionalista del computacionalismo sobre la que en parte se funda constituye un *non-sequitur*. En la medida en que esta crítica tenga validez concierne únicamente a las formas de computacionalismo comprometidas con una concepción semántica de la computación. Incluso si la noción de *representación* no tiene ninguna función sustantiva en la explicación de la cognición, esto no supone evidencia contraria con respecto al computacionalismo *per se*. Como expondré en el siguiente capítulo, el computacionalismo puede interpretarse y justificarse como una hipótesis empírica al margen de cualquier compromiso con el representacionalismo sobre la mente.

4.6. Conexionismo y computacionalismo

En parte en virtud de las dificultades reseñadas en este capítulo para justificar teóricamente la analogía entre la mente y los computadores digitales programables, en los años ochenta del siglo pasado tuvo lugar en el ámbito de la ciencia cognitiva el resurgimiento de ideas inspiradas en el modelo computacional de McCulloch y Pitt. Este resurgimiento, asociado con el nombre de "conexionismo" se percibió inicialmente como un evento "revolucionario" en el estudio científico de la cognición. Los conexionistas cuestionaban la estrategia clásica de modelar la cognición en términos de programas computacionales, basada en la concepción de la mente como el software del cerebro (Churchland y Sejnowski 1992). En su opinión, los modelos explicativos de la cognición debían ajustarse de manera más estricta a los datos conocidos sobre el funcionamiento del sistema nervioso y las herramientas formales

empleadas en estos modelos debían estar al servicio de la descripción de los mecanismos explicativos de ese funcionamiento (Stinson 2019).

Conviene recordar en este contexto que las máquinas de Turing son solo uno de entre varios tipos posibles de mecanismos computacionales abstractos. No todos los mecanismos computacionales, en particular, efectúan computaciones mediante el seguimiento de instrucciones ni la ejecución de programas. McCulloch y Pitt idearon, como describí en el segundo capítulo, un tipo de mecanismo computacional abstracto que computaba funciones lógicas cuya operación, a diferencia de una máquina de Turing, no estaba guiada por instrucciones. Estos mecanismos se entendían como descripciones abstractas de una neurona, por lo que un conjunto adecuadamente ensamblado de los mismos representaba una especie de red neuronal abstracta. El objetivo de McCulloch y Pitt era modelar mediante estos mecanismos abstractos el comportamiento de las neuronas y de las redes neuronales que constituían el sistema nervioso y explicar las capacidades cognitivas del cerebro como un producto de la actividad neuronal. Las redes neuronales abstractas descritas por McCulloch y Pitt fueron denominadas posteriormente "sistemas conexionistas" (Piccinini 2008c).

Un sistema conexionista está conformado por unidades funcionales que reciben *inputs* (unidades de entrada), unidades que exhiben *outputs* (unidades de salida) y unidades que se comunican con otras unidades del sistema (unidades ocultas). Estas unidades están organizadas en estratos o "capas", de manera que un sistema conexionista puede tener una o varios estratos de "unidades ocultas" que median entre las unidades de entrada y las de salida. Cada unidad recibe estímulos de entrada y produce estímulos de salida como una función del tipo de entrada y del estado de la unidad. Cada unidad del sistema –cada "neurona artificial"– tiene "valores de activación" que determinan la manera en que sus *inputs* afectan su conducta. Como un resultado de las actividades de sus unidades y de su organización, los sistemas conexionistas convierten el *input* recibido por sus unidades de entrada en el *output* producido por sus unidades de salida (Sejnowski, Koch y Churchland 1988; Bermúdez 2014). Un sistema conexionista es en este sentido un tipo de mecanismo abstracto que produce *outputs* a partir de ciertos procesos definidos de manipulación de sus *inputs*. En el caso de los sistemas ideados por los conexionistas de los ochenta, las manipulaciones en cuestión eran susceptibles de descripción algorítmica, por lo que estos sistemas eran entendidos como mecanismos computacionales abstractos, en el mismo sentido que las máquinas de Turing. Los mapeos de *inputs* a *outputs* producidos por los sistemas conexionistas paradigmáticos pueden caracterizarse mediante el mismo tipo de formalismos empleados por la teoría matemática de la computación en la descripción de los sistemas de computación "clásicos".

Los sistemas conexionistas son mecanismos de computación no clásicos. Los sistemas de computación clásicos son mecanismos susceptibles de "descomposición computacional" en el siguiente sentido. En su conjunto, el mecanismo efectúa una

tarea de computación. Por ejemplo, computa la función factorial de su *input*. La efectuación de tal computación puede explicarse en términos de sus componentes y de las tareas que estos efectúan. La computación de un factorial puede explicarse por ejemplo mediante la acción conjunta de un componente de memoria que almacena un programa y un procesador que ejecuta ese programa. Esta estrategia de explicación, tal es el punto crucial, puede repetirse. La capacidad del procesador para ejecutar programas puede explicarse mediante las computaciones que sus componentes (las llamadas unidades de "control" y de "proceso") efectúan. Y las computaciones efectuadas por estos componentes pueden explicarse a su vez por sus propios componentes (los llamados circuitos booleanos) y la manera en que están organizados. Dado que las operaciones efectuadas por estos últimos componentes del sistema son computacionalmente primitivas, la descomposición computacional termina en ese punto. Un elemento distintivo de la computación clásica asociada a esta posibilidad de descomposición es su carácter "paso-a-paso". Un proceso de computación clásica procede realizando una operación a la vez, donde una "operación" consiste en una modificación definida de los vehículos del proceso. Puesto que estos vehículos son entidades discretas, las operaciones en cuestión y la dinámica en su conjunto de un mecanismo de computación clásico son discretas (Piccinini 2008c).

Los sistemas conexionistas paradigmáticos se caracterizan por carecer de estas dos propiedades descritas. Estos sistemas no producen sus *outputs* en virtud de procesos "paso a paso". En estos sistemas, cada unidad afecta la activación de las unidades con las que está conectada según sus diferentes pesos de activación y en virtud de relaciones dinámicas que varían en tiempo continuo. Es gracias a esta característica que su operación no es susceptible de descomposición computacional. Esto es, aunque estos sistemas tienen capacidades computacionales, estas no pueden explicarse en términos de computaciones más simples efectuadas por sus componentes junto con la manera en que estos componentes están organizados. Por otra parte, a diferencia de los sistemas clásicos, los sistemas conexionistas no tienen una estructura fija. Las conexiones entre sus unidades pueden modularse en el tiempo sin alterar la arquitectura del sistema para producir diferentes *outputs* dados los mismos *inputs*. En estas modificaciones los valores de activación de las unidades cambian de manera que el comportamiento de todo el sistema se modifica sin alterar cuáles unidades están conectadas con cuáles otras. En este sentido, se dice que los sistemas conexionistas son susceptibles de "entrenamiento". Después de un periodo de tiempo en el que sistema experimenta un cambio en los valores de activación de sus unidades ("periodo de entrenamiento"), estos valores se estabilizan y el sistema en su conjunto genera *outputs* definidos dados *inputs* definidos de maneras susceptibles de descripción mediante los formalismos de la teoría de la computación (Bringsjord y Govindarajulu 2018).

4.6.1. El conexionismo y el problema de la justificación del computacionalismo

Un sistema conexionista, al igual que una máquina de Turing, puede ser un tipo de sistema físico concreto o un sistema matemático abstracto (Piccinini 2008c). En su dimensión abstracta, un sistema conexionista, también al igual que una máquina de Turing, puede entenderse como un modelo o representación de un mecanismo concreto. La pretensión de los conexionistas en los años ochenta era que los sistemas conexionistas constituían un mejor tipo de modelo del cerebro que los sistemas clásicos y, a diferencia de estos, podían emplearse para explicar de manera realista sus capacidades cognitivas. Las diferentes capacidades cognitivas del cerebro serían el resultado de diferentes tipos de conexiones y de la actividad neuronal determinada por estas conexiones entre las unidades del sistema nervioso. Los sistemas conexionistas abstractos se pretendían descriptivos de redes neuronales concretas y en ese sentido se afirmaba que eran modelos biológicamente más realistas. El problema al que se enfrentaban los modelos clásicos de justificar la atribución al sistema nervioso de procesos computacionales explicativos de su conducta no sería un problema para los modelos conexionistas, supuesto que los mecanismos postulados a través de los modelos conexionistas fueran mecanismos fundados en una descripción realista de la actividad del cerebro. En la retórica conexionista, esto suponía un abandono del marco de referencia funcionalista de la ciencia cognitiva clásica y de su compromiso con la idea de *autonomía*. Los modelos conexionistas se pretendían modelos descriptivos y explicativos de las estructuras físicas que realizan la cognición, de manera que esta podía de hecho reducirse a fenómenos bioquímicos y físicos (Sejnowski, Koch y Churchland 1988).

La premisa en que se respaldaban estas pretensiones carecía sin embargo de fundamento. Los modelos conexionistas propuestos en los años ochenta en contraposición a los modelos clásicos *no* eran modelos en modo alguno biológicamente realistas. En el diseño de estos modelos, los conexionistas se sirvieron de presuposiciones arbitrarias sobre el número de neuronas y de estratos de unidades, sobre los mecanismos que mediaban la conectividad entre las neuronas y que subyacían a su respuesta a estímulos por parte de otras neuronas, así como sobre las formas en que las redes neuronales reales alteran sus patrones de conectividad (Boone y Piccinini, 2015). En resumen, los modelos propuestos por los conexionistas no estaban fundados en los mecanismos y procesos neuronales conocidos por la neurociencia del momento. Aunque el cerebro está sin duda constituido por neuronas organizadas en redes interconectadas, es seguro que no está constituido por las neuronas y las redes descritas por los modelos conexionistas (Marcus, 2014). Como consecuencia, estos modelos no eran menos abstractos que los modelos clásicos y la actividad teórica desplegada por los conexionistas manifestaba en último término un compromiso implícito con la autonomía del estudio científico de la cognición no

menor que el de los funcionalistas. El conexionismo supuso por tanto no una revolución sino un cuestionamiento de aspectos relativamente periféricos del programa de la ciencia cognitiva clásica. Que es de hecho así puede confirmarse a partir del compromiso de algunos de los autores más representativos del conexionismo con tesis nucleares del computacionalismo funcionalista. Los autores más prominentes del movimiento suscribían por ejemplo la concepción semántica de la computación (Sejnowski, Koch y Churchland 1988, p. 1300) e incurrían en la falacia Church-Turing (Churchland y Churchland 1990, p. 32). Por otra parte, una evidencia de la congruencia entre el programa clásico fundado y el conexionismo es que los sistemas conexionistas terminaron siendo incorporados en el programa de la inteligencia artificial como herramientas tanto en la construcción de agentes artificiales en el sentido descrito antes como en la construcción de artefactos tecnológicos capaces de realizar tareas más circunscritas como la clasificación de imágenes y el reconocimiento de la voz (Marcus 2014; Bringsjord y Govindarajulu 2018).

El conexionismo, en consecuencia, no resuelve la cuestión central que deja abierta el computacionalismo funcionalista y de la que dependen las aspiraciones teóricas de la ciencia cognitiva: ¿cómo justificar la tesis de que el cerebro –o cualquier sistema físico con capacidades cognitivas– es un sistema que efectúa computaciones?

4.7. Computación física, funcionalismo y nihilismo computacional

El proyecto de mecanización de la cognición y del cerebro se fundó en la idea de que los formalismos matemáticos desarrollados por la teoría de la computación aportaban las herramientas necesarias para modelar la cognición en términos mecánicos y hacerla así un fenómeno susceptible de estudio y explicación científica. En sí mismos, sin embargo, los formalismos en cuestión son descripciones de entidades abstractas. Antes de que estos formalismos puedan emplearse para explicar –como algo opuesto a modelar en términos matemáticos abstractos– el comportamiento de sistemas físicos concretos, es preciso desarrollar una teoría o explicitación adecuada de la relación entre los formalismos y el mundo físico (Ritchie y Piccinini 2019).

El reconocimiento de los problemas asociados al concepto positivista de reducción llevó a diferentes filósofos de la ciencia en la segunda mitad del siglo pasado a defender una concepción anti-reduccionista de la ciencia y de la explicación científica, de acuerdo con el cual el mundo natural podía estudiarse en dos niveles disjuntos e irreducibles uno al otro: el nivel físico-estructural y el nivel funcional. Las ciencias no básicas como la biología y la ciencia cognitiva estudiarían de acuerdo con esto el mundo natural desde una perspectiva funcional. En este contexto, el computacionalismo funcionalista emergió como una propuesta de conceptualización

de la cognición de acuerdo con la cual las capacidades cognitivas son capacidades de ciertos sistemas físicos: aquellos cuya "organización funcional" es computacional. La organización funcional de un sistema puede entenderse como un mecanismo complejo en el sentido amplio propio de la filosofía mecanicista, si bien un mecanismo individuado en términos abstractos, cuyos componentes y organización son múltiplemente realizables. La especificación de una organización funcional es la especificación de los componentes de un sistema y de la manera en que estos interactúan para generar un cierto tipo de *outputs* del sistema. Estos componentes e interacciones no se identifican sin embargo, tal es el punto crucial, con estructuras físicas y procesos físicos específicos. Lo importante desde una perspectiva funcional es "qué hace" un sistema, no "cómo" o mediante qué recursos físicos lo hace.

La única relación relevante entre la organización funcional de un sistema y las entidades y estructuras físicas que lo individúan *qua* sistema físico es que estas últimas "realizan" la organización funcional descrita por los científicos que estudian el sistema desde una perspectiva funcional. En ausencia de una vinculación estrecha entre la organización funcional y las estructuras y procesos que individúan el sistema *qua* sistema físico, sin embargo, resulta imposible sustanciar la tesis de que la organización funcional que explica un conjunto de las capacidades del sistema es una organización de un tipo o de otro. En particular, resulta imposible sustanciar la tesis de que la organización funcional que explica las capacidades del cerebro o de un sistema cognitivo es una organización computacional. Tal ha sido el principal resultado de las consideraciones desplegadas en este capítulo.

Supuesto un anti-reduccionismo de partida, el computacionalismo funcionalista se propuso explicitar la noción de *mecanismo* en términos de la noción de *computación*. Esta última noción era adecuadamente abstracta para impedir cualquier compromiso con una realización física. Una consecuencia de esta estrategia es lo que algunos autores han denominado "nihilismo computacional" (Glennan 2017; Piccinini 2008a). El nihilismo computacional es un tipo de relativismo con respecto a la idea de computación física de acuerdo con la cual no hay nada intrínseco ni distintivo de los procesos físicos categorizados como computaciones que funde esa categorización. Si algo es una computación física, de acuerdo con esto, depende de la manera en que el observador decida interpretarlo. Basta que las transiciones de estado de un sistema físico –su evolución dinámica– puedan mapearse de alguna manera con transiciones de un proceso de computación abstracta para que el sistema pueda categorizarse como un sistema computacional. En palabras de dos representantes insignes del conexionismo: "[…] Sieves and threshing machines could be construed as computers if anyone has reason to care about the specific function reflected in their *input-output* behavior" (Churchland y Sejnowski 1992, pp. 65–66).

El nihilismo computacional entraña una posición eminentemente contraintuitiva. Si fuera correcta, por ejemplo, supondría que la invención de los computadores

electrónicos programables no fue en modo alguno un logro intelectual y técnico y que disciplinas como la ciencia de la computación y la ingeniería computacional carecen de un objeto de estudio circunscrito. La única razón para aceptar este tipo de relativismo es la ausencia de una alternativa viable. Por supuesto, existe tal alternativa; pero su reconocimiento, como expondré en el próximo capítulo, depende del abandono del marco de referencia anti-reduccionista del funcionalismo.

CAPÍTULO 5

5.1. Introducción

Una computación es un proceso "mecánico" de determinación de un valor o resultado definido. Dado un valor o un conjunto de valores de entrada, el proceso consiste en el seguimiento automático de una serie de reglas como resultado del cual se genera un valor de salida. Los valores en cuestión son valores de funciones matemáticas y las reglas son procedimientos formales diseñados para resolver este tipo de funciones. Las computaciones representan en este sentido un ámbito de estudio matemático. Los formalismos abstractos desarrollados en este campo, sin embargo, han inspirado el desarrollo de hipótesis *explicativas* en las ciencias no-formales. De acuerdo con estas hipótesis, algunos fenómenos naturales pueden *explicarse* como el resultado causal de computaciones efectuadas por sistemas físicos. En una *explicación computacional* el explanandum es una capacidad de un sistema y el explanans es un proceso o mecanismo de computación. Como cualquier explicación científica, una explicación computacional consiste en la exhibición de una relación de dependencia entre un explanandum y un explanans. ¿Cuál es exactamente la relación de dependencia en juego en este caso? En particular, dado el carácter abstracto de los procesos de computación, una cuestión central es: ¿cómo pueden estos procesos ser explanantia de fenómenos naturales?

De acuerdo con el modelo nomológico-deductivo dominante en la filosofía de la ciencia en los años de la posguerra, en las explicaciones científicas los explanantia son siempre fundamentalmente leyes y las relaciones de dependencia explicativa son relaciones de subsunción deductiva: un explanandum *depende* y es explicable por un explanans en el sentido de que este subsume y permite deducir la ocurrencia de aquel. Para que una explicación computacional sea una explicación científica, en consecuencia, los procesos de computación deben ser tipos de leyes de las que pueden derivarse capacidades de un sistema.

