Universidad de Valladolid

UNIVERSIDAD DE VALLADOLID

Escuela de Doctorado

Máster de Lógica y Filosofía de la Ciencia

Simulación y explicación psicológica: evaluación de la validez de los modelos cognitivos basados en simulaciones

Simulation and psychological explanation: Assessing the validity of simulation-based models of human cognition

Laura Senis Villacé

Tutor: José Vicente Hernández Conde

Departamento de Filosofía

Curso 2024-2025

Resumen

La tesis principal que se pretende defender en el presente trabajo es que los modelos cognitivos basados en simulaciones constituyen una explicación psicológica, si y solo si, dicha explicación es plural y normativamente informada. Su validez depende de su capacidad para incorporar el componente normativo inherente a lo mental, e integrar diferentes niveles de explicación, no reducibles entre sí, en consonancia con el pluralismo científico que se defiende. Para fundamentar esta tesis, la investigación se estructura en tres partes. Primero, se revisa la teoría de la modelización en las ciencias cognitivas para caracterizar los modelos basados en simulaciones. Segundo, se analizan las diferentes concepciones de la explicación científica y sus limitaciones. Finalmente, se evalúa la validez de estos modelos como explicaciones psicológicas a través de un análisis transversal que aborda la validez experimental de la simulación, cuestiones ontológicas y epistémicas sobre el emergentismo, la brecha explicativa, y la defensa del pluralismo científico. El trabajo concluye con una propuesta de explicación normativa de la mente que pueda servir de condición necesaria, aunque no suficiente, para toda explicación psicológica válida futura.

Palabras clave

Modelos cognitivos; simulación; explicación psicológica; pluralismo científico; normatividad

Abstract

This paper argues that cognitive simulation-based models are valid psychological explanations only if they are both plural and normatively informed. The explanatory power of these models depends on their capacity to incorporate the normative component inherent to mental phenomena while also integrating multiple, irreducible levels of explanation, a stance consistent with scientific pluralism. To support this thesis, the paper first reviews the theory of modeling in the cognitive sciences to characterize simulation-based models. Next, it analyzes different conceptions of scientific explanation and their limitations. Finally, it assesses the validity of these models as psychological explanations through a multifaceted analysis, which includes an examination of the experimental validity of simulation, ontological and epistemic issues concerning emergentism the explanatory gap, and a defense of scientific pluralism. The paper concludes by proposing a normative explanation of the mind, which serves as a necessary, though not sufficient, condition for any future valid psychological explanation.

Keywords

Cognitive models; simulation; psychological explanation; scientific pluralism; normativity

A la logopeda que me dio *voz,* Beatriz Penedo Vázquez. *In memoriam.*

A Daniel, por ser un compañero de vida inmejorable.

A mi familia, amigos y tutor, José. *Natürlich*.

ÍNDICE DE CONTENIDOS

1. INTRODUCCIÓN

If you can't model a process, you don't understand it.

John von Neumann

John von Neumann formuló su célebre aforismo —"si no puedes modelar un proceso, entonces no lo entiendes" — en el contexto de sus trabajos pioneros sobre simulación numérica y teoría de autómatas a mediados del siglo XX. Consciente de que las ecuaciones analíticas a menudo resultaban inaccesibles para los fenómenos complejos, von Neumann impulsó el desarrollo de métodos de simulación por computadora (como los métodos de Monte Carlo) para "ensayar" procesos físicos y biológicos en el ordenador. Para él, la construcción de un modelo operativo —incluso si simplificado— revelaba la estructura causal subyacente del sistema y permitía explorar su comportamiento en condiciones variables. Esta perspectiva supuso un cambio de paradigma al pasar de la simple resolución de fórmulas a la indagación experimental por medio de simulaciones, con la convicción de que solo mediante esa traducción a un modelo manipulable se alcanzaba una comprensión genuina (von Neumann & Burks, 1966). Tomando como punto de partida este principio, la pregunta que guía nuestro trabajo es análoga pero centrada en la esfera de la mente humana: si logramos construir un modelo capaz de reproducir un proceso psicológico —ya sea de razonamiento, atribución de intenciones, control motor, etc.—, ¿hemos obtenido una explicación psicológica de ese proceso? En otras palabras, ¿modelar implica explicar? Más aún, ¿constituye una explicación cuando se consigue experimentalmente simular la conducta observable asociada al proceso que se ha pretendido modelar?