A pesar de su rechazo del modelo positivista de reducción inter-teórica, los funcionalistas no cuestionaron esta concepción de las explicaciones científicas. La descripción de un sistema como una entidad capaz de efectuar computaciones es desde el punto de vista funcionalista explotable para efectos explicativos en la medida en que esta descripción permita deducir y predecir el comportamiento del sistema. El funcionalismo computacional, sin embargo, no ofrece una explicitación satisfactoria de la manera en que los procesos de computación pueden ser explicativos. En los términos empleados en el capítulo anterior, el funcionalismo carece de las herramientas necesarias para resolver el *problema de la computación física*. Una consecuencia de esta impotencia es la aceptación de un tipo de nihilismo

computacional de acuerdo con el cual cualquier sistema puede concebirse y explicarse computacionalmente.

En este capítulo expondré una concepción no funcionalista de las explicaciones computacionales que elude las dificultades mencionadas y ofrece una explicitación satisfactoria del sentido en el que los procesos de computación pueden tener potencial explicativo. Esta concepción es el resultado de la aplicación al ámbito de la computación de ideas sobre la naturaleza de las explicaciones en biología desarrolladas en el seno de una nueva filosofía de la ciencia: el nuevo mecanicismo.

En la sección 2 expongo el componente central de la crítica a la filosofía positivista de la ciencia efectuada por Thomas Kuhn, con un énfasis en la idea de *adecuación descriptiva* como un criterio evaluativo básico de las afirmaciones meta-teóricas sobre la ciencia. En la parte final de la sección presento el nuevo mecanicismo como una filosofía de la ciencia heredera de la crítica kuhniana. La sección 3 pasa revista a la discusión post-positivista en filosofía de la ciencia sobre las nociones de *causalidad* y *explicación*. Como se verá, esta discusión recapitula en buena medida la discusión reseñada en los capítulos previos entre el mecanicismo clásico y el Newtonianismo sobre la ciencia. En la sección 4 deescribo la concepción de las explicaciones científicas como explicaciones *constitutivas* desarrollada por el nuevo mecanismo. Esta concepción, según se expondrá, representa tanto una elucidación descriptivamente más adecuada del tipo de explicaciones típicas de las ciencias biológicas como una puntualización de la idea de explicación implícita en el mecanicismo clásico. En la sección 5 defiendo la idea según la cual la perspectiva meta-teórica sobre la ciencia desarrollada por los nuevos mecanicistas permite articular una posición intermedia en la discusión entre el reduccionismo positivista y el anti-reduccionismo fundado en la idea de *realizabilidad múltiple*. De acuerdo con esta posición, las explicaciones científicas comportan un componente reductivo pero esto no supone el carácter dispensable de las ciencias especiales. En la sección final expongo una concepción mecanicista de la computación física y de las explicaciones computacionales como una alternativa a la interpretación funcionalista del computacionalismo discutida en los dos capítulos previos.

.

5.2. La crítica kuhniana a la filosofía positivista de la ciencia

La tarea de la filosofía de la ciencia consistía para los positivistas lógicos no en el examen de los procesos y actividades en virtud de los cuales los científicos acopian conocimiento y proponen representaciones del mundo natural sino en el análisis directo de esas representaciones, reconstruidas en términos lógicos como enunciados de un lenguaje formal. De acuerdo con la famosa distinción introducida por Hans Reichenbach, el ámbito de la reflexión filosófica sobre la ciencia era el "contexto de

justificación" y no el "contexto de descubrimiento" del conocimiento. Los factores culturales, institucionales y en general contextuales relevantes para entender la aparición de una teoría o representación científica como un producto histórico se estimaban desde este punto de vista como irrelevantes para la filosofía de la ciencia. La distinción de Reichenbach consagraba una división del trabajo en el estudio de la ciencia: la filosofía se ocuparía del contexto de justificación y disciplinas como la historia, la sociología o la psicología se ocuparían del contexto de descubrimiento (Bailer-Jones 2009). La presuposición crucial que subyacía y respaldaba este esquema de trabajo era la de que las propiedades epistémicas de las representaciones científicas podían delimitarse y evaluarse independiente y autónomamente de consideraciones relativas al contexto de descubrimiento. Esta presuposición era natural para los positivistas lógicos en virtud de su compromiso con un tipo de empirismo humeano de acuerdo con el cual todo conocimiento del mundo natural se deriva de la experiencia, entendida como una fuente de datos teóricamente neutros sobre la realidad. Así, si un enunciado o representación expresa conocimiento empírico, *debe* poder determinarse el conjunto de experiencias posibles de las cuales se deriva ese conocimiento (Carnap 1991). Con prescindencia de las virtudes de esta concepción del conocimiento desde la perspectiva de la epistemología filosófica, un resultado perdurable de la filosofía de la ciencia del siglo XX es que su uso en el análisis y estudio de la actividad científica es problemático.

Thomas Kuhn mostró que la adopción de una perspectiva más *empírica* en el estudio de la ciencia obligaba a reconsiderar la distinción entre contexto de justificación y contexto de descubrimiento y socavaba uno de los núcleos de la filosofía positivista de la ciencia: su concepción empirista de la evaluación epistemológica de las representaciones científicas. De acuerdo con Kuhn, el estudio de la ciencia como una actividad e institución histórica pone de manifiesto que la práctica científica obedece a esquemas generales implícitos que orientan el trabajo de los científicos y que carecen del tipo de justificación empirista consagrado por los positivistas. Los practicantes competentes de una disciplina científica exhiben en cada momento de la historia de la misma un acuerdo general sobre cuáles problemas merecen investigación, cómo ha de proceder esa investigación y qué cuenta como una resolución de esos problemas. Este acuerdo general resulta de presupuestos de tipo teórico, experimental, metodológico y axiológico que conjuntamente constituyen lo que Kuhn llamó "paradigmas". El punto crucial es que los paradigmas no son propiamente objetos susceptibles de justificación empírica. La justificación empírica de cualquier conjetura científica, por el contrario, depende del paradigma en términos de cuyos supuestos se formula. Evidencia de esto de acuerdo con Kuhn era la manera en la que tiene lugar el cambio teórico en la historia de las disciplinas científicas. (Kuhn 1996).

Los paradigmas desempeñan una serie de funciones críticas en la actividad científica que los positivistas asignaban a la contrastación de las representaciones científicas a través de conjuntos de datos empíricos: establecen las condiciones de *validación* de las conjeturas científicas, fijan los términos mismos en los cuales se formulan estas conjeturas y definen las circunstancias del cambio histórico en la ciencia. Cuáles son las pautas cognitivas en términos de las cuales se *construyen* las representaciones científicas, en qué circunstancias estas representaciones son epistémicamente *aceptables* o no y en consecuencia cuándo corresponde modificar estas representaciones son todos factores que en la concepción kuhniana de la ciencia determinan los paradigmas y no la experiencia (Kuhn 1996). Una heurística, entendida en el sentido explicitado en el primer capítulo como una serie de directrices que orientan la investigación en un ámbito de estudio determinado, puede concebirse de esta manera como parte de un paradigma compartido por una población de investigadores. El mecanicismo clásico discutido en los primeros capítulos, comprometido metodológicamente con una heurística de descomposición y localización, es de acuerdo con esto un tipo de paradigma general

Más allá de los detalles específicos de la propuesta positiva de la filosofía de Kuhn, dos conclusiones de su crítica a la imagen positivista de la ciencia tuvieron un impacto perdurable: la disolución de una distinción clara entre contexto de justificación y contexto de descubrimiento, y el reconocimiento de la importancia ineludible del estudio de la ciencia desde una perspectiva empírica e histórica. Después de Kuhn, la filosofía de la ciencia ya no fue concebible como una disciplina independiente y autónoma con respecto al estudio de la ciencia como un fenómeno histórico concreto. La filosofía de la ciencia hace parte de una empresa más amplia de estudio de la ciencia, entendida como un conjunto de instituciones creadas por los seres humanos para representar, entender, predecir y controlar el mundo natural. En consecuencia, las afirmaciones que sobre la ciencia haga la filosofía de la ciencia deben evaluarse y validarse con atención a las características de la ciencia como un fenómeno empírico. Después de Kuhn, dicho brevemente, la *adecuación descriptiva* de las afirmaciones filosóficas y en general meta-teóricas sobre la ciencia con respecto a las prácticas científicas se convirtió en un principio evaluativo básico. Si los modelos desarrollados para describir e interpretar la práctica científica no se ajustan a esta, aun cuando estén respaldados por ideas filosóficas aparentemente correctas, la razón de este desajuste debe buscarse en primer lugar en los modelos y no en la práctica científica.

El nuevo mecanicismo es un movimiento filosófico heredero de las ideas de Kuhn. Los nuevos mecanicistas enfatizaron la importancia del estudio de los procesos de descubrimiento, de la descripción de las prácticas científicas y del análisis de los conceptos metacientíficos a la luz de la evidencia empírica aportada por el examen de casos concretos de la historia de la ciencia (Bechtel y Hamilton 2007, Glennan 2016,

Craver 2007). Este movimiento tuvo su origen en el contexto de la revisión crítica de las implicaciones de la imagen positivista de la ciencia con respecto a las ciencias biológicas. Desde principios de los años setenta, autores como David Hull, Stuart Kauffman y, en especial, William Wimsatt, llamaron al desarrollo de una filosofía de la biología que, en contra de la concepción general de la ciencia consagrada por el positivismo lógico, diera cuenta de manera más fiel de las prácticas teóricas que conformaban la disciplina. De acuerdo con estos autores, la interpretación positivista de la ciencia en términos de conceptos como *teoría*, *ley* y *reducción interteórica* podría tal vez tener sentido en el contexto de la física pero no en el ámbito de la caracterización de los fenómenos biológicos y de las actividades de los biólogos. Más que el concepto de *teoría* –en el sentido lógico-formal explotado por los positivistas lógicos–, sería el concepto de *modelo* el que resulta adecuado para describir la manera en que los biólogos representan sus fenómenos de interés; y más que el concepto de *ley*, es el concepto de *mecanismo* el que permite interpretar adecuadamente el tipo de conocimiento acopiado en biología y empleado en la formulación de explicaciones (Glennan 2016).

Aunque inicialmente el nuevo mecanicismo fue solo una filosofía de la biología, en los últimos treinta años las herramientas teóricas desarrolladas por este movimiento se han empleado en la caracterización de las prácticas descriptivas y explicativas de la neurociencia, la ciencia cognitiva y la psicología, las ciencias sociales e incluso la física (Glennan 2017). De esta manera, el nuevo mecanicismo se ha transfigurado en una filosofía general de la ciencia.

Mientras la filosofía positivista, en la tradición newtoniana descrita en el tercer capítulo, ataba el concepto de *explicación* a los conceptos de *ley* y *teoría*, el nuevo mecanicismo concibe las explicaciones científicas en términos de *mecanismos* y *modelos*. Entender las razones y el trasfondo histórico de esta transición es indispensable para calibrar y evaluar tanto las ambiciones teóricas generales del nuevo mecanicismo como la manera en que este es empleado en el ámbito de la computación. En la próxima sección examinaré la manera en que la revisión crítica de la filosofía positivista a partir de los años setenta condujo a una reconsideración de la naturaleza de la causalidad y de su relación con las prácticas explicativas de la ciencia.

5.3. Causalidad y explicación: entre leyes y mecanismos

En el fondo de la oposición referida entre el positivismo lógico y el nuevo mecanicismo hay un desacuerdo fundamental sobre la naturaleza de la *causalidad* y el uso de la noción de *causa* en la ciencia. Esta oposición, como se verá, reproduce en buena medida la discusión originada por la crítica newtoniana del mecanicismo clásico reseñada en los capítulos 1 y 3.

5.3.1. La concepción regularista de causalidad y sus limitaciones

Los positivistas lógicos dieron por sentada la corrección del análisis humeano de la idea de *ley natural* fundamental en la concepción newtoniana de la ciencia y del correspondiente concepto de *causalidad*. En su opinión, el logro de Hume fue articular una explicitación *reductiva* de la relación de causalidad entre eventos. Una explicitación, en particular, que liberó el discurso sobre relaciones causales de cualquier compromiso con una noción no epistémica de *necesidad* entre causas y efectos. La relación de causalidad no es de acuerdo con esto nada más que un tipo de dependencia funcional entre variables o estados temporalmente sucesivos. Morris Schlick, uno de los fundadores del positivismo lógico, expresó esta idea de causalidad de manera particularmente clara: "the difference between a mere temporal sequence and a causal sequence is the regularity, the uniformity of the latter. If *C* is *regularly* followed by *E*, then *C* is the cause of *E*; if *E* only 'happens' to follow *C* now and then, the sequence is called mere chance" (citado por Psillos 2007, pp. 121-122). Aquello que *hace verdaderas* las aserciones causales es el conjunto de patrones de regularidad constatados en la experiencia de la naturaleza. El concepto de *causalidad* se reduce en esta tradición empirista al de *ley* y este al de *sucesión regular*. Aunque las relaciones de dependencia causal y de dependencia explicativa tengan un componente modal, como se indicó en el tercer capítulo, esta modalidad es estrictamente epistémica. La *necesidad* de las sucesiones regulares que se toman como leyes y de las relaciones que se estiman causales son proyectadas por la mente, no descubiertas en los fenómenos.

El razonamiento que subyace a la explicitación empirista de estos conceptos puede entenderse como un razonamiento basado en el principio conocido como la "Navaja de Ockham" (Psillos 2007). De acuerdo con una versión simple de este principio, es preferible una explicación más simple o "económica" de un fenómeno o explanandum –esto es, una explicación cuyo explanans implique menos recursos teóricos– que una explicación más compleja o "costosa", dado por supuesto que la evidencia disponible no respalde decisivamente a ninguna de las dos. En este caso, el explanandum en cuestión es el uso de los conceptos de *ley* y *causalidad* en la práctica científica y en el discurso cotidiano. El explanans más simple es el que da cuenta de estos conceptos en términos del concepto de *sucesión* regular. Los explanantia más costosos son aquellos que, como los propuestos por los mecanicistas, postulan la existencia de entidades y propiedades no observables o bien observables solo bajo ciertas condiciones especiales. Para los mecanicistas, como he indicado repetidamente, las regularidades no son aquello que hace verdadera a una aserción causal sino solo *evidencia* de la existencia de relaciones causales. Evidencia, en particular, que requiere de respaldo y explicación a través de la determinación de mecanismos físicos en la naturaleza. ¿Por qué recurrir a propiedades y procesos no observables si es posible explicitar satisfactoriamente la noción de *causalidad* y con

ella la de *explicación* en términos de patrones regulares empíricamente constatables? La navaja de Ockham prescribe suprimir explanantia superfluos.

La tradición newtoniana representada y sistematizada por el positivismo lógico dio por sentada la corrección del análisis más económico de la causalidad como un tipo de propiedad de regularidades empíricamente constatables. Esta concepción dominó la reflexión sobre la ciencia desde finales del siglo XIX hasta la primera mitad del siglo XX. La inadecuación descriptiva del modelo positivista de reducción inter-teórica y la reconsideración subsecuente del análisis positivista de la ciencia, sin embargo, aportaron evidencia en el sentido de que la concepción humeana de la causalidad era inadecuada para efectos de dar cuenta de la práctica científica. El resultado es que los explanantia más "costosos" desde una perspectiva empirista se han revelado después de todo necesarios y han sido reivindicados por la filosofía de la ciencia post-kuhniana.

Una de las conclusiones resultantes de la evaluación del positivismo a la luz de la práctica científica fue justamente el reconocimiento de que la concepción "regularista" de la causalidad no captura algunos de los usos nucleares que la noción tiene en manos de los científicos (Illari y Russo 2014). Además del hecho conocido y aceptado por los positivistas de la existencia de regularidad sin causalidad (por ejemplo, la sucesión regular del día con respecto a la noche), en la actualidad es ampliamente aceptado también que existen relaciones causales sin sucesión regular. Un ejemplo claro de esto es la relación causal entre el consumo de tabaco y el desarrollo de cáncer de pulmón. Aunque la relación causal entre estas dos variables es aceptada ampliamente, su "regularidad" es en el mejor de los casos parcial. La regularidad en cuestión es no-estricta o no universal: muchos fumadores nunca desarrollan cáncer de pulmón mientras que algunos no-fumadores de hecho lo desarrollan.

En el marco del regularismo, del carácter no estricto de la regularidad existente entre fumar y desarrollar cáncer de pulmón puede concluirse una de dos cosas. Una es afirmar que la regularidad en cuestión no es una ley y en consecuencia que la relación entre ambas variables no es una relación causal. Esta opción contradice una opinión científica bien establecida y resulta por tanto difícil de aceptar. La segunda opción es afirmar que la regularidad en cuestión es una ley probabilística. Una ley probabilística es una correlación estadística entre variables. Una ley de este tipo es *causal* si la ocurrencia o instanciación de una variable eleva significativamente la probabilidad de la ocurrencia de otra; esto es, si, dado un conjunto de datos aceptable, se evidencia una "dependencia robusta" de una variable con respecto a la otra (Díez y Moulines 1999). Qué cuenta como una incremento significativo de la probabilidad de una variable y qué cuenta como una dependencia robusta son cuestiones técnicas y su respuesta depende en parte de los conjuntos de datos analizados. En general, la relación entre causalidad y probabilidad es especialmente compleja y para efectos de

la presente discusión carece de importancia. El punto relevante es la constatación del hecho de que la noción de *sucesión regular* no agota la noción de *causalidad*, como esta se presenta y es empleada en la práctica científica. Una consecuencia crucial de esto es que si la explicación científica de un fenómeno consiste en la exhibición de sus causas, el modelo nomológico-deductivo no representa adecuadamente los explanantia de las explicaciones científicas.

Lo anterior no significa sin embargo que las nociones de *causalidad* y *sucesión regular* no estén importantemente vinculadas. La constatación de regularidades es *una* de las fuentes de evidencia de la existencia de relaciones causales; *una* de las razones por las que se juzga que existe una relación causal (Illari y Russo 2014). Dada la existencia de patrones no estrictos de regularidad juzgados causales en la práctica científica, sin embargo, resulta claro que la mera constatación de una regularidad no es una razón suficiente para afirmar la existencia de una relación causal. Hay algo más en el establecimiento de una relación causal que la constatación de una regularidad. ¿Existen principios que permitan discriminar los factores que justifican la atribución de causalidad a correlaciones entre variables?

5.3.2. Correlaciones invariantes, causalidad y explicación: la propuesta manipulacionista

De acuerdo con una tradición en la reflexión filosófica sobre la causalidad asociada con John Stuart Mill, los principios en que permiten discriminar los factores que justifican la atribución de causalidad a correlaciones entre variables pueden derivarse de las normas que subyacen a la práctica científica de realización de experimentos controlados.

En un experimento controlado, se alteran y manipulan variables con el propósito de explorar las circunstancias en las que la presencia o ausencia de un factor *hace una diferencia* con respecto a la ocurrencia o no de un efecto. Supuesta una hipótesis en el sentido de que una variable *X* es la causa de una variable *Y*, en un experimento se *manipulan* o alteran factores relativos a *X* y se observa qué ocurre con la variable *Y*. El propósito general es estudiar si algunas cosas son diferentes (los efectos) bajo un rango diferente de circunstancias en las que otras cosas (las causas) están o no presentes (Psillos 2007). De acuerdo con esta tradición, el establecimiento de relaciones causales y el razonamiento científico sobre causas está orientado no solo por la búsqueda de correlaciones robustas entre variables sino por el rastreo de *hacedores de diferencias* en esas correlaciones y la determinación de "estabilidad o invariancia a través de variaciones".

La propuesta mejor articulada en la literatura reciente sobre la metodología que orienta el establecimiento de relaciones causales en la práctica científica es el llamado "manipulacionismo" de James Woodward (Woodward y Hitchcock 2003). La idea general de esta propuesta es que una variable *C* (un evento, hecho o generalidad)

causa una variable *E* si en caso de que se "manipulara" o "interviniera" *C*, *E* cambiaría también. De acuerdo con Woodward, la metodología que subyace a la determinación de relaciones causales en la mayor parte de la práctica científica es la manipulación experimental de variables. El aspecto clave de la propuesta de Woodward es la idea de "estabilidad a través de intervenciones". Para determinar si una correlación o generalización empírica es causal es preciso examinar el grado de variación en las variables que introduce un conjunto de intervenciones. Considérense las correlaciones existentes entre: (a) fumar y desarrollar cáncer de pulmón; (b) fumar y tener los dedos amarillos; (c) tener los dedos amarillos y desarrollar cáncer de pulmón. Aunque la tercera correlación pueda ser estadísticamente robusta, no es el caso que exista una relación causal entre las dos variables en juego: no es el caso que tener los dedos amarillos esté causalmente relacionado con desarrollar cáncer de pulmón. ¿Qué diferencia esta correlación de las dos primeras? De acuerdo con la propuesta manipulacionista, una correlación es causal si es estable o invariante a través de un conjunto suficientemente amplio de intervenciones (Woodward y Hitchcock 2003). Una intervención es una manipulación que modifica una variable pretendidamente causal.

La noción de *intervención* en juego es una noción técnica que designa condiciones sobre manipulaciones posibles. En particular, de acuerdo con Woordward, una intervención *I* sobre una variable *X* debe satisfacer tres condiciones: (i) el cambio en el valor de *X* se debe únicamente a *I*; (ii) *I* afecta el valor de *Y* solo en virtud del cambio en el valor de *X*; (iii) *I* no está correlacionada con otras posibles causas de *Y* (Woodward y Hitchcock 2003, Illari y Russo 2014). La idea general es que si una variable *Y* –la variable "efecto"– es afectada, lo es únicamente a través de la acción de *X*, desencadenada por la intervención *I*. Cualquier otra influencia posible de *I* sobre *Y*, o de otros factores *Z* sobre *X* o *Y* quedan excluidos si el evento de cambio de *X* es una intervención. Las intervenciones son de esta manera medios para determinar si cambios en una variable-causa generan o generarían cambios en una variable-efecto, sin socavar la relación entre la causa y el efecto. Si se interviene en un grupo de población para alterar la variable "dedos amarillos" y si esta intervención no conlleva una alteración de la variable "fumar", no se verá ninguna incidencia en el desarrollo de cáncer de pulmón. Si se interviene en la variable "fumar", en cambio, y se satisfacen los criterios indicados, entonces se verá una alteración en la variable "desarrollo de cáncer de pulmón". En esto consiste el carácter *causal* de esta correlación.

Esta idea de "estabilidad bajo intervenciones", de acuerdo con Woordward, arroja luz también sobre la naturaleza de las explicaciones científicas. Una explicación aporta información sobre aquello de lo que *depende* la ocurrencia de un explanandum; esto es, sobre una relación de *dependencia* entre variables. En el caso de la ciencia, las relaciones de dependencia en cuestión son típicamente relaciones de

dependencia causal. En estos términos, explicar un explanandum *e* es de acuerdo con Woodward situarlo en una generalización empírica adecuada: asociarlo con una variable *c* con la que esté relacionado de manera estable bajo un rango amplio de intervenciones y que en este sentido *haga una diferencia* sobre su ocurrencia. Una explicación de *e* aporta información sobre la variable *c* con cuya presencia *e* coincide bajo un rango amplio de circunstancias posibles y de la que en consecuencia podemos inferir que *depende* causalmente. De acuerdo con esto, la determinación de correlaciones estables bajo intervenciones y la explicación científica de fenómenos son dos aspectos de la misma tarea (Woodward y Hitchcock 2003). La búsqueda de la explicación de un fenómeno motiva la búsqueda de correlaciones estables que cubran el fenómeno y a su vez el establecimiento de estas correlaciones es susceptible de empleo en la formulación de diferentes fenómenos.

5.3.3. Manipulacionismo y Newtonianismo

El manipulacionismo tiene la virtud de conectar el análisis filosófico de la noción de causalidad con la metodología y el uso de la noción en la práctica científica. En línea con una aproximación más empírica al examen filosófico de la ciencia, este se propone como una explicitación descriptivamente más adecuada del carácter de las relaciones causales y del uso de las mismas en la explicación científica. En este sentido, supone una enmienda a las deficiencias percibidas en el análisis positivista, con su énfasis en las ideas de *ley estricta*, *generalización universal* y *deducción*.

Desde una perspectiva más general, sin embargo, el manipulacionismo es una forma de Newtonianismo. Desde un punto de vista manipulacionista, la explicación y la causalidad siguen siendo nociones estrictamente epistémicas, subsidiarias de la detección y constatación de patrones de regularidad en la experiencia de la naturaleza. De acuerdo con Woodward, de hecho, su idea de "estabilidad a través de un rango de intervenciones" puede entenderse como una explicitación de la noción empirista de *ley* (Woodward y Hitchcock 2003). En la tradición newtoniana, el manipulacionismo enfatiza que las generalizaciones susceptibles de empleo en el establecimiento de relaciones causales y en la formulación de explicaciones no requieren ellas mismas de explicación en términos de procesos y propiedades causales más fundamentales. Una vez que una generalización se ha establecido de acuerdo con las directrices intervencionistas como una generalización causal, no es necesario ni oportuno preguntar a su vez por los factores en virtud de los cuales esta generalización tiene lugar.

El elemento común de todas las versiones del Newtonianismo es su renuencia a explicar o *fundar* las regularidades (juzgadas) causales. La provincia de teorización sobre la noción de causalidad se restringe a la explicitación de los factores a través de los cuales una regularidad puede juzgarse causal; al establecimiento de los criterios que rigen la *detección* de relaciones causales. Para detectar una relación causal entre

dos variables, no es preciso examinar los posibles procesos o factores físicos que median entre ambas variables. De la misma manera en la que Newton afirmaba que había establecido el carácter *causal* de la gravedad aun cuando desconociera los posibles procesos físicos que subyacían a este hecho, desde una perspectiva manipulacionista puede afirmarse que la determinación del carácter causal de la relación entre el consumo de tabaco y el desarrollo de cáncer de pulmón puede efectuarse al margen del establecimiento de los posibles factores físicos que median entre ambas variables. El vínculo entre causa y efecto puede de acuerdo con esto tratarse como una caja negra en la medida en que la relación entre causa y efecto pueda detectarse de manera fiable. El intento de explicitar los contenidos de esta caja negra –la naturaleza general de los factores que median entre las variables vinculadas por una generalización– supone de acuerdo con esta perspectiva el paso a un ámbito de especulación metafísica pretendidamente ajeno a las normas que rigen la teorización científica (Woodward 2017).

5.3.4. Capacidades y explicación causal

La inspección de las prácticas explicativas en diferentes sectores de la ciencia revela sin embargo que los científicos están concernidos con el contenido de esta caja negra (Illari y Russo 2014). Además de establecer experimentalmente la existencia de generalizaciones causales, uno de los objetivos de la ciencia es desentrañar los factores en virtud de los cuales estas generalizaciones son verdaderas. Esto ocurre en particular cuando, dado el conocimiento disponible sobre un sector de la realidad, no resulta claro *cómo* una variable dada puede ser causalmente responsable de algún efecto. Además de estar causalmente vinculado con el desarrollo de cáncer de pulmón, es un hecho establecido que el consumo de tabaco es también causa de diferentes enfermedades cardiacas. En el momento en el que empezó a acopiarse evidencia en este sentido, no era claro en virtud de qué tipo de procesos podía ser esto el caso; no se sabía *cómo* el consumo de tabaco podía afectar el sistema cardiaco. Este hecho representó una perplejidad teórica que espoleó el desarrollo de la investigación. En términos generales, cuando una correlación es suficientemente robusta pero el vínculo entre las variables que cubre es motivo de perplejidad dado el conocimiento disponible, un explanandum científicamente legítimo es, en contra de las directrices del Newtonianismo, la ocurrencia de la correlación en cuestión. Hasta que un explanans adecuado sea establecido, por otra parte, el estatus de la correlación para efectos de su uso teórico en la predicción y explicación es problemático. Si no se supiera cómo, en virtud de qué procesos biológicos, el consumo de tabaco puede afectar el sistema cardiaco, la creencia en el vínculo causal entre las dos variables carecería de justificación sólida y no podría excluirse la hipótesis de que la correlación entre ambas variables se debiera a una causa común diferente.

Este estado de cosas es común en el ámbito de las ciencias biológicas. Los fenómenos estudiados por estas disciplinas exhiben patrones de regularidad *locales* y restringidos, de una manera en que el mero establecimiento de generalizaciones causales carece de fuerza explicativa. En este ámbito, dicho brevemente, las generalizaciones causales son explananda más que explanantia.

Considérese el siguiente explanandum: "la metabolización de lactosa por parte de una bacteria E. Coli en un organismo". Esta expresión designa un tipo de evento o proceso situado espacio-temporalmente y, de acuerdo con la tradición newtoniana, en la medida en que sea un evento explicable científicamente, su explicación consistiría en su subsunción bajo una generalización. El explanans de este explanandum desarrollado por la biología molecular, sin embargo, no consiste en su subsunción bajo una generalización (Illari y Russo 2014). Un evento del tipo descrito participa ciertamente de correlaciones y el establecimiento de las mismas es un aspecto importante de su estudio científico. Por ejemplo, fue importante en la comprensión de este fenómeno reconocer que la ausencia de glucosa en el tracto gastrointestinal de un organismo es una variable relacionada con la metabolización de lactosa por parte de una bacteria E. Coli. El establecimiento de esta generalización, sin embargo, no es en sí misma explicativa sino más bien un explanandum que requiere de un explanans.

Robert Cummins ha argumentado con agudeza que esta situación es igualmente típica en la psicología y en general en el estudio científico de los fenómenos cognitivos. En la medida en que haya leyes en este dominio, estas son generalizaciones estadísticas que describen fenómenos cuya explicación es precisamente el principal objetivo teórico (Cummins 1983; 2000). La llamada "Ley de Stevens", por ejemplo, es una relación experimentalmente establecida entre dos variables: la magnitud de un estímulo físico y la intensidad o fuerza en la sensación creada por el estímulo. Esta ley permite hacer predicciones robustas relativas a la manera en que los seres humanos reportan la intensidad percibida de un rango de estímulos. Estas predicciones, sin embargo, no tienen ninguna fuerza explicativa. Aunque puede afirmarse que la medida en la que un sujeto grita de dolor al ser quemado es predicha y puede explicarse citando esta ley, este no es ciertamente un explanandum típico de las disciplinas que estudian la cognición. Los explanandum típicos en estos casos, al igual que en las ciencias biológicas, son más bien la ocurrencia de generalizaciones como la ley de Stevens o en general de los patrones de regularidad que caracterizan el comportamiento de los sistemas biológicos en diferentes circunstancias. ¿Por qué, en virtud de qué factores, la ley de Stevens es una generalización verdadera de la respuesta de ciertos organismos a ciertos estímulos? ¿Por qué el sistema digestivo de ciertos organismos metaboliza lactosa en un rango más o menos definido de circunstancias?

La respuesta adecuada a estas preguntas, las explicaciones que la ciencia ha propuesto para las mismas, como han enfatizado los críticos del modelo nomológico-deductivo y de la concepción newtoniana de la ciencia, no consiste en la subsunción de estas generalizaciones en otras generalizaciones sino en la especificación de los procesos causales *locales* que subyacen a la ocurrencia de estas generalizaciones. Los explananda típicos en este caso son *capacidades* y la explicación de una capacidad consiste en la especificación de la manera en que la capacidad se despliega en un sistema específico. Explicar por qué ocurre una capacidad es aportar información causal relativa a *cómo* "funciona" esa capacidad en un contexto específico.

Una capacidad puede identificarse con una *propiedad* de un sistema. En particular, una propiedad de la variedad que en el tercer capítulo denominé "disposicional". Una disposición es una propiedad de una entidad que no se manifiesta en cada momento en que la entidad existe sino solo en los momentos en los que se dan ciertas condiciones. Estas condiciones se denominan típicamente "condiciones de manifestación". Los ejemplos típicos de estas "potencialidades" son la *fragilidad* y la *elasticidad*. Un sistema es un tipo de entidad que puede reaccionar a un rango restringido de circunstancias comportándose de ciertas maneras o desencadenando ciertos procesos. Solo si en el tracto gastrointestinal de un organismo hay ausencia de glucosa, simplificando brutalmente las condiciones de manifestación en juego, el proceso de metabolización efectuado por la bacteria E. Coli se desencadena. El factor "interno" al sistema responsable de estas respuestas discriminadas a ciertos estímulos "externos" es una *capacidad*. La metabolización de lactosa por parte de la bacteria E. Coli es en este sentido una capacidad de un sistema, y de esta manera es conceptualizada por la biología molecular. Antes de que se descubrieran los procesos en virtud de los cuales esta bacteria efectuaba esta tarea, esta capacidad era un explanandum de la biología. En el caso del estudio científico de la cognición, explananda típicos son la capacidad de percibir, de aprender y usar una lengua, de recordar, razonar, entre otras. A diferencia de lo que ha ocurrido con la capacidad de metabolizar lactosa, los procesos en virtud de los cuales estas capacidades se despliegan en ciertos organismos son poco entendidos. Cubrir esta deficiencia, entender este tipo de capacidades es lo que motiva en primer lugar la investigación (Cummins 2000). El objetivo teórico central es también acá, como he dado por supuesto desde el primer capítulo, la explicación de capacidades.