El uso de simulaciones computacionales en las últimas décadas con respecto al estudio de la mente humana ha adquirido una centralidad creciente dentro de las ciencias cognitivas. Desde los modelos conexionistas hasta los algoritmos bayesianos o las arquitecturas cognitivas como ACT-R, se ha intentado modelar funciones mentales complejas mediante sistemas artificiales que reproducen, con distintos grados de éxito, comportamientos observables, patrones de decisión, errores sistemáticos o incluso formas rudimentarias de aprendizaje. Estas simulaciones se han utilizado normalmente como herramientas predictivas o descriptivas, pero en algunos casos son presentadas además como explicación psicológica, es decir, como instrumentos capaces de darnos cuenta de por qué ciertos procesos mentales se producen, y bajo qué condiciones emergen. No obstante, el estatuto epistemológico de las simulaciones no está libre de debate. ¿Explican realmente los fenómenos que modelan o solo los imitan? ¿Qué tipo de relación hay entre la simulación de una conducta y la comprensión de los mecanismos que la originan?

La relevancia filosófica de estas preguntas estriba en que se inscriben en los debates generales sobre la naturaleza de los modelos científicos, la noción de representación, y los criterios de validez epistémica en la ciencia. Autores como van Fraassen, Giere, Frigg o Cartwright han problematizado

la idea de que los modelos representen fielmente los sistemas que modelan, y han propuesto en su lugar visiones más pragmáticas, instrumentales o estructuralistas. La naturalización de la mente, entendida como su reducción a procesos naturales modelizables, ha sido defendida por algunos como un ideal regulativo de la ciencia cognitiva, pero ha sido también objeto de crítica por parte de quienes consideran que la mente humana no puede ser plenamente comprendida desde una perspectiva exclusivamente externa, computacional o causal. Aunque el modelo sea el que da cuenta del posible proceso o mecanismo que tiene su análogo en la cognición humana, tendría este un carácter demasiado especulativo sin el testeo experimental que facilita la simulación, con la cual se puede poner a prueba lo que propone el modelo.

La aparición de la Inteligencia Artificial, junto con los recientes Modelos de Lenguaje Grande (*Large Language Models* o LLMs), nos obliga, no tanto a dar un volantazo de 180 grados, pero sí, cuando menos, a replantear y repensar algunos conceptos y teorías de la filosofía de la mente y del lenguaje. El hecho de que puedas mantener una conversación con Chat GPT —esto es, que sea capaz de "imitar" o "simular" ser un interlocutor como podría serlo un amigo tuyo, un profesor o tu vecino–, no debería de tomarse a la ligera y negar de antemano que pueda tener una repercusión teórica. El desarrollo de sistemas que aprenden, predicen y toman decisiones con base en grandes volúmenes de datos y estructuras algorítmicas complejas depende, en buena medida, de la capacidad para interpretar esos sistemas como representaciones válidas de procesos psicológicos reales. Consiguientemente, las decisiones clínicas, educativas o incluso legales que puedan derivarse de estos modelos están condicionadas por la confianza epistémica que se les otorgue. De este modo, evaluar con precisión cuál es el alcance explicativo y predictivo de los modelos basados en simulaciones no es solo un ejercicio teórico, sino una necesidad urgente para garantizar la responsabilidad y la transparencia en el uso de tecnologías cognitivas. Los niveles de sofisticación, completamente sin precedentes, que ha alcanzado la inteligencia artificial, pues estos sistemas no solo imitan conductas humanas, sino que parecen interactuar, aprender y adaptarse de manera autónoma, hacen hoy más que nunca conveniente reflexionar sobre estas cuestiones que no suelen ser, desgraciadamente, el tópico preferido, ni mucho menos, dentro de la ingente literatura sobre IA. Aunque estos sean fascinantes, cabe asimismo atemperar nuestro entusiasmo ante el posible riesgo de una suerte de "ilusión epistémica". ¿Se está confundiendo simulación con explicación? ¿Apariencia de inteligencia con inteligencia *efectiva*? ¿Modelización formal con comprensión del fenómeno mental?