Un rasgo característico de las capacidades, de acuerdo con lo dicho, es su *localidad*. Al afirmar que los patrones de regularidad característicos de los sistemas biológicos son locales lo que se afirma es que las generalizaciones que son verdaderas del comportamiento de estos sistemas tienen una aplicación restringida. El comportamiento característico de un tipo de célula o tejido puede y es en muchos casos diferentes al comportamiento de otro tipo de célula o tejido. Lo mismo puede ocurrir con respecto a diferentes sistemas o redes neuronales causalmente

responsables de capacidades cognitivas. La descripción de patrones de regularidad locales y de capacidades de sistemas es efectuada en la práctica científica no mediante *teorías* sino mediante *modelos*. Un modelo es una descripción interpretativa de un sistema que tiene la función de facilitar el acceso epistémico a los componentes y procesos que conforman el sistema (Bailer-Jones 2009). La explicación de una capacidad, en particular, es una representación a través de un modelo de los factores que subyacen o de los que *depende* causalmente la manifestación de la misma en un sistema definido. Estos factores son típicamente componentes y procesos internos de un sistema cuya representación tiene el potencial de especificar la manera *cómo* este hace lo que hace. Para los nuevos mecanicistas, tales componentes y procesos pueden conceptualizarse provechosamente en la descripción de la práctica científica mediante el uso de la vieja noción de *mecanismo*. Un modelo de un mecanismo es de acuerdo con esto el vehículo representacional por excelencia mediante el que la ciencia aporta información causal relevante explicativa de una capacidad.

5.4. El nuevo mecanicismo y la explicación constitutiva de capacidades

En el contexto de la descripción del comportamiento de diferentes sistemas, es habitual en las disciplinas biológicas el uso de la noción de *mecanismo*. Los patólogos hablan de los "mecanismos de una enfermedad", los farmacólogos de los "mecanismos de acción de una droga", los biólogos del desarrollo de "mecanismos genéticos", los biólogos moleculares de los "mecanismos fundamentales de la vida", etc. (Ioannidis y Psillos 2017). Los psicólogos y los teóricos de la cognición hablan también habitualmente de los mecanismos responsables de la visión, la memoria, el lenguaje el razonamiento, etc. En todos los casos, la noción de *mecanismo* cumple una función tanto descriptiva como explicativa. Los autores que dieron forma al nuevo mecanismo tomaron nota de este uso extendido de la vieja noción de mecanismo y construyeron alrededor de ella un análisis de las explicaciones científicas. Este análisis, en el espíritu de los estudios post-kuhnianos de la ciencia, tiene la pretensión de ser descriptivamente adecuado y de suplantar al menos en el ámbito de las ciencias no físicas el modelo nomológico-deductivo. De acuerdo con los nuevos mecanicistas, las explicaciones típicas en este ámbito son explicaciones mecánicas de un tipo especial: explicaciones composicionales o "constitutivas". Una explicación constitutiva, como indicaré, exhibe *cómo* ciertas entidades de interés científico tienen las capacidades causales que tienen apelando a la manera en que están constituidas por conjuntos organizados de componentes (Povich y Craver 2018; Ylikovski 2013).

5.4.1. Las explicaciones constitutivas como una forma de explicación "vertical"

Una explicación constitutiva es, en términos de una metáfora espacial empleada en la literatura reciente, un tipo de explicación "vertical" de un fenómeno natural. Estas

explicaciones dan cuenta de un explanandum citando factores en un "nivel inferior de organización" de los que este explanandum *depende*. Las explicaciones horizontales, del tipo sancionado por la tradición newtoniana, exhiben en cambio relaciones de dependencia entre un evento o una regularidad con respecto a otros eventos o regularidades que no están en ningún sentido teóricamente relevante en un nivel de organización diferente. Ejemplos de explicaciones verticales son la explicación de la herencia de "rasgos" entre organismos y sus descendientes recurriendo a la composición de estos rasgos en términos moleculares; la explicación del índice refractivo de un cristal usando las propiedades y relaciones entre los átomos que componen el cristal; o la explicación del movimiento de la superficie de la tierra recurriendo a las placas tectónicas y corrientes de magma que componen la tierra. Dado que en todos los casos el explanandum en cuestión se explica exhibiendo la manera en la que el mismo está "compuesto" o "constituido" por entidades, propiedades y procesos, se dice que estas explicaciones son "explicaciones constitutivas" y que las relaciones verticales entre diferentes fenómenos estudiados por la ciencia son relaciones de "constitución científica" (Aizawa y Gillett 2016). Las relaciones de dependencia entre explanandum y explanans en las explicaciones constitutivas es por tanto mereológica y conlleva un componente metafísico: el objetivo de las explicaciones verticales es determinar la estructura de la realidad en diferentes niveles y la manera en que diferentes fenómenos de interés científico están situados y pueden explicarse en virtud de esta estructura de niveles.

En la tradición newtoniana anti-mecanicista, los positivistas lógicos trataron de eludir el componente metafísico de las relaciones verticales al analizarlas en términos *deductivos* como relaciones de subsunción entre regularidades. Los "niveles" en los que se sitúan los fenómenos estudiados por disciplinas diferentes son niveles epistémicos. De acuerdo con este análisis, un explanandum *x* está en un nivel "inferior" o es más básico a un explanandum *y* si *y* puede inferirse deductivamente a partir de *x*. En la medida en que sea legítimo hablar de niveles no epistémicos, existe un único nivel de realidad, descrito por la física. Hablar de niveles diferentes en la realidad es un resultado de nuestras limitaciones epistémicas; del carácter parcial de nuestro conocimiento de la manera en que cada evento y regularidad es un evento físico. A pesar de las apariencias, así, todas las explicaciones científicas legítimas son de acuerdo con esto explicaciones horizontales.

Este modelo empirista y fisicalista de las relaciones verticales entre fenómenos como relaciones entre regularidades constitutivas entre teorías, como se indicó en el tercer capítulo, es descriptivamente inadecuado. Los explananda de las ciencias no básicas no pueden inferirse deductivamente de regularidades empíricas más "básicas". Dada esta inadecuación, la filosofía post-positivista ha explorado alternativas para dar cuenta de las relaciones verticales existentes entre diferentes explananda de interés científico y de las maneras en que los modelos propuestos para

explicar estos explananda se relacionan entre sí. La noción mecanicista de "explicación constitutiva" representa una alternativa en este sentido. Una, en particular, que abandona de manera resuelta el marco de referencia del Newtonianismo e incorpora una dimensión metafísica como un componente necesario de la descripción adecuada de la práctica científica.

5.4.2. Niveles de organización y mecanismos

Los explananda típicos de las explicaciones verticales son capacidades de sistemas que exhiben cierta complejidad interna: totalidades compuestas de partes en diferentes *niveles de organización*. Un nivel de organización, de acuerdo con el nuevo mecanicismo, está definido siempre en términos de un mecanismo explicativo de una capacidad (Craver 2007; Povich y Craver 2018).

Considérense ejemplos de sistemas con este tipo de complejidad interna: organismos vivos, sistemas nerviosos, cerebros, redes neuronales, neuronas, entre otros. Como totalidades, estas entidades tienen propiedades definidas y se relacionan con entidades numéricamente diferentes a ellas mismas: los organismos se comunican con otros organismos similares, los cerebros están encapsulados en cráneos, las neuronas se vinculan unas con otras a través de sinapsis. Las totalidades relacionadas de esta manera están en el *mismo nivel* de "realidad". Existen en cambio otro tipo de relaciones que ocurren entre entidades que son *parte* de otras entidades: las neuronas hacen parte de redes neuronales; estas redes hacen parte de cerebros; los cerebros hacen parte de sistemas nerviosos y estos de organismos. *Ser parte* de algo y *estar compuesto* de algo son relaciones que ocurren entre entidades en un *diferente nivel* de "realidad". Estas relaciones entre totalidades y partes son relaciones de *constitución* (Maley y Piccinini 2014).

Una totalidad en este sentido puede ser *parte* de una totalidad mayor. Un hipotálamo, por ejemplo, es una totalidad que hace parte de otra totalidad (el cerebro). Algunas totalidades exhiben una característica especial: sus partes están organizadas de tal manera que su interacción genera capacidades específicas. Estas totalidades son lo que los nuevos mecanicistas llaman "mecanismos". Los mecanismos se contrastan con meras "agregaciones" de entidades en la medida en que las partes de un mecanismo están organizadas a través de una serie de dimensiones. Las partes de un mecanismo tienen relaciones *espaciales*, *temporales* y *organizacionales* con las demás partes, en virtud de las cuales estas efectúan conjuntamente tareas (Povich y Craver 2018). Es en virtud de esto que un mecanismo se denomina un nivel de *organización* de la realidad.

Un mecanismo es en este sentido algo diferente no solo de cualquier agregación de entidades sino también de un sistema. Un único sistema es una totalidad que puede albergar diferentes mecanismos explicativos de diferentes de sus capacidades. Un ser humano, por ejemplo, es una sistema capaz de pensar, moverse, digerir

alimentos, respirar, reproducirse, etc. Estas capacidades son *explananda* de explicaciones constitutivas y sus explanans son mecanismos diferentes entendidos como sub-sistemas especiales del sistema. Una explicación de una capacidad específica consiste en consecuencia en la identificación y representación (a través de un modelo) del mecanismo responsable de la misma *en un tipo de sistema*: las partes de este implicadas en la *producción* de la capacidad y la manera en que las interacciones entre esas partes lo hacen. La formulación de explicaciones constitutivas requiere así de la *descomposición* tanto estructural como funcional de un sistema, y de la *localización* de las capacidades del sistema en los componentes que estructurales que lo conforman (Bechtel y Hamilton 2007).

5.4.3. Jerarquías de mecanismos y reduccionismo explicativo

Dado que un mecanismo es típicamente un sub-sistema de un sistema mayor, sus capacidades pueden explicarse también de manera mecánica. Como totalidades, los mecanismos exhiben capacidades que pueden explicarse verticalmente a través de sub-mecanismos. El explanans de una capacidad de un sistema en un nivel *n* es un mecanismo en un nivel *n*-1. Este mecanismo puede exhibir a su vez una capacidad cuyo explanans es un mecanismo en un nivel *n*-2 Esta característica, de acuerdo con los nuevos mecanicistas, es típica: los sistemas de interés científico exhiben una jerarquía de mecanismos incrustados en mecanismos, por lo que las explicaciones de sus capacidades deben ser explicaciones *multi-nivel* (Craver 2007).

El establecimiento del número de niveles que exhibe un sistema, al igual que la determinación del mecanismo explicativo de algunas de sus capacidades, es siempre una cuestión *empírica*, que debe resolverse caso por caso. La noción mecanicista de *nivel* es una noción *local*, *relativa* a un sistema definido objeto de interés científico. El punto de referencia de la *descomposición* de un sistema desde esta perspectiva es siempre la explicación científica de propiedades y capacidades del mismo. Un *nivel* está definido en este contexto por un *mecanismo* y un mecanismo es un componente explicativo de una capacidad de un sistema. De esta manera, no tiene sentido preguntar de dos entidades arbitrarias diferentes (como una herradura y un hipocampo) si están o no en el mismo nivel. Aunque esta explicitación de la idea de *nivel* resulta inadecuada para efectos de reflexiones metafísicas sobre la estructura *global* de la realidad y de la *reducción* de todos los explananda de interés científico a un conjunto restringido de explanans, su utilidad en la descripción de las prácticas explicativas de la ciencia es suficiente para otorgarle legitimidad desde la perspectiva del nuevo mecanicismo (Povich y Craver 2018; Walter y Eronen 2014).

En contraste además con un tipo de reduccionismo explicativo estricto de acuerdo con el cual debe haber un *único* nivel de realidad fundamental en términos de cuyas propiedades y regularidades podrían explicarse todos los fenómenos situados en niveles superiores, la concepción mecanicista enfatiza el carácter *local* de las

explicaciones constitutivas. Una capacidad de un sistema en un nivel *n* se explica a través de componentes organizados del sistema en un nivel adyacente *n*-1. Aunque las capacidades de estos componentes pueden explicarse en virtud de mecanismos en un nivel *n*-2, esta segunda explicación no conlleva necesariamente información explicativa relevante sobre la capacidad en el nivel *n*. Tratar de explicar una capacidad a través de un modelo mecánico que describe partes componentes de un sistema situadas en un nivel no adyacente (por ejemplo, tratar de explicar el ritmo cardiaco regular apelando a componentes a nivel atómico del sistema cardiaco) supondría un error. Su poder explicativo –la relevancia explicativa de la información causal aportada– sería en el mejor de los casos menor que el de un modelo mecánico descriptivo de componentes en un nivel adyacente (Franklin-Hall 2016). Evidencia de esto es la inexistencia de explicaciones de este tipo en la práctica científica. La inspección del tipo de explicaciones características de capacidades formuladas en las ciencias "especiales" revela, de acuerdo con los nuevos mecanicistas, la mayor adecuación descriptiva de su modelo con respecto al modelo reduccionista.

5.5. Reduccionismo y anti-reduccionismo: realizabilidad múltiple y la metafísica de las explicaciones

La diferencia introducida en el modelo positivista por el funcionalismo anti-reduccionista de los años sesenta se refirió al tipo de propiedades relevantes en la individuación de estas regularidades implicadas en la explicación nomológico-deductiva de los fenómenos de interés científico –al tipo de variables vinculadas por las mismas. Las propiedades en cuestión serían disposicionales o funcionales, deductivamente irreductibles a propiedades micro-físicas. La renuencia de los funcionalistas a fundar los poderes causales de las propiedades funcionales de un sistema –esto es, sus capacidades– en propiedades y procesos físicos resulta de la afirmación del carácter no *subsumible* de las regularidades en que intervienen estas propiedades con respecto a regularidades físicas más "básicas" (Fodor 1974; Ruphy 2016).

Las propiedades funcionales son propiedades de sistemas; esto es, poderes causales de totalidades. Las totalidades, como se indicó antes, pueden entrar en relaciones de *composición* con otras totalidades. Un sistema puede *ser parte* de otro sistema o *estar compuesto* por otros sistemas. Tanto una totalidad como sus partes tienen poderes causales (los cerebros tanto como las redes neuronales de las que están compuestos exhiben poderes causales característicos). La noción de *realización* designa una relación entre propiedades de totalidades y propiedades de las partes que las componen. Las propiedades de una totalidad (propiedades de nivel superior) son *realizadas* por propiedades de las partes de esa totalidad (propiedades de nivel inferior). El anti-reduccionismo consiste en la tesis según la cual las propiedades de

nivel superior son causalmente eficaces por derecho propio y *diferentes* a las propiedades de nivel inferior que las realizan. Las propiedades o estados mentales, por ejemplo, serían *diferentes* a las propiedades o estados neuronales. Una totalidad, en resumen, puede exhibir de acuerdo con esta tesis propiedades que no son *idénticas* a las propiedades de sus partes (Maley y Piccinini 2014).

¿Cómo puede sin embargo esta diferencia ser el caso? Supuesto que un sistema *S* está compuesto de las partes S_1, S_2 y S_3, ¿cómo puede *S* exhibir poderes causales que no sean también exhibidos por sus partes? En contra de formas radicales de holismo del tipo reseñado en el primer capítulo (como el vitalismo), el anti-reduccionismo funcionalista *no* afirma que las totalidades exhiban poderes causales no exhibidos por sus partes. La diferencia en cuestión resulta del hecho de que ciertas propiedades de nivel superior de un sistema pueden ser realizadas en diferentes circunstancias por diferentes propiedades de nivel inferior. Diferentes tipos de entidades con diferentes propiedades pueden realizar la propiedad funcional de nivel superior *ser un motor*, *ser un computador*, *ser un estado mental* o *ser un ala*. Esta circunstancia justifica de acuerdo con los funcionalistas la asignación de un estatus metafísico robusto a las propiedades funcionales, que conformarían un ámbito de la realidad distinto al ámbito de sus propiedades realizadoras.

Esta asignación cumple para los funcionalistas la tarea de asegurar la *autonomía explicativa* de las ciencias especiales que se ocupan de su estudio. En el modelo reduccionista, las ciencias especiales tenían un rol subsidiario y provisional: estas disciplinas estudian regularidades en la naturaleza que *aún* no han podido subsumirse e identificarse en términos de propiedades y regularidades físicas. El progreso de estas disciplinas consistiría en su gradual desaparición. Para los funcionalistas, así, la defensa de la diferencia *metafísica* de las propiedades funcionales representaba un medio de asegurar la independencia *teórica* de las disciplinas que las estudian. En la explicación de una capacidad funcional resultaría irrelevante así el acopio de información sobre las entidades y procesos físicos que realizan la capacidad en cada caso específico. Lo "estructural" y lo "funcional" conforman ámbitos de investigación separados y disjuntos. Los explanantia relevantes de explananda funcionales son factores funcionales (Fodor 1974; Elber-Dorozko y Shagrir 2019). Desde esta perspectiva, la heurística mecanicista de descomposición y localización no difiere del modelo positivista de reducción inter-teórica: en ambos casos se trata de programas de explicación de lo funcional en términos de lo estructural.

5.5.1. Realizabilidad múltiple y propiedades funcionales: un dilema

La defensa de la diferencia metafísica de las propiedades funcionales a través del uso de la noción de *realizabilidad múltiple* se enfrenta a un dilema agudamente formulado por Jaegwon Kim (Kim 1992). Considérese una propiedad *P* múltiplemente realizable

y sus propiedades realizadoras R_1, R_2,…R_n. Supuesto que un sistema que posea P no exhibe poderes causales no exhibidos por sus propiedades realizadoras (de acuerdo con el fisicalismo genérico aceptado por los funcionalistas), el partidario de la diferencia metafísica de las propiedades funcionales se enfrenta a un dilema: o bien P es meramente una disyunción de propiedades de nivel inferior $R_1 \vee R_2 \vee \ldots R_n$ o P es una propiedad genuina con poderes causales propios. Si P no es nada más que una disyunción de propiedades de nivel inferior, entonces no es una propiedad en un sentido relevante para efectos del trabajo científico. Esto es, en este caso P sería solo una manera de designar una multiplicidad de poderes causales diferentes que intervendrían en generalizaciones diferentes y explicarían fenómenos diferentes. La realizabilidad múltiple sería solo una relación entre una disyunción y uno de sus disyuntos. Por otra parte, si P es una propiedad genuina y exhibe poderes causales propios, entonces debe ser idéntica a alguno(s) de sus realizadores. Habría algo en los realizadores de P que sería idéntico a P, con lo que la realizabilidad múltiple no sería un fenómeno teórico genuino. En ambos casos, P no puede ser una propiedad metafísicamente diferente.

Este resultado, como no resulta difícil advertir, concuerda con el análisis mecanicista de las explicaciones científicas como explicaciones constitutivas. La explicación constitutiva de una capacidad consiste en la exhibición de la manera en que la misma resulta de entidades y procesos de nivel inferior. Así, parecería que el nuevo mecanicismo conlleva una recaída en una forma de reduccionismo. De todo lo dicho en las secciones precedentes, sin embargo, resulta claro que el nuevo mecanicismo no afirma el carácter dispensable ni provisional de las ciencias especiales. Como expondré en las próximas secciones, el componente metafísico de esta filosofía de la ciencia representa una posición intermedia en el debate post-positivista entre reduccionismo y anti-reduccionismo.

5.5.2. Nuevo mecanicismo, metafísica y unidad de la ciencia

El establecimiento de la posición intermedia representada por la metafísica mecanicista depende del rechazo de una falsa disyuntiva definitoria del debate entre reduccionismo y anti-reduccionismo: la disyuntiva según la cual las propiedades funcionales de orden superior de un sistema son o bien *idénticas* a sus propiedades realizadores de orden inferior o bien estrictamente *diferentes* a estas. De acuerdo con el nuevo mecanicismo, ninguna de las dos alternativas es correcta: las propiedades funcionales de nivel superior no son metafísicamente diferentes ni idénticas a las propiedades realizadoras de nivel inferior. El énfasis en esta disyuntiva revela desde la perspectiva de los nuevos mecanicistas un compromiso subyacente con la concepción newtoniana –horizontal– de las explicaciones científicas. Una vez que una de las dos opciones de la disyuntiva es aceptada, las explicaciones pueden proceder proponiendo como explanantia regularidades en las que participan un único tipo de

propiedades en un único nivel de organización. El funcionalismo sustituye el modelo positivista de un único nivel explicativo por un modelo con dos niveles entre los que no tienen lugar relaciones explicativas. En consecuencia, desde la perspectiva funcionalista, no hay explicaciones verticales. Esta posición, como se vio, es descriptivamente inadecuada con respecto a la práctica científica. Una vez que se acepta la existencia de explicaciones verticales, la disyuntiva se revela como falsa.

La opción mecanicista en este debate metafísico resulta de una reflexión sobre la naturaleza de las explicaciones constitutivas; en particular, sobre el tipo de explanantia postulados en estas explicaciones. En las explicaciones constitutivas, una propiedad funcional de un sistema se explica como el resultado de un subconjunto de los poderes causales de nivel inferior del sistema. La propiedad de nivel superior es *constituida* por la interacción de un conjunto restringido de componentes de un sistema. La opción mecanicista consiste en concebir las propiedades de nivel superior como *subconjuntos* o *partes propias* de los poderes causales de sus propiedades realizadoras de nivel inferior (Maley y Piccinini 2014).

Las propiedades funcionales representan formas de *individuar* sistemas, totalidades. Al atribuirle a *x* la propiedad de *ser un reloj*, se individúa a *x* en términos de cierto conjunto de tareas que puede desempeñar, al margen de otras muchas propiedades que *x* tiene y que, desde esta perspectiva, no son relevantes para su identificación. La entidad *x* podría individuarse a través de otras de sus propiedades, funcionales o no: como *ser de metal*, *ser inflamable*, *tener forma esférica*, etc. Aproximadamente cualquier combinación de propiedades puede emplearse para individuar una porción de la realidad como un *tipo* de objeto. Algunas de estas formas de individuación tienen interés científico, mientras que otras no. En cualquier caso, al individuar un objeto *x* de acuerdo con un conjunto de propiedades como un tipo *M* se establece un criterio con respecto a los tipos de modificaciones de las propiedades de *x* que afectan la identidad de *x* como un *M*. No cualquier modificación en las propiedades de *x* hace que este objeto deje de ser un ejemplar del tipo *reloj*. Cambiar su color, por ejemplo, o suprimir algunas de sus piezas no hacen que algo deje de ser un reloj. Al individuar un sistema como un tipo funcional, así, se identifica al sistema en términos de un subconjunto de sus propiedades: algunas de las propiedades del sistema se toman en cuenta mientras que se hace abstracción de otras (Maley y Piccinini 2014).

Los poderes causales característicos de los sistemas identificados como tipos funcionales son en este sentido siempre un subconjunto de los poderes causales de los sistemas en tanto totalidades físicas concebidas con sus partes y su organización interna. Desde el punto de vista mecanicista, al explicar un tipo funcional –los factores en virtud de los cuales un sistema exhibe una capacidad–, el explanans buscado es un *segmento* de las *partes* que componen el sistema como una totalidad física; esto es, un mecanismo. La tarea explicativa de las ciencias "especiales" consiste en la

determinación del segmento de las propiedades en nivel *n*-1 que *constituyen* (o, en la jerga funcionalista, realizan) una propiedad funcional.

Las propiedades realizadoras de una propiedad funcional individúan desde esta perspectiva un mecanismo explicativo de esa propiedad, con lo que, como se verá, la noción funcionalista de *realizabilidad múltiple* puede reinterpretarse perspicua y provechosamente a través de la noción de mecanismo.

5.5.2.1 Realizabilidad múltiple y mecanismos

En la literatura funcionalista clásica, la exposición y defensa de la realizabilidad múltiple, como expuse en el tercer capítulo, se da a través de la consideración de ejemplos intuitivos. *Motores*, *computadores*, *saca-corchos*, *ojos*, *alas*, *estados mentales*, entre otros, son ejemplos típicamente empleados para estos efectos. En cada uno de estos ejemplos, pretendidamente, se evidencian casos de "estructuras" diferentes posibles que "realizan" una misma "función" (Bechtel y Hamilton 2007). Qué cuentan como propiedades "realizadoras", propiedades "realizadas" y en consiste la relación de "realización" son sin embargo cuestiones que no recibieron tratamiento sistemático ni respuestas satisfactorias en esta literatura (Walter y Eronen 2014). Desde el punto de vista funcionalista, en particular, aproximadamente cualquier arreglo de propiedades micro-físicas de bajo nivel podían "realizar" un tipo funcional. Cualquier *partición* de la realidad *individuada* en términos no-funcionales podía fungir como la realización de un tipo funcional. Dado el compromiso tácito con una concepción "horizontal" de las explicaciones científicas, las relaciones entre los tipos funcionales y las porciones de la realidad que los realizan eran concebidos como *no explicativas*. Desde el punto de vista del nuevo mecanicismo, en cambio, estas relaciones son relaciones explicativas cuyo establecimiento es un objetivo teórico básico de las ciencias especiales.

Supuesta la corrección del análisis mecanicista de las explicaciones científicas, la concepción funcionalista clásica de realizabilidad múltiple no caracteriza de manera descriptivamente adecuada las formas en que los fenómenos funcionales pueden estar causalmente vinculados con diferentes configuraciones micro-físicas. La noción de realizabilidad múltiple adquiere mayor fuerza descriptiva con respecto a la práctica científica cuando sus *relata* se conciben como particiones de la realidad identificadas por las ciencias especiales en virtud de su valor explicativo. Estas particiones son, por supuesto, *mecanismos*. De acuerdo con la reinterpretación mecanicista, la existencia de realizabilidad múltiple depende de *mecanismos* en la naturaleza. Si una misma capacidad de dos sistemas se explica en virtud de dos mecanismos relevantemente diferentes, entonces los dos sistemas cuentan como diferentes realizaciones de la capacidad en cuestión. La realizabilidad múltiple es la relación que tiene lugar cuando hay tipos relevantemente diferentes de propiedades

de nivel inferior que *constituyen* la misma propiedad funcional de nivel superior (Maley y Piccinini 2014)[2].

Dado que un mecanismo es un conjunto de entidades que interactúan de maneras sistemáticas en virtud de su organización para generar una capacidad de un sistema, hay un grupo de dimensiones a través de las cuales puede identificarse un mecanismo como de un *tipo* diferente a otro mecanismo generador de la misma capacidad. Un mecanismo m_1 es un tipo diferente a un mecanismo m_2 generador de la misma capacidad si se cumplen algunas de las siguientes tres condiciones: (i) los componentes de m_1 son de un tipo diferente a los de m_2 pero están organizados de la misma manera; (ii) los componentes de m_1 son del mismo tipo a los de m_2 pero están organizados de maneras diferentes; (iii) los componentes de m_1 son de un tipo diferente y están organizados de maneras diferentes a los de m_2. Estas opciones designan tipos o maneras diferentes en las que puede ocurrir la realizabilidad múltiple. Estos tipos pueden denominarse RM_1, RM_2 y RM_3.

Como un ejemplo de RM_1 considérese el fenómeno de los ciclos circadianos. Un conjunto de organismos diversos exhiben ritmos circadianos y la organización de los sistemas responsables de generar las oscilaciones constitutivas de estos ritmos es la misma para la amplia mayoría de estos organismos. Las proteínas y los genes que participan en el proceso, sin embargo, son de un *tipo* diferente en diferentes organismos. Otro ejemplo de RM_1, mencionado antes, es el de las puertas lógicas. Una puerta lógica exhibe siempre la misma organización pero puede estar realizada por tipos de componentes diferentes.

Como un ejemplo simple de RM_2 considérese lo que puede hacerse con una superficie esférica y tres barras rígidas de igual longitud. Si las tres barras están dispuestas para sostener la superficie en tres puntos diferentes suficientemente lejanos del centro de la superficie esférica, el resultado es una entidad del tipo funcional *mesa*. De manera alternativa, dos de las barras pueden estar dispuestas para formar una cruz y la barra restante puede usarse para conectar el centro de la cruz con el centro de la superficie esférica. El resultado de esta diferente organización es aún una mesa. Un ejemplo menos simple de RM_2 es el de dos programas diferentes (almacenados en la misma computadora) para multiplicar enteros. Estos programas computan la misma función usando los mismos componentes de hardware. La organización temporal de la participación de estos componentes en la ejecución de ambos programas puede diferir de manera considerable. Así, los procesos generados por los dos programas cuentan como dos realizaciones diferentes de la operación de multiplicación.

Un ejemplo de MR_3 es el tipo funcional *ojo*. Existen muchas maneras diferentes en los que componentes de un sistema biológico pueden estar organizados para formar

[2] Las consideraciones presentadas en el resto de esta sección siguen de cerca la exposición de Piccinini y Maley 2014.

ojos y también muchos tipos de componentes organizados de esta manera. Otros ejemplos son *motores*, *saca-corchos* y *rifles*. Como con los tipos anteriores de realizabilidad múltiple, también los procesos de computación física exhiben MR_3: el mismo diseño computacional puede ser realizado por diferentes tecnologías, que a su vez pueden estar organizadas de maneras diferentes para efectuar las mismas computaciones.

En todos los casos posibles de realizabilidad múltiple, de acuerdo con lo anterior, los *relata* vinculados por esta relación son propiedades funcionales de un sistema y mecanismos *explicativos* de esas propiedades. Una propiedad funcional es, como se ha indicado repetidamente una *capacidad*: una propiedad en virtud de la cual un sistema responde de maneras discriminadas a un rango definido de estímulos bajo un conjunto específico de circunstancias. Un *motor* transforma energía de algún tipo en energía mecánica capaz de efectuar un trabajo; un *saca-corchos* remueve corchos de una botella bajo ciertas circunstancias; un *ojo* detecta ondas lumínicas y las convierte en impulsos electroquímicos; un *computador* produce ciertos *outputs* datos ciertos *inputs*. Todas estas capacidades son múltiplemente realizables en virtud de ser susceptibles de explicación constitutiva mediante mecanismos diferentes. Las capacidades computacionales, sin embargo, exhiben un tipo de realizabilidad múltiple especial.

Sacar corchos, producir energía mecánica capaz de efectuar un trabajo o convertir ondas lumínicas en impulsos electroquímicos son capacidades definidas en términos de entidades físicas específicas con propiedades físicas específicas. Los mecanismos que realizan y explican estas capacidades deben satisfacer restricciones relativas al tipo de propiedades físicas que sus componentes poseen. Un objeto que no tenga la *rigidez* adecuada para penetrar corchos, por ejemplo, no puede ser un saca-corchos. En contraste, las capacidades computacionales, como indiqué en la exposición de la concepción mecanicista de la computación, están definidas sin hacer referencia a ningún *medio físico* específico. Aunque un sistema computacional se define en términos de manipulaciones de vehículos físicos, estos vehículos no necesitan estar constituidos por ningún medio físico específico. Un sistema de computación digital, por ejemplo, manipula ristras de *dígitos*. Cualquier medio físico puede contar como una ristra de dígitos siempre que venga en el número adecuado de tipos, pueda concatenarse en ristras y exista una manera de que un mecanismo manipule estas ristras y produzca las relaciones correctas entre ellas. Las capacidades computacionales son en este sentido *independientes del medio*. La independencia del medio es una forma estricta de *realizabilidad múltiple*: no solo la capacidad de producir los *outputs* correctos a partir de los *inputs* correctos es múltiplemente realizable sino que también (a diferencia de las capacidades no computacionales) los *inputs y* los *outputs* son múltiplemente realizables.

5.5.2.2. Nuevo mecanicismo y ciencias especiales

De acuerdo con este marco de referencia mecanicista, las propiedades funcionales carecen de la independencia metafísica que el funcionalismo clásico les asignó como un medio para asegurar la legitimidad teórica y el carácter no dispensable de las ciencias especiales que las estudian. Explicar estas propiedades no consiste, en línea con el modelo nomológico-deductivo, en el establecimiento de regularidades funcionales con poder predictivo sino en la exhibición de cómo estas están en cada caso *constituidas* por mecanismos en un nivel inferior de realidad. Esta posición representa un tipo de reduccionismo restringido que se opone a la *autonomía explicativa* defendida por el funcionalismo. Supuesto que los explanantia de una capacidad funcional son mecanismos físicos, la corrección de un modelo explicativo de una capacidad de este tipo depende de la inclusión en el mismo de información relativa a los procesos físicos que *realizan* la capacidad en un sistema definido (Povich y Craver 2018).

Esto no supone sin embargo que la investigación de las propiedades funcionales presentes en la naturaleza no desempeñe un papel fundamental en la ciencia. La noción de propiedad funcional de nivel superior (y la noción correspondiente de realizabilidad múltiple) es una herramienta útil en la empresa científica por al menos tres razones. (a) Estas propiedades permiten *individuar* fenómenos de interés científico que podrían ser difíciles o imposibles de individuar en términos de propiedades de nivel inferior. (b) El uso de propiedades funcionales posibilita una taxonomía de sistemas que difieren en sus propiedades de nivel inferior pero que exhiben capacidades similares. (c) Esta forma de individuación y sistematización de la naturaleza, por último, hace posible buscar los explanantia correctos de diferentes fenómenos: los mecanismos explicativos de ciertas capacidades que, en ausencia de una individuación y sistematización funcional, sería también o difícil o imposible establecer. Esta es la razón por la que las ciencias "especiales" son necesarias y no dispensables: estas ciencias establecen los tipos de abstracciones que proporcionan las mejores explicaciones de las capacidades de nivel superior de muchos segmentos de la naturaleza, en circunstancias en las que rastrear detalles de nivel inferior puede ser imposible, menos informativo o ambas cosas (Maley y Piccinini 2014; Boone y Piccinini 2016).

5.5. La concepción mecanicista de la computación física

Los nuevos mecanicistas han desarrollado en las últimas décadas estudios de caso detallados en los que han puesto de manifiesto la utilidad descriptiva de las nociones de *mecanismo* y *explicación constitutiva* para dar cuenta de las prácticas teóricas en el ámbito de las ciencias biológicas (Bechtel 2013). Las explicaciones en biología, de acuerdo con esto, son en la mayoría de los casos explicaciones constitutivas multi-

nivel. En los últimos años, Carl Craver (2005; 2007) y William Bechtel (2009) han defendido la tesis de que las explicaciones en la neurociencia y en las ciencias cognitivas son también explicaciones constitutivas. Estos ejemplos inspiraron el proyecto de ampliar el alcance del nuevo mecanicismo también al ámbito de la ciencia de la computación.