La estructura del trabajo se organiza en tres capítulos principales, además de esta introducción y una conclusión final. El primer capítulo examina el concepto de modelo en las ciencias cognitivas, diferenciando entre modelos representacionales e instrumentales, y discutiendo sus tipos y funciones, con especial atención a los modelos computacionales y simulativos. Asimismo este se dedica a caracterizar los modelos cognitivos basados en simulaciones y a presentar los argumentos que se han formulado a favor y en contra de su capacidad para representar adecuadamente la mente humana. El segundo capítulo aborda principalmente la noción de explicación en ciencia, y, en concreto, en psicología. Finalmente, el tercer capítulo discute las condiciones que debe cumplir una explicación psicológica y evalúa en qué medida las simulaciones pueden satisfacerlas. Este último capítulo concluye con una reflexión sobre los límites de la simulación en relación con la comprensión de la mente, abordando los problemas de la emergencia, la brecha explicativa y la dimensión fenomenológica del sujeto. La tesis principal que se pretende defender en el presente trabajo es que

los modelos cognitivos basados en simulaciones son una buena herramienta para aproximarnos a ciertos fenómenos psíquicos y que son capaces de constituirse como una explicación psicológica completa, si y solo si, se comprende la explicación de forma plural y normativamente informada. Así, la validez explicativa de las simulaciones cognitivas depende de su capacidad para incorporar el componente normativo que se defenderá que es inherente a lo mental, e, igualmente, integrar diferentes niveles de explicación, no reducibles entre sí, y que son resultado del pluralismo científico que se sostiene en el trabajo.

2. MODELIZACIÓN CIENTÍFICA Y MODELOS COGNITIVOS BASADOS EN SIMULACIONES

A program, no matter how sophisticated, only simulates the mind; it is not thereby the mind.

Hilary Putnam

Los modelos, cual metáforas, no poseen un mero significado figurado, no poseen un mero carácter ornamental, sino que tienen un poder creador, representacional. Las metáforas, como desde el siglo XIX se han comprendido, pueden reivindicarse como una suerte de instrumentos cognitivos que de manera primigenia nos permiten acercarnos a la realidad externa, a nosotros mismos. El lenguaje es un ejemplo de ello; intermediario que a veces se ha considerado entre mente y mundo. Es, precisamente, a través del lenguaje, de las metáforas, de los modelos, los conceptos, que podemos asimilar nuestras experiencias, otorgarlas sentido. La metáfora no solo puede caracterizarse por el desplazamiento del significado, sustituyéndose así el término habitual por otro menos frecuente, sino que cumplen una función no meramente estilística cuando se usan para designar aquello que no posee ninguna etiqueta asignada por convención social alguna. Cumpliendo de este modo la función que según Aristóteles en su *Retórica* y *Poética* debiese tener la metáfora, esto es, poner la cosa ante los ojos. Los modelos, cual metáforas, son una forma de aproximación, de hacer patente lo desconocido mediante lo conocido; lo abstracto, mediante lo concreto; lo oscuro, mediante lo claro.