De acuerdo con la concepción mecanicista de la computación, desarrollada y defendida principalmente por Gualtiero Piccinini y Marcin Milkowski, los procesos de computación física son la manifestación de capacidades de un tipo especial de sistemas físicos y su explicación consiste en la exhibición de mecanismos que subyacen a esas capacidades (Piccinini 2007a; 2008a; 2008b; 2010; 2018; Milkowski 2018b). Esta concepción, como intentaré mostrar, aporta las herramientas necesarias para resolver el problema de la computación física del que, de acuerdo con lo expuesto en el capítulo anterior, depende la posibilidad del empleo de las herramientas formales de la computación en la explicación de procesos físicos. Este es el problema de explicitar las condiciones bajo las cuales cabe atribuir a un sistema físico la capacidad de efectuar computaciones, en el sentido abstracto definido por la teoría matemática de la computación (Ritchie y Piccinini 2019). Mientras que en la tradición funcionalista se daba por supuesto que la noción de *mecanismo de computación* desarrollada por Turing permitía esclarecer el uso de la noción de *mecanismo* para efectos explicativos, la concepción mecanicista de la computación invierte este orden de dependencia. De acuerdo con esta concepción, es la noción de *mecanismo* desarrollada por el nuevo mecanicismo la que permite entender la intersección entre la noción abstracta de computación y los procesos de computación físicos.

La concepción mecanicista de la computación sostiene que los procesos de computación física son la manifestación de ciertas capacidades de un tipo especial de sistemas físicos: sistemas cuya *función* es procesar variables físicas de acuerdo con reglas definidas sobre esas variables. Un sistema computacional es un tipo de entidad que, al igual que los sistemas biológicos, posee complejidad interna y exhibe ciertas capacidades. Diferentes de estas capacidades pueden depender en el caso de sistemas complejos como los computadores programables de subsistemas en diferentes niveles de organización. Esta concepción de la computación física se denomina "mecanicista" en la medida en que concibe los procesos de computación como el resultado causal de la operación de mecanismos constitutivos de ciertos sistemas. Si un proceso físico *p* es un proceso computacional, lo es *en virtud* de mecanismos físicos que subyacen a la ocurrencia del proceso. Supuesto que todo proceso computacional tiene lugar en el contexto de un sistema físico, la explicación de estos procesos depende de la explicitación de los componentes del sistema que subyacen al ejercicio de sus capacidades de procesar variables físicas de acuerdo con reglas definidas sobre esas variables. Las computaciones efectuadas por componentes más básicos

son constituyentes de computaciones más complejas por mecanismos de los que son *parte*. De esta manera, las capacidades computacionales de un sistema como un computador son explicadas mecánicamente.

Lo que una respuesta mecanicista al problema de la computación física afirma es que un sistema físico "implementa" una computación cuando es un sistema constituido por componentes cuya función es manipular de maneras sistemáticas ciertas variables físicas. ¿Qué tipo de sistemas están constituidos de esta manera?

5.5.1. Computación digital

Los sistemas computacionales mejor conocidos son artefactos diseñados y construidos de manera deliberada sobre la base de los formalismos desarrollados en el seno de la teoría matemática de la computación. La forma clásica de esta teoría especifica procesos de computación denominados "digitales" (Piccinini 2008a). Estos procesos comprenden una amplia clase de formalismos, los más conocidos de los cuales son los formalismos que especifican el ámbito de las funciones recursivas y las máquinas de Turing (Sieg 2009). Son estos procesos los que, como he indicado en los capítulos previos, constituyen el trasfondo teórico de las diferentes formas de computacionalismo.

La computación digital puede definirse tanto en términos abstractos como en términos físicos concretos. Los procesos de computación digital abstractos consisten en la manipulación de ristras de elementos discretos; en particular, ristras de símbolos o letras de un alfabeto finito. Los procesos de computación digital concretos se definen sobre vehículos físicos que pueden concebirse como la *realización* de estos símbolos abstractos. Una manera usual de denominar a estos vehículos es "dígitos" (Piccinini 2012; 2018a). Un dígito es un estado macroscópico de un sistema físico (o de un componente de este) susceptible de ser identificado por el sistema como un *tipo* de estado y distinguido de otros estados macroscópicos de diferente *tipo*. Para asegurar la manipulación fiable de dígitos basada en el tipo de estos, un sistema físico debe manipular a lo sumo un número finito de tipos de dígitos. Los artefactos de computación digital más usuales, por ejemplo, contienen solo dos tipos de dígitos, usualmente referidos como "0" y "1". En artefactos de este tipo a cada dígito le corresponden un amplio número de estados microscópicos posibles. Por ejemplo, a diferentes conjuntos de distintas disposiciones de electrones (estados microscópicos) les corresponde la misma carga almacenada en un condensador eléctrico (*capacitor*). Los artefactos de computación digital están diseñados para tratar todos estos estados microscópicos de una misma manera; esto es, la manera que corresponde a su tipo de dígito macroscópico (Piccinini y Bahar 2013). Los dígitos así entendidos son entidades *individuadas* en términos de propiedades relativas a su posible manipulación por componentes adecuados de un sistema físico y no, como en el

funcionalismo computacional de Fodor, en términos de ninguna propiedad semántica que pueda atribuírseles.

Los dígitos pueden concatenarse para formar secuencias. Son estas secuencias las que conforman en sentido estricto los vehículos de las computaciones digitales concretas. Un proceso físico de computación digital consiste en el procesamiento de secuencias de dígitos de acuerdo con una "regla". Una regla en este sentido es un tipo mapeo de ciertas secuencias de dígitos (*inputs*) en cierta secuencia diferente de dígitos (*outputs*). Ejemplos de reglas en este sentido son la adición, la multiplicación y la clasificación (*sorting*). Aunque estas reglas pueden definirse en términos abstractos, en un proceso de computación física la manipulación de dígitos es efectuada por componentes concretos de un sistema de manera que los *inputs* del proceso se convierten en los *outputs* adecuados de acuerdo con la regla en virtud de factores físicos. En el caso de los artefactos computacionales electrónicos, los factores en cuestión son factores relativos a cargas eléctricas manipuladas por ciertos componentes concretos de estos artefactos. Las reglas de computación física, en consecuencia, pueden entenderse como maneras de describir la forma en que ciertos componentes de un sistema físico reciben, transforman y transmiten tipos de dígitos (Piccinini 2008a; Piccinini 2018a).

Los sistemas computacionales digitales concretos se individúan así en virtud de su capacidad de manipular ciertos vehículos físicos de maneras sistemáticas. Los vehículos en cuestión, sin embargo, no se individúan en virtud de ninguna propiedad física específica. Los efectos físicos necesarios para un proceso de computación concreto se individúan en términos de relaciones entre vehículos físicos cualesquiera siempre que estos posean ciertas dimensiones de variación o "grados de libertad". Cualquier medio físico puede contar como una secuencia de dígitos en la medida en que venga en el número correcto de *tipos*, haya una manera de concatenarlos en secuencias y un sistema pueda manipular estas secuencias y producir el tipo correcto de relaciones entre las mismas. Considérese el caso de los sistemas de computación digital concretos más elementales: las puertas lógicas. Una puerta lógica NO, como se indicó en el segundo capítulo, es un dispositivo que recibe como *input* un dígito y produce como *output* un dígito del tipo "opuesto". En la medida en que un sistema tenga dos *inputs* posibles, pueda producir dos *outputs* posibles y de hecho genere de manera automática el *output* del tipo opuesto al *input* recibido en cada caso específico, cuenta como una puerta lógica NO. No importa si los vehículos del proceso son eléctricos, mecánicos, acústicos, o de algún otro tipo. En este sentido, se dice que los vehículos de los procesos de computación digital concretos son "independientes del medio" (Piccinini y Maley 2014). La "independencia de medio", como indiqué previamente, es una forma de lo que los funcionalistas llaman "realizabilidad múltiple". La realizabilidad múltiple de los vehículos de las computaciones digitales, sin embargo, tal es el punto crucial, no supone que cualquier sistema físico, con

prescindencia de su composición y de las estructuras físicas concretas que lo constituyen, pueda ser un sistema computacional.

Los computadores digitales programables, el tipo de artefactos cuyas capacidades inspiraron la analogía entre mente y máquina en la tradición funcionalista, son sistemas de computación digital del tipo descrito que, en virtud de la complejidad de su composición interna, efectúan sofisticadas tareas de procesamiento de dígitos. Estos artefactos tienen dos tipos de componentes especiales: las llamadas "unidades de memoria" y las "unidades procesamiento". Las unidades de procesamiento son componentes que tienen la función de ejecutar operaciones primitivas de cómputo sobre cadenas de dígitos denominadas "datos" de acuerdo con instrucciones definidas. Las unidades de memoria, por su parte, son componentes que almacenan datos e instrucciones de computación. Las instrucciones en cuestión no son más que secuencias de dígitos que, en virtud de la manera en que los artefactos en cuestión están diseñados, tienen el poder de *causar* la ejecución de operaciones de computación por parte de unidades de procesamiento. Una lista de instrucciones constituye un *programa*; con lo que la "ejecución de un programa" es la efectuación de una lista de instrucciones en un orden definido.

De esta manera, la *capacidad* de ejecutar programas característica de los computadores digitales programables se *explica* en virtud de sus *componentes* y las *interacciones* organizadas entre los mismos. A diferencia de otros sistemas de computación digital, como las calculadoras –para no hablar ya de sistemas no computacionales–, solo los computadores tienen unidades de procesamiento capaces de ejecutar programas y unidades de memoria capaces de almacenar estos programas. Este es el factor en virtud del cual las capacidades de estos artefactos pueden *explicarse* en términos de la ejecución de programas. Cuando la conducta de un computador se explica de esta manera, el programa al que se apela en la explicación no es solo una descripción abstracta. El programa es un (estado estable de) un *componente físico* del artefacto, cuya función es generar alguna de sus capacidades (Piccinini 2010).

En resumen, un sistema físico es un sistema de computación digital si y solo si es un sistema cuyos componentes están organizados para manipular secuencias de tipos de dígitos de acuerdo con una regla definida sobre las secuencias y sobre posibles estados internos del sistema.

5.5.2. Computación analógica

Los procesos de computación digital representan con mucho nuestra mejor comprensión de la idea de computación y constituyen el tipo de computación implementado en casi la totalidad de los sistemas físicos que efectúan computaciones. Existen sin embargo procesos matemáticos definidos sobre vehículos no discretos

que no cabe caracterizar como símbolos y sistemas que implementan estos procesos. Tales sistemas son conocidos como "computadores analógicos" (Fresco 2014).

Los computadores analógicos son artefactos físicos empleados principalmente para resolver sistemas de ecuaciones diferenciales. Estas ecuaciones están definidas sobre variables reales continuas, que pueden tomar cualquier número real como un valor. Para implementar variables de este tipo, los computadores analógicos deben manipular variables físicas que puedan representar o tomar cualquier valor real, dentro de ciertos intervalos. A diferencia de los dígitos, este tipo de variables físicas solo pueden medirse y manipularse en un grado finito de aproximación, con la consecuencia de que los computadores analógicos producen *outputs* que representan solo aproximaciones de los resultados deseados. Mientras que los sistemas de computación digital pueden siempre distinguir de manera inequívoca los tipos de sus vehículos, los sistemas de computación analógica no pueden hacerlo. Siempre es posible que la diferencia entre dos porciones de una variable continua sea suficientemente pequeña para que el sistema no pueda detectarla. Los vehículos de los procesos de "computación analógica" no son así secuencias de dígitos, por lo que estos procesos no cuentan como computaciones digitales (Piccinini 2012).

A pesar de esto, existen similitudes suficientes entre ambos tipos de procesos para incluirlos dentro de una categoría más amplia de computación física. En particular, tanto los sistemas de computación digital como los sistemas de computación analógica se emplean para resolver problemas matemáticos y ambos lo hacen en virtud del "seguimiento" de un tipo de procedimiento o regla definida sobre las variables físicas que manipulan. De manera crucial, los vehículos de los sistemas de computación analógica, al igual que los vehículos de las computaciones digitales, son "independientes del medio". Estos sistemas son sensibles a diferencias entre sus vehículos solo en dimensiones específicas de variación, por lo que un proceso de computación analógica puede implementarse en múltiples medios físicos (mecánicos, electrónicos, magnéticos) siempre que el medio posea un número suficiente de dimensiones de variación o "grados de libertad" a los que el sistema pueda acceder y manipular (Piccinini 2018a).

5.5.3. Computación genérica y la concepción mecanicista de la explicación de las capacidades computacionales

El resultado de conjugar la especificación de la naturaleza de los procesos de computación digital y de los procesos de computación analógica es la articulación de una concepción *genérica* de computación física, que caracteriza todos los procesos de computación física conocidos y concebibles hasta el momento. Un sistema de computación física es un sistema cuya función es manipular variables independientes del medio de acuerdo con reglas definidas sobre esas variables (Piccinini 2018a).

Esta concepción genérica de computación delimita de manera precisa las propiedades distintivas de los sistemas de computación físicos. Esta delimitación, como indiqué en el capítulo anterior, resultaba imposible dentro del marco de referencia del funcionalismo, cuyo compromiso con un anti-reduccionismo definido en términos del modelo positivista de reducción inter-teórica conducía a un tipo de nihilismo computacional de acuerdo el cual cualquier sistema físico podía categorizarse como un sistema computacional. Aun cuando muchos sistemas físicos sean susceptibles de descripción mediante modelos computacionales, la mayoría no efectúa computaciones en el sentido explicitado en las páginas previas. El nihilismo computacional se revela como una posición infundada una vez que las condiciones en virtud de las cuales un sistema físico efectúa computaciones se explicitan adecuadamente (Piccinini 2008a). Todos los sistemas físicos conocidos que efectúan computaciones y solo ellos caen bajo la caracterización presentada.

La *explicación* de las capacidades que exhibe un sistema físico de efectuar computaciones consiste en la descripción de la manera en que sus componentes en diferentes niveles de organización manipulan ciertos vehículos físicos de maneras discriminadas, de acuerdo con reglas algorítmicas. Una explicación de este tipo es una explicación constitutiva, en el mismo sentido en el que lo son las explicaciones de las capacidades no-computacionales de los sistemas biológicos. La concepción mecanicista de la computación aporta así un tratamiento unificado de las explicaciones de capacidades como explicaciones mecánicas. En contra de lo pretendido por los funcionalistas, en el contexto de la explicación de fenómenos naturales es la noción general de mecanismo la que permite dar cuenta de la noción de computación.

La consecuencia crucial de todo lo anterior es que si la cognición en su conjunto o alguna capacidad cognitiva en particular puede *explicarse*, como algo opuesto a modelarse abstractamente en términos computacionales, entonces el cerebro –o algún sistema físico que exhiba capacidades del tipo en cuestión– debe tener el tipo correcto de organización y estructura física para implementar un proceso computacional. Si esto es así o no es una cuestión empírica. Si el cerebro es un sistema conexionista no programable –como pretendían tanto McCulloch y Pitt como los conexionistas–, un sistema de computación digital programable –como pretendían el paradigma clásico en ciencias cognitivas–, o bien un sistema de computación analógico, es algo que solo puede establecerse en último término por medios empíricos, estudiando las estructuras y procesos físicos que subyacen al ejercicio de las capacidades cognitivas en sistemas complejos como el cerebro.

CONCLUSIÓN

El escrutinio crítico efectuado por la filosofía de la ciencia reciente del uso de la noción de *causalidad* y de la naturaleza de las explicaciones causales a la luz de la práctica científica (reseñado parcialmente en el último capítulo) puso de manifiesto la inadecuación descriptiva de imagen newtoniana del conocimiento científico basada en las categorías de *ley* y *teoría*. Esta imagen representó el presupuesto tácito en el que se desarrolló la discusión entre las concepciones reduccionistas y anti-reduccionistas de la ciencia a partir de los años sesenta del siglo pasado. Dada la inadecuación de este presupuesto no resulta sorprendente que ninguna de las dos posiciones en el debate represente una concepción general adecuada de la ciencia. El ideal positivista y newtoniano de unidad de la ciencia como un sistema deductivo en el que a partir de unos pocos principios fundamentales podría deducirse todo el conocimiento del mundo natural supone un artefacto filosófico sin respaldo empírico en la práctica científica. Los diferentes sectores de la empresa científica no están relacionados de conformidad con este modelo y una historia causal completa del mundo natural no puede contarse, como se seguiría de la corrección del mismo, con los principios y estructuras disponibles en un nivel micro-físico "fundamental". Tales fueron las conclusiones razonablemente establecidas por el anti-reduccionismo. La inferencia a partir de estos resultados de un pluralismo general conforme al cual la ciencia estaría bifurcada en la investigación de propiedades y procesos en dos niveles metafísicamente diferentes, con resultados *explicativamente autónomos*, se reveló sin embargo igualmente infundada.

El nuevo mecanicismo aporta una imagen general de la ciencia alternativa a la concepción newtoniana y descriptivamente más adecuada con respecto a la práctica científica. Esta imagen general, como expuse en el capítulo anterior, representa una posición intermedia en el debate entre reduccionismo y el anti-reduccionismo.

En contra de lo pretendido por el anti-reduccionismo funcionalista, los diferentes sectores de la ciencia no conforman islas teóricas. La descripción y explicación de los diferentes fenómenos constitutivos del mundo natural exige en muchas ocasiones la colaboración de diferentes disciplinas científicas. El trabajo interdisciplinario es una marca distintiva de la ciencia. La complejidad causal del mundo y las interconexiones evidenciales entre diferentes procesos naturales no se atiene a fronteras disciplinarias establecidas en muchos casos de acuerdo con factores sociales e institucionales contingentes. Un resultado en una disciplina *x* puede socavar o respaldar un resultado en la disciplina *y*, aunque inicialmente los fenómenos estudiados por cada una parezcan distantes. Un hallazgo en biología molecular, por ejemplo, puede obligar a reconsiderar una tesis en la neurología de la visión. Las diferentes disciplinas científicas se proponen desentrañar patrones y relaciones

causales en el mundo y en el proceso de hacerlo no pueden sino recurrir al conocimiento acopiado en disciplinas vecinas. Es en virtud de esta colaboración, encaminada a establecer implicaciones evidenciales de cara a la complejidad causal del mundo, que la ciencia comprende una empresa unificada (Potochnik 2010).

En la descripción de estas formas de colaboración las categorías meta-teóricas desarrolladas por la nueva filosofía mecanicista se revelan más útiles que las categorías de la tradición newtoniana. Considérense unos pocos ejemplos: (a) la físico-química estudia la estructura de las moléculas mientras que la bioquímica se ocupa de estudiar sus funciones; (b) la citología aporta información sobre la localización física en las células de las estructuras genéticas postuladas por la genética; (c) la bioquímica estudia los procesos e interacciones que constituyen los patrones hereditarios de expresión génica postulados por los genetistas; (d) la neurolingüística se ocupa de los procesos neuronales que constituyen las capacidades descritas por la lingüística teórica (Bechtel y Hamilton 2007). Estas formas típicas de colaboración se corresponden con relaciones entre capacidades descritas en un nivel de organización y (la localización) de mecanismos explicativos de las mismas en niveles inferiores más que con la deducción de tipos de eventos y regularidades a partir de leyes y principios físicos. Las nociones de *mecanismo*, *descomposición estructural* y *funcional*, y *localización* son así herramientas descriptivas útiles para exhibir la manera en que la ciencia representa una empresa *integrada* de generación de conocimiento.

Los mecanismos postulados en las explicaciones exitosas de una disciplina pueden representar de esta manera tanto explanantia como explananda de otras disciplinas. El ideal de la investigación científica es el desarrollo de conocimiento integrado en la forma de lo que los nuevos mecanicistas denominan "explicaciones constitutivas multi-nivel" de las propiedades y capacidades características de diferentes sectores del mundo natural (Craver 2007). La formulación de explicaciones multi-nivel representa una medida de la integración de un conjunto de disciplinas. El grado de integración de los resultados de una disciplina incipiente con respecto a los resultados de otras disciplinas mejor establecidas, por otra parte, supone en este marco de referencia una medida de su progreso. Cuando los mecanismos genéticos y bioquímicos responsables de ciertos procesos y mecanismos postulados por la biología evolutiva fueron establecidos, por ejemplo, el conocimiento acopiado por esta disciplina pudo integrarse con el de otras disciplinas más desarrolladas, como resultado de lo cual experimentó un progreso como campo de estudio.

El potencial de integración de las hipótesis explicativas distintivas de una disciplina o grupo de disciplinas está íntimamente vinculado con lo que en los dos primeros capítulos denominé la "mecanización" de un ámbito de estudio. El progreso de la biología a partir del siglo XIX, como expuse en esos capítulos, tuvo lugar a través de la gradual mecanización de diferentes capacidades biológicas como la fecundación,

la comunicación entre células nerviosas, la transmisión de información genética y la síntesis de proteínas. El computacionalismo, entendido como la hipótesis según la cual las capacidades cognitivas del cerebro resultan de mecanismos de computación en algún nivel de organización del sistema nervioso, supuso en este mismo sentido un esfuerzo de mecanización de un segmento de la realidad natural.

La convicción de que el cerebro *debe ser* un sistema computacional se deriva del hecho de que hasta el momento los formalismos matemáticos desarrollados por la teoría de la computación constituyen la única idea razonable de concebir mecanismos que mediante medios finitos produzcan sistemáticamente *outputs* de una variedad potencialmente ilimitada. Dado que el cerebro es el asiento causal de las capacidades cognitivas de los organismos inteligentes y que estas capacidades, a diferencia del resto de capacidades de los organismos vivos, son capacidades de propósito general y, al menos en el caso de los seres humanos, de potencial "universalidad" (en el sentido definido por Descartes y discutido en el primer capítulo), la mejor apuesta científica hasta el momento es conjeturar que el cerebro debe realizar sus tareas de alguna manera mediante procesos de computación física (Marcus 2014). Esta apuesta, sin embargo puede y ha sido interpretada de diferentes maneras.

El resultado alcanzado al final del último capítulo es que el computacionalismo, entendido como una hipótesis empírica y explicativa de ciertas capacidades biológicas, debe interpretarse en términos mecanicistas. El computacionalismo sobre la cognición, en resumen, es una hipótesis mecanicista. Los modelos computacionales propuestos por los partidarios de esta hipótesis son descripciones de mecanismos posibles que aspiran a representar los mecanismos que subyacen a los procesos físicos constitutivos de la cognición en el funcionamiento del cerebro. Las explicaciones computacionales son por tanto una forma de explicación mecánica y en este sentido pueden *integrarse* con otro tipo de explicaciones mecánicas desarrolladas en niveles inferiores de organización neurobiológica: niveles celulares, moleculares, bioquímicos, electrofisiológicos, etc. (Serban 2015).

La ciencia cognitiva se desarrolló sin embargo alrededor de una interpretación funcionalista no-mecanicista del computacionalismo. De acuerdo con esta interpretación, la descripción computacional de un proceso o sistema representa una descripción del mismo en un nivel funcional único, irreductible al nivel micro-físico, y la explicación computacional de las capacidades cognitivas es en consecuencia un tipo de explicación cualitativamente diferente y autónomo al tipo de explicación de las estructuras neuronales. Dada la incorrección de sus supuestos teóricos fundacionales, como se discutió con detalle en el capítulo 4, no resulta sorprendente que el ímpetu de esta iniciativa interdisciplinar se haya agotado y que la idea misma de una ciencia cognitiva unificada, como fue concebida en los años sesenta, se haya en buena medida disuelto en las últimas décadas (Núñez et al. 2019). Las diferentes disciplinas que conformaban esta iniciativa han seguido cursos independientes y no convergentes. La

aspiración de articular una ciencia unificada de la cognición alrededor de una interpretación funcionalista de la hipótesis computacional se reveló tan infundada como prematura.

La interpretación mecanicista del computacionalismo representa un marco de referencia descriptiva y teóricamente más adecuado para una ciencia de la cognición que el funcionalismo anti-reduccionista anclado en la concepción newtoniana de la ciencia. Tal es la tesis que me propuse defender en los capítulos previos. En los dos primeros capítulos presenté la concepción mecanicista de la ciencia y de las explicaciones científicas desarrollada a partir del siglo XVII como el contexto teórico en el que tuvo lugar el surgimiento del computacionalismo sobre la cognición. La falta de una explicitación meta-teórica clara de los conceptos de *causalidad, explicación* y *reducción* empleados por los mecanicistas motivó, como expuse en el tercer capítulo, el desarrollo de una filosofía de la ciencia anti-mecanicista en términos de la cual se interpretó el computacionalismo a partir de la segunda mitad del siglo pasado. La inadecuación de este marco de referencia fue expuesta el objeto del capítulo cuatro. La interpretación funcionalista del computacionalismo, según indiqué, conduce a un tipo de nihilismo computacional conforme al cual el uso de modelos computacionales en la investigación de la cognición carece en último término de fuerza explicativa. En el último capítulo expuse el resurgimiento de la filosofía mecanicista en el contexto de las discusiones post-positivas sobre la naturaleza de las explicaciones causales características de las ciencias "especiales". El trabajo realizado por los nuevos mecanicistas aporta las explicitaciones meta-teóricas ausentes en el mecanicismo clásico y ofrece un análisis de las explicaciones científicas en términos del cual resulta posible eludir las dificultades reseñadas del funcionalismo y justificar el potencial explicativo del uso de modelos computacionales en la investigación de las capacidades cognitivas.

El desarrollo de una ciencia de la cognición es sin embargo aún hoy, como en el tiempo de Descartes, una asignatura pendiente y dado el estado actual de nuestro conocimiento no resulta claro qué forma podría tomar esta ciencia del futuro. Todo lo que categorizamos en términos funcionales como capacidades cognitivas es sin duda el resultado de mecanismos físicos en algún nivel de organización neurobiológico. Estos mecanismos son muy probablemente también mecanismos que efectúan tareas de computación. El computacionalismo entraña en este sentido un programa de representación de mecanismos biológicos y una ciencia de la cognición fundada en el mismo debe entenderse, en contra de las pretensiones de "autonomía" de los funcionalistas, como un capítulo de la biología y de la ciencia natural en su conjunto. La tarea de desentrañar el nivel o los niveles correctos de organización en el que operan estos mecanismos y el tipo exacto de tareas de computación que efectúan apenas empieza sin embargo. El conocimiento de la manera en que el cerebro despliega sus capacidades cognitivas es, dicho en otros términos, aún fragmentario. El

trabajo realizado en las últimas décadas en las neurociencias y en algunas de las disciplinas agrupadas en la ciencia cognitiva clásica (como la lingüística) ha permitido acopiar información valiosa en este sentido. Conocemos la identidad de muchas de las moléculas que componen las neuronas, algunos de los mecanismos mediante los que estas se comunican con otras neuronas y algunas de las funciones que las estructuras neuronales así conformadas efectúan. Sabemos también de la existencia de estructuras neuroanatómicas repetidas a través del neocórtex y de correlaciones entre estas estructuras y la ejecución de diferentes capacidades cognitivas. Lo que ignoramos aún es cómo estas piezas encajan en un modelo general de la manera en que el cerebro genera las diferentes capacidades cognitivas características de la conducta inteligente (Marcus y Freeman 2014). Si los mecanismos que subyacen a diferentes de estas capacidades son del mismo tipo es por otra parte también una cuestión abierta. Nada en el estado actual de nuestro conocimiento excluye que los mecanismos explicativos del lenguaje sean por ejemplo muy diferentes a los mecanismos explicativos de la memoria o la percepción, de manera que las tareas de explicación de estas capacidades sean después de todo muy diferentes y la ciencia de la cognición una empresa teórica diferente a lo que hemos imaginado.

Referencias

Abraham, T. (2019) Cibernetics. En M. Sprevak y M. Colombo (Eds.), *The Routledge Handbook of the Computational Mind* (pp. 52-64). Londres: Routledge.

Aizawa, K. y Gillett, C. (2016). Vertical Relations in Science, Philosophy, and the World: Understanding the New Debates over Verticality. En K. Aizawa y C. Gillett (Eds.), *Scientific Composition and Metaphysical Ground* (pp. 1-38). Londres: Palgrave Macmillan UK.

Akagi, M. (2017). Rethinking the problem of cognition. *Synthese*, 195(8), 3547-3570.

Allen, G. (2005). Mechanism, vitalism and organicism in late nineteenth and twentieth-century biology: the importance of historical context. *Studies In History And Philosophy Of Science Part C*, *36*(2), 261-283.

Allen, G. (2018). Mechanism, Organicism, and Vitalism. En S. Glennan y P. Illari (Eds.), *The Routledge Handbook of Mechanisms and Mechanical Philosophy* (pp. 59-73). Londres: Routledge.

Bailer-Jones, D. (2009). *Scientific Models in Philosophy of Science*. Pittsburgh: University of Pittsburgh Press.

Bechtel, W. (2009). Constructing a Philosophy of Cognitive Science. *Topics in Cognitive Science*, 1, 548-569.

Bechtel, W. (2013). Understanding Biological Mechanisms: Using Illustrations from Circadian Rhythm Research. En K. Kampourakis (Ed.) *The Philosophy of Biology*: *A Companion for Educators* (pp. 487-510). Dordrecth: Springer.

Bechtel, W., Abrahamsen, A., y Graham, G. (1998). The life of cognitive science. En W. Bechtel y G. Graham (Eds.), *A companion to cognitive science* (pp. 1-104). Oxford: Blackwell.

Betchel, W. y Hamilton, A. (2007). Reduction, Integration, and the Unity of Science: Natural, Behavioral, and Social Sciences and the Humanities. En T. Kuipers (Ed.), *General philosophy of science: Focal Issues* (pp. 377-430). Amsterdam: North Holland.

Betchel, W. y Herschbach, M. (2010). Philosophy of the Cognitive Sciences. En F. Allhoff (Ed.), *Philosophies of the Sciences: A Guide* (pp. 239-261). Malden, MA: Wiley-Blackwell.

Bermúdez, J. (2014). *Cognitive Science. An Introduction to the Science of the Mind*. New York: Cambridge University Press.

Block, N. (1978). Troubles with Functionalism. *Minnesota Studies in the Philosophy of Science,* 9, 261-325.

Block, N. (1995). The Mind as the Software of the Brain. En: D. Osherson, L. Gleitman, S. Kosslyn, E. Smith y S. Sternberg (Eds.), *An Invitation to Cognitive Science*, *Vol.* 3, (pp. 377-425). Cambridge, MA: MIT Press.

Boden, M. A. (2006). *Mind as Machine: A History of Cognitive Science*. New York: Oxford University Press.

Boone, W., y Piccinini, G. (2015). The Cognitive Neuroscience Revolution. *Synthese*, *193* (*5*), 1509-1534.

Boone, W., y Piccinini, G. (2016). Mechanistic Abstraction. *Philosophy of Science, 83*(5), 686-697.

Bringsjord, S. y Govindarajulu, N. (2018). Artificial Intelligence. En E. Zalta (Ed.), *Stanford Encyclopedia of Philosophy.*

Brooks, R. (1997). Intelligence without Representation. En J. Haugeland (Ed.), *Mind Design II. Philosophy, Psychology, Artificial Intelligence* (pp. 395-420). Cambridge, MA: MIT Press.

Buckner, C. y Garson, J. (2019). Connectionism and post-connectionist models. En M. Sprevak y M. Colombo (Eds.), *The Routledge Handbook of the Computational Mind* (pp. 76-90). Londres: Routledge.

Carnap, R. (1963). Carl G. Hempel on Scientific Theories. En P. A. Schilpp (Ed.), *The Philosophy of Rudolf Carnap* (pp. 958-966). La Salle, IL: Open Court.

Carnap, R. (1991). Logical Foundations of the Unity of Science. En: R. Boyd y P. Gasper (Eds.), *The Philosophy of Science* (pp. 393-404). Cambridge, MA: MIT Press.

Cat, J. (2017). The Unity of Science. En E. Zalta (Ed.), *Stanford Encyclopedia of Philosophy.*

Copeland, J. (2000). Narrow Versus Wide Mechanism: Including a Re-Examination of Turing's Views on the Mind-Machine Issue. *The Journal of Philosophy*, XCVI (1), 5–33.

Copeland, J. (2006). The Modern History of Computing. En E. Zalta (Ed.), *Stanford Encyclopedia of Philosophy.*

Copeland, J. (2017). The Church-Turing Thesis. En E. Zalta (Ed.), *Stanford Encyclopedia of Philosophy.*

Chomsky, N. (1959). Review of B. F. Skinner's *Verbal behavior. Language*, 35, 26–58.

Chomsky, N. (1997). Language and Cognition. En D. Martel y C. Erneling (Eds.), *The Future of Cognitive Revolution* (pp. 15-31). New York: Oxford University Press.

Churchland, P. y Churchland, P. (1990). Could a Machine Think? *Scientific American*, 262(1), 32-37.

Churchland, P. y Sejnowski, T. (1992). *The Computational Brain*. Cambridge, MA: MIT Press.

Cummins, R. (1983). *The Nature of Psychological Explanation*. Cambridge, MA: MIT Press.

Cummins, R. (2000). "How Does it Work" versus "What are the Laws?": Two Conceptions of Psychological Explanation. En F. Keil y R. Wilson (Eds.), *Explanation and Cognition* (pp. 117-145). Cambridge, MA: MIT Press.

Crane, T. (2016). *The Mechanical Mind*. Londres: Routledge.

Craver, C. (2005). Beyond reduction: mechanisms, multifield integration and the unity of neuroscience. *Studies In History And Philosophy Of Biological And Biomedical Sciences*, *36*(2), 373-395.

Craver, C. (2007). *Explaining the Brain*. Oxford: Oxford University Press.

Dennett, D. (1981). Artificial Intelligence as Philosophy and as Psychology. En *Brainstorms*. *Philosophical Essays on Mind and Psychology* (pp. 109-126). Cambridge, MA, MIT Press.

Descartes, R. (2011). *Descartes*. Madrid: Gredos.

De Mol, L. (2018). Turing Machines. En E. Zalta (Ed.), *Stanford Encyclopedia of Philosophy.*

Díez, J. y Moulines, U. (1999). *Fundamentos de filosofía de la ciencia*. Barcelona: Ariel.

Egan, F. (2010). Computational models: a modest role for content. *Studies In History And Philosophy Of Science*, *41*(3), 253-259.