La noción de modelo ocupa un lugar central en la epistemología contemporánea de la ciencia. Lejos de ser meras herramientas ilustrativas o secundarias, los modelos científicos cumplen una doble función: por un lado, sirven como representaciones estructuradas de fenómenos complejos que permiten hacer inferencias, predicciones y manipulaciones teóricas; por otro, actúan como instrumentos epistémicos que median entre la teoría y el mundo, facilitando la comprensión, exploración y comunicación del conocimiento. Esta doble dimensión —representacional y heurística— está en el núcleo del debate sobre qué significa modelar en ciencia, y resulta especialmente relevante cuando se aborda el estudio de procesos mentales en psicología y ciencias cognitivas. Pero ¿qué es un modelo científico? La literatura filosófica ha ofrecido múltiples definiciones del concepto de modelo, reflejando la diversidad de usos que tiene en las distintas disciplinas científicas. Algunos autores han defendido una concepción representacional según la cual un modelo es una estructura que representa, de manera más o menos idealizada, un sistema real o teórico. Esta idea está presente, por ejemplo, en los enfoques estructuralistas, como el de Frederick Suppe o Patrick Suppes, donde los modelos son sistemas formales que comparten una estructura isomórfica con la realidad que intentan

representar. Otros han adoptado perspectivas más pragmáticas o instrumentalistas, donde el valor de un modelo no reside en su correspondencia ontológica con el mundo, sino en su utilidad para generar predicciones exitosas, organizar datos o guiar la acción experimental. Una definición reciente (2018) del astrofísico y filósofo de la ciencia Gustavo E. Romero, discípulo de Mario Bunge, los ha pretendido definir formalmente del siguiente modo: un modelo fáctico es la representación conceptual de un mecanismo. Un modelo no sería una aplicación de las matemáticas a la realidad, cuanto una matematización de nuestras ideas sobre la realidad. El avance de la matemática permitiría construir modelos alternativos pero empíricamente equivalentes de un proceso o mecanismo dado. Y, como cada modelo es simbólico, tiene ciertos elementos convencionales y, asimismo, como la matematización implica idealización, los modelos siempre son defectuosos en algún aspecto u otro. De modo que, en el mejor de los casos, son buenas aproximaciones, pero no deben confundirse con la realidad. No cabe confundir los modelos de las teorías, las cuales son un conjunto lógicamente organizado de enunciados referidos a objetos de una misma clase.

En el marco actual de la filosofía de la ciencia se ha vuelto común distinguir entre diferentes tipos de modelos: modelos formales (matemáticos o computacionales); modelos analógicos (como los modelos mecánicos); y modelos conceptuales o cualitativos. Frigg y Hartmann (2021) insisten en que los modelos no deben entenderse como meras descripciones reducidas de la realidad, sino como construcciones idealizadas que incorporan supuestos, simplificaciones y objetivos epistémicos específicos. Así, la representación que ofrece un modelo no debe interpretarse como un espejo del mundo, sino como una herramienta que destaca ciertos aspectos relevantes del sistema estudiado y omite otros, según las intenciones del investigador. Esta concepción da lugar a una pregunta fundamental: ¿qué significa que un modelo represente algo? Para algunos, representar implica fidelidad estructural o semántica; para otros, como Ronald Giere (2004), los modelos representan cuando los científicos los usan como representaciones, es decir, cuando son tratados como tales dentro de prácticas epistemológicas determinadas. Esto introduce una dimensión contextual y práctica en la noción de representación científica, la cual resulta especialmente útil para analizar los modelos en psicología, donde el grado de correspondencia directa entre modelo y fenómeno suele ser bajo, y se dan exclusiones necesarias de fenómenos psíquicos no susceptibles de ser abordados científicamente sin una aparente pérdida significativa de su contenido semántico. Pero esta pérdida o exclusión no debe de tomarse como un hándicap, pues para que un modelo científico tenga una funcionalidad como herramienta epistemológica debe de ser una representación idealizada de un sistema o fenómeno. Para ilustrar este punto es pertinente sacar a colación la distinción de Weisberg (2013) entre idealización galileana y minimalista. Esta primera consiste en la práctica de simplificar modelos y teorías mediante la introducción de distintas distorsiones con el objetivo de hacerlos más manejables en la práctica. Se sostiene en la idea de Galileo Galilei de poder avanzar en un problema utilizando un montaje experimental similar a la situación prevista; de modo que, una vez se ha trabajado sobre el modelo más simple, se eliminan los factores distorsionantes. Esta idealización es considerada por Weisberg como pragmática, pues considera los modelos como provisionales para ir sucesivamente elaborando modelos más acertados, *aproximadamente* verdaderos. La idealización minimalista en cambio es la práctica de construir modelos y teorías que incluyan solo los factores y partes necesarias y esenciales que dan lugar a un fenómeno; solo representan lo que realmente importa. Así, se pueden justificar las partes idealizadas demostrando que únicamente distorsionan características que son irrelevantes, que no suponen diferencias, que no son contextualmente destacadas o que no son de interés.

La psicología ha desarrollado históricamente distintos estilos de modelización. Durante el auge del conductismo, el modelo dominante era el de caja negra, esto es: se describían relaciones entre estímulos y respuestas sin intentar acceder a los procesos internos. En este marco, los modelos eran esencialmente funcionales o estadísticos, y su valor se medía por su capacidad de predecir la conducta observada. Desde este punto de vista, los modelos simulativos actuales podrían ser vistos como una sofisticación tecnológica de esta tradición, pues, si lo único que se espera de un modelo es que reproduzca la conducta observable, entonces una simulación exitosa puede ser considerada suficiente, sin necesidad de acceder al *interior* de la mente. Sin embargo, con el giro cognitivo de mediados del siglo XX, la psicología comenzó a interesarse por los procesos internos, y con ello surgió la necesidad de modelos que representaran no solo lo observable, sino también lo inferido: estructuras mentales, esquemas, reglas de procesamiento, memoria, etc. En este nuevo contexto, el modelo no es simplemente una herramienta predictiva, sino una hipótesis estructurada sobre cómo opera la mente. Muchos de los modelos actuales, incluidos los basados en simulación, operan en un terreno intermedio, ya que pretenden representar estructuras funcionales internas, pero lo hacen mediante sistemas artificiales cuya correspondencia con la mente humana es, en el mejor de los casos, indirecta o parcial. En algunos casos, como los modelos conexionistas, se busca una analogía estructural con las redes neuronales; en otros, como los modelos simbólicos, se apuesta por un isomorfismo lógico con las reglas del pensamiento. Pero en todos los casos, la brecha entre modelo y fenómeno persiste, y con ella la necesidad de preguntarse en qué sentido —si alguno— estos modelos representan realmente la mente humana.

Las distintas formas de teorizar sobre la mente no se excluyen mutuamente, sino que en muchos casos se superponen, se integran o se utilizan de manera complementaria en el intento de captar la complejidad del funcionamiento mental. Dado que la mente no es directamente observable, y que la psicología opera a menudo con constructos teóricos de difícil verificación empírica directa, los diferentes modelos desempeñan un papel fundamental en la formalización, contrastación y visualización de hipótesis cognitivas. De hecho, la modelización de la cognición ha evolucionado a través de varios enfoques. Uno de ellos son los modelos computacionales clásicos que conceptualizan la mente como un sistema simbólico que manipula representaciones formales mediante reglas; los modelos conexionistas, inspirados en las redes neuronales biológicas, los cuales procesan la información de forma distribuida y aprenden asociaciones ajustando pesos sinápticos —aunque su opacidad estructural dificulta la interpretación de sus mecanismos internos–; los modelos bayesianos, que conciben la cognición como un proceso inferencial de actualización de creencias basado en la evidencia –con gran rigurosidad formal, pero quizá demasiado *idealizados*, ya que los seres humanos rara vez aplican inferencias bayesianas de forma explícita; los modelos dinámicos, que describen la cognición como un acoplamiento continuo entre organismo y entorno, inspirados en la teoría de sistemas no lineales que, aun siendo prometedores en áreas como el control motor, su complejidad matemática y la falta de una teoría psicológica unificada representan un gran desafío en su fundamentos; asimismo, los modelos híbridos buscan combinar las fortalezas de los distintos enfoques anteriores, integrando, por ejemplo, estructuras simbólicas con aprendizaje conexionista. Pero, aunque ofrecen estos mayor flexibilidad, corren el riesgo de perder claridad conceptual al mezclar marcos teóricos diversos.

Por último, los modelos basados en simulaciones utilizan la simulación como principio constitutivo y no solo como medio de ejecución. Estos modelos no se limitan a representar formalmente funciones cognitivas, sino que buscan generar conductas o procesos análogos a los humanos mediante agentes artificiales, entornos virtuales o sistemas multi-agente. Su valor reside en su capacidad para producir fenómenos emergentes, explorar escenarios contrafácticos y testear hipótesis complejas de manera controlada. En algunos casos, estas simulaciones se presentan como sustitutos de experimentos empíricos, y en otros como formas de explicación de fenómenos cognitivos reales; pero, precisamente por ello, exigen una evaluación crítica más rigurosa.

Antes de exponer los cuatro tipos principales de modelos cognitivos basados en simulaciones que existen, es importante diferenciar entre simular una teoría y construir una teoría por simulación. Cuando se simula una teoría partimos de hipótesis formales previamente establecidas (p. ej. reglas de producción de ACT-R) y las implementamos en un ordenador para derivar predicciones cuantitativas. En este caso la validez empírica sigue dependiendo del ajuste entre datos y predicciones, no de la simulación *per se*. Así, simular una teoría implica disponer de un conjunto de conocimientos previos acerca de la mente o del comportamiento y utilizar la simulación interna simplemente como un mecanismo de aplicación de esas reglas. En esta perspectiva, el sujeto posee una teoría tácita —constituida por principios y esquemas causales sobre creencias, deseos e intenciones— y ejecuta mentalmente un "ensayo" de la situación concreta para ver cómo esas reglas se despliegan dinámicamente en un caso particular. De este modo, la simulación funciona como un motor de elaboración de inferencias: toma como entrada los datos contextuales (*inputs* ficticios) y los procesa según la teoría preexistente, generando predicciones o explicaciones que son inmediatamente atribuibles a la estructura teórica subyacente (Stich & Nichols, 1994). Pero construir una teoría por simulación consiste en que la propia dinámica del modelo —habitualmente agentes artificiales que interactúan— genera regularidades que después se interpretan teóricamente. Construir una teoría por simulación supone partir sin un paquete fijo de reglas y emplear la simulación interna de manera exploratoria y generativa. Aquí, la mente crea diversos modelos candidatos —simulaciones sencillas de posibles mecanismos de atribución— y, tras ejecutar esos modelos, compara los resultados de cada simulación con los datos observados en el mundo real. A partir de las divergencias y coincidencias entre predicción y observación, se infieren patrones recurrentes que, poco a poco, conforman un cuerpo de reglas o esquemas causales. Este proceso de ajuste iterativo, por ejemplo, recuerda los métodos de "model fitting" en psicología computacional, donde varias hipótesis se prueban por medio de la simulación, y solo aquellas que mejor explican los datos pasan a formar parte de la teoría definitiva (Rabinowitz et al., 2018). Aquí la explicación no precede sino que *emerge* de la simulación (Grimm, V. et al., 2005), constituyendo de este modo los agentes artificiales el componente fundamental. Según Sun, R. (2005), un agente cognitivo sintético sería un sistema computacional cuyo diseño intenta captar los principios organizativos de la mente humana. Para materializarlos se emplean al menos tres grandes familias de técnicas: (1) Sistemas multi-agente (MAS), los cuales permiten representar interacciones