Egan, F. (2012). Representationalism. En E. Margolis, R. Samuels y S. Stich (Eds.), *The Oxford Handbook of Philosophy of Cognitive Science* (pp. 250-272). New York: Oxford University Press.

Egan, F. (2013). How to think about mental content. *Philosophical Studies*, *170*(1), 115-135.

Egan, F. (2019). The Nature and Function of Content in Computational Models. En M. Sprevak y M. Colombo (Eds.), *The Routledge Handbook of the Computational Mind* (pp. 247-258). Londres: Routledge.

Elber-Dorozko, L. y Shagrir, O. (2019). Computation and Levels in the Cognitive and Neural Sciences. En M. Sprevak y M. Colombo (Eds.), *The Routledge Handbook of the Computational Mind* (pp. 205-222). Londres: Routledge.

Falcon, A. (2019). Aristotle on Causality. En E. Zalta (Ed.), *Stanford Encyclopedia of Philosophy.*

Faries, F. y Chemero, A. (2019). Dynamic Information Processing. En M. Sprevak y M. Colombo (Eds.), *The Routledge Handbook of the Computational Mind* (pp. 134-148). Londres: Routledge.

Flanagan, O. (1991). *The Science of the Mind*. Cambridge, MA: MIT Press.

Fodor, J. (1968). *Psychological Explanation*. New York: Random House.

Fodor, J. (1968b). The Appeal to Tacit Knowledge in Psychological Explanation. *Journal of Philosophy*, 65, 627-640.

Fodor, J. (1974). Special sciences (or: the disunity of science as a working hypothesis). *Synthese*, 28(2), 97–115.

Fodor, J. (1975). *The Language of Thought*. Cambridge, MA: Harvard University Press.

Franklin-Hall, L. (2016). New Mechanistic Explanation and the Need for Explanatory Constraints. En K. Aizawa y C. Gillett (Eds.), *Scientific Composition and Metaphysical Ground* (pp. 41-74). Londres: Palgrave Macmillan UK.

French, S. y Saatsi, J. (2014). Travelling in New Directions. En J. Saatsi y S. French (Eds.), *The Bloomsbury Companion to the Philosophy of Science* (pp. 357-378). New York NY: Bloomsbury.

Fresco, N. (2014). *Physical Computation and Cognitive Science*. Berlin, Heidelberg: Springer.

Frigg, R. y Hartmann, S. (2012). Models in Science. En E. Zalta (Ed.), *Stanford Encyclopedia of Philosophy.*

Gardner, H. (1985). *The Mind's New Science*: *A History of the Cognitive Revolution*. New York: Basic Books.

Glennan, S. (2016). Mechanisms and Mechanical Philosophy. En P. Humphreys (Ed.), *The Oxford Handbook of Philosophy of Science* (pp. 796-816). New York, NY: Oxford University Press.

Glennan, S. (2017). *The New Mechanical Philosophy*. New York. Oxford University Press.

Grush, R. y Damm, L. (2012). Cognition and the Brain. En E. Margolis, R. Samuels y S. Stich (Eds.), *The Oxford Handbook of Philosophy of Cognitive Science* (pp. 273-290). New York: Oxford University Press.

Halina, M. (2018). Mechanistic Explanation and its Limits. En S. Glennan y P. Illari (Eds.), *The Routledge Handbook of Mechanisms and Mechanical Philosophy* (pp. 213-224). Londres: Routledge.

Harnish, R. (2002). *Minds, Brains*, *Computers*. *An historical Introduction to the Foundations of Cognitive Science*. Malden, MA: Blackwell.

Harper, W. (2016). Newton's Argument for Universal Gravitation. En: R. Iliffe y G. Smith (Eds.), *The Cambridge Companion to Newton* (pp. 229-260). Cambridge: Cambridge University Press.

Hempel, C. (1965). Aspects of Scientific Explanation. En *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science* (pp. 331-496). New York: Free Press.

Hume, D. (1999). *An Enquiry Concerning Human Understanding*. Oxford: Oxford University Press.

Illari, P. y Russo, F. (2014). *Causality. Philosophical Theory Meets Scientific Practice*. New York: Oxford University Press.

Ioannidis, S. y Psillos, S. (2017). In Defense of Methodological Mechanism: The Case of Apoptosis. *Axiomathes*, 27(6), pp.601-619.

Isaac, A. (2019). Computational Thought from Descartes to Lovelace. En M. Sprevak y M. Colombo (Eds.), *The Routledge Handbook of the Computational Mind* (pp. 9-22). Londres: Routledge.

Janiak, A. (2014). Newton's Philosophy. En E. Zalta (Ed.), *Stanford Encyclopedia of Philosophy.*

Kaplan, D. (2017). Integrating Mind and Brain Science. A Field Guide. En D. Kaplan (Ed.), *Explanation and Integration in Mind and Brain Science* (pp. 1-28). New York: Oxford University Press.

Kim, J. (1992). Multiple Realization and the Metaphysics of Reduction. *Philosophy and Phenomenological Research*, *52*(1), 1–26.

Kuhn, T. (1996). *The Structure of Scientific Revolutions*. Chicago: The University of Chicago Press.

Loeb, J. (1912). The Mechanistic Conception of Life. En *The Mechanistic Conception of Life. Biological Essays* (pp. 3-34). Chicago: The University of Chicago Press.

Maley, C. y Piccinini, G. (2017). A Unified Mechanistic Account of Teleological Functions for Psychology and Neuroscience. En D. Kaplan (Ed.), *Explanation and Integration in Mind and Brain Science* (pp. 236-256). New York: Oxford University Press

Marcus, G. (2014). The Computational Brain. En G. Marcus y J. Freeman (Eds.), *The Future of the Brain* (pp. 205-215). Princeton: Princeton University Press.

Marcus, G. y Freeman, J. (Eds.) (2014). *The Future of the Brain*. Princeton: Princeton University Press.

Marraffa, M., y Paternoster, A. (2012). Functions, levels, and mechanisms: Explanation in cognitive science and its problems. *Theory & Psychology*, *23*(1), 22-45.

Martel, D. y Emeling, C. (1997). *The Future of Cognitive Revolution*. New York: Oxford University Press.

Martel Johnson, D. (1997). What is the Purported Discipline of Cognitive Science and Why it Does It Need to Be Reassessed at the Present Moment? En D. Martel y C. Erneling (Eds.), *The Future of Cognitive Revolution* (pp. 3-11). New York: Oxford University Press

Matthewson, J. (2018). Models of Mechanisms. In: S. Glennan and P. Illari, ed., *The Routledge Handbook of Mechanisms and Mechanical Philosophy*. Londres: Routledge.

McGilvray, J. (2017). Cognitive Science: What Should It Be? En J. McGilvray (Ed.), *The Cambridge Companion to Chomsky* (pp. 175-195). New York: Cambridge University Press.

McCulloch, W. y Pitts, W. (1943). A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics*, 7, 115–133.

Miłkowski, M. (2018a). Mechanisms and the Mental. En S. Glennan y P. Illari (Eds.), *The Routledge Handbook of Mechanisms and Mechanical Philosophy* (pp. 74-88). Londres: Routledge.

Miłkowski, M. (2018b). From Computer Metaphor to Computational Modeling: The Evolution of Computationalism. *Minds and Machines*, 28(3), pp.515-541.

Miller, G. (2003). The Cognitive Revolution: a Historical Perspective. *Trends in Cognitive Sciences*, *7*(3), 141-144.

Nagel, E. (1961). *The Structure of Science*: *Problems in the Logic of Scientific Explanation*. New York: Harcourt, Brace.

Newell, A. (1990). *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.

Núñez, R., Allen, M., et al. (2019). What happened to cognitive science? *Nature Human Behaviour*, 3, 782-791

Oppenheim, P. y Putnam, H. (1958). Unity of science as a working hypothesis. *Minnesota Studies in the Philosophy of Science*, 2, 3–36.

Patterson, D. y Hennesy, J. (1998). *Computer Organization and Desing*. *The Hardware/Software Interface*. San Francisco: Morgan Kauffman.

Piccinini, G. (2004a). Functionalism, Computationalism, and Mental States. *Studies In History And Philosophy Of Science*, *35*(4), 811-833.

Piccinini, G. (2004b). Functionalism, Computationalism, and Mental Contents. *Canadian Journal Of Philosophy*, *34*(3), 375-410.

Piccinini, G. (2004c). The First Computational Theory of Mind and Brain: A Close Look at Mcculloch and Pitts's "Logical Calculus of Ideas Immanent in Nervous Activity". *Synthese*, *141*(2), 175-215.

Piccinini, G. (2007a). Computational modelling vs. Computational explanation: Is everything a Turing Machine, and does it matter to the philosophy of mind? *Australasian Journal of Philosophy*, 85(1), 93-115.

Piccinini, G. (2007b). Computationalism, The Church–Turing Thesis, and the Church–Turing Fallacy. *Synthese*, *154*(1), 97-120.

Piccinini, G. (2008a). Computers. *Pacific Philosophical Quarterly*, *89*(1), 32-73.

Piccinini, G. (2008b). Computation without Representation. *Philosophical Studies*, *137*(2), 205-241.

Piccinini, G. (2008c). Some neural networks compute, others don't. *Neural Networks*, *21*(2-3), 311-321.

Piccinini, G. (2010). The Mind as Neural Software? Understanding Functionalism, Computationalism, and Computational Functionalism. *Philosophy and Phenomenological Research*, 81(2), pp.269-311.

Piccinini, G. (2011). The Physical Church–Turing Thesis: Modest or Bold? *The British Journal for the Philosophy of Science*, *62*(4), 733-769.

Piccinini, G. (2012). Computationalism. En E. Margolis, R. Samuels y S. Stich (Eds.), *The Oxford Handbook of Philosophy of Cognitive Science* (pp. 222-249). New York: Oxford University Press.

Piccinini, G. (2018a). Computational Mechanisms. En S. Glennan y P. Illari (Eds.), *The Routledge Handbook of Mechanisms and Mechanical Philosophy* (pp. 435-446). Londres: Routledge.

Piccinini, G. (2018b). Computation and Representation in Cognitive Neuroscience. *Minds And Machines*, *28*(1), 1-6.

Piccinini, G., y Bahar, S. (2013). Neural Computation and the Computational Theory of Cognition. *Cognitive Science*, *37*(3), 453-488.

Piccinini, G. y Craver, C. (2011). Integrating psychology and neuroscience: functional analyses as mechanism sketches. *Synthese*, 183(3), 283-311.

Piccinini, G. y Maley, C. (2014). The Metaphysics of Mind and the Multiple Sources of Multiple Realizability. En M. Sprevak y J. Kallestrup (Eds.), *New Waves in Philosophy of Mind* (pp. 125-152). Londres: Palgrave Macmillan.

Piccinini, G., y Scarantino, A. (2010). Information processing, computation, and cognition. *Journal Of Biological Physics*, *37*(1), pp. 1-38.

Piccinini, G., y Shagrir, O. (2014). Foundations of computational neuroscience. *Current Opinion In Neurobiology*, *25*, 25-30.

Pinker, S. (1997). *How the Mind Works*. New York: W. W. Norton.

Potochnik, A. (2010). A Neurathian Conception of the Unity of Science. *Erkenntnis*, *74*(3), 305-319.

Povich, M. y Craver, C. (2018). Mechanistic Levels, Reduction and Emergence. En S. Glennan y P. Illari (Eds.), *The Routledge Handbook of Mechanisms and Mechanical Philosophy* (pp. 185-197). Londres: Routledge.

Proudfoot, D. y Copeland, J. (2012). Artificial Intelligence. En E. Margolis, R. Samuels y S. Stich (Eds.), *The Oxford Handbook of Philosophy of Cognitive Science* (pp. 147-182). New York: Oxford University Press.

Proudfoot, D. y Copeland, J. (2019). Turing and the first electronic brains: What the papers said. En M. Sprevak y M. Colombo (Eds.), *The Routledge Handbook of the Computational Mind* (pp. 23-37). Londres: Routledge.

Psillos, S. (2007). Past and Contemporary Perspectives on Explanation. En T. Kuipers (Ed.), *General philosophy of science: Focal Issues* (pp. 97-173). Amsterdam: North Holland.

Psillos, S. (2016). Having Science in View: General Philosophy of Science and Its Significance. En P. Humphreys (Ed.), *The Oxford Handbook of Philosophy of Science* (pp. 137-161). New York, NY: Oxford University Press.

Putnam, H. (1960). Minds and Machines. En S. Hook (Ed.) *Dimensions of Mind: A Symposium* (pp. 138-164). New York, Collier.

Putnam, H. (1967). The Mental Life of Some Machines. En H. Castañeda (Ed.), *Intentionality, minds, and perception* (pp. 177–200). Detroit: Wayne State University Press.

Putnam, H. (1975a). Philosophy and Our Mental Life. En *Philosophical Papers Vol.2* (pp. 291-303). Cambridge: Cambridge University Press.

Putnam, H. (1975b). The Nature of Mental States. En *Philosophical Papers Vol.2* (pp. 429-440). Cambridge: Cambridge University Press.

Rescorla, M. (2015). The Computational Theory of Mind. En E. Zalta (Ed.), *Stanford Encyclopedia of Philosophy.*

Ritchie, J. B. y Piccinini, G. (2019). Computational Implementation. En M. Sprevak y M. Colombo (Eds.), *The Routledge Handbook of the Computational Mind* (pp. 192-204). Londres: Routledge.

Rorty, R. (1979). *Philosophy and the Mirror of Nature*. Princeton, NJ: Princeton University Press, 1979.

Roux, S. (2018). From the Mechanical Philosophy to Early Modern Mechanisms. En S. Glennan y P. Illari (Eds.), *The Routledge Handbook of Mechanisms and Mechanical Philosophy* (pp. 26-45). Londres: Routledge.

Ruphy, S. (2016). *Scientific Pluralism Reconsidered*. Pittsburgh: University of Pittsburgh Press.

Samuels, R. (2019). Classical Computational Models. En M. Sprevak y M. Colombo (Eds.), *The Routledge Handbook of the Computational Mind* (pp. 103-119). Londres: Routledge.

Sejnowski, T., Koch, C., y Churchland, P. (1988). Computational neuroscience. *Science*, 241(4871), 1299–306.

Serban, M. (2015). The scope and limits of a mechanistic view of computational explanation. *Synthese*, 192(10), pp.3371-3396.

Shagrir, O. (2006). Why we view the brain as a computer. *Synthese*, *153*(3), 393-416.

Shannon, C. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, *27*(3), 379-423.

Sieg, W. (2009). On Computability. En A. Irvine (Ed.), *Philosophy of Mathematics* (Handbook of the Philosophy of Science) (pp. 535-630). Amsterdam: North Holland.

Sprevak, M. (2016). Philosophy of the Psychological and Cognitive Sciences. En P. Humphreys (Ed.), *The Oxford Handbook of Philosophy of Science* (pp. 92-114). New York, NY: Oxford University Press.

Sklar, L. (1999). The Reduction (?) of Thermodynamics to Statistical Mechanics. *Philosophical Studies*, 95, 187-202.

Stinson, C. (2019). Explanation and Connectionist Models. En M. Sprevak y M. Colombo (Eds.), *The Routledge Handbook of the Computational Mind* (pp. 120-133). Londres: Routledge.

Turing, A. (1936). On Computable Numbers, with an Application to the *Entscheidungsproblem*. *Proceedings of the London Mathematical Society* (Series 2), 42, 230–265.

Uebel, T. (2007). Carnap and the Vienna Circle: Rational Reconstructionism Refined. En M. Friedman y R. Creath (Ed.), *The Cambridge Companion to Carnap* (pp. 153-175). Cambridge: Cambridge University Press.

Van Gelder, T. (1995). What Might Cognition Be, if not Computation? *The Journal of Philosophy*, 92, 345-81.

Walter, S. y Eronen, M. (2014). Reduction, Multiple Realizability and Levels of Reality. En J. Saatsi y S. French (Eds.), *The Bloomsbury Companion to the Philosophy of Science* (pp. 138-156). New York NY: Bloomsbury.

Weisberg, R. y Reeves, L. (2013). *Cognition. From Memory to Creativity*. New Jersey: John Wiley & Sons.

Weiskopf, D. (2017). The Explanatory Autonomy of Cognitive Models. En D. Kaplan (Ed.), *Explanation and Integration in Mind and Brain Science* (pp. 44-69). New York: Oxford University Press.

Weiskopf, D. (2019). Reductive explanation between psychology and neuroscience. En M. Sprevak y M. Colombo (Eds.), *The Routledge Handbook of the Computational Mind* (pp. 223-236). Londres: Routledge.

Winsberg, E. (2015). Computer Simulations in Science. En E. Zalta (Ed.), *Stanford Encyclopedia of Philosophy.*

Woodward, J. y Hitchcock, C. (2003). Explanatory Generalizations I: A Counterfactual Account. *Nous*, *37*(1), 1-24.

Woodward, J. (2017). Interventionism and the Missing Metaphysics. En M. Slater y Z. Yudell (Ed.), *Metaphysics and the Philosophy of Science* (pp. 193-228). New York, NY: Oxford University Press.

Ylikoski, P. (2013). Causal and Constitutive Explanation Compared. *Erkenntnis*, *78*(S2), 277-297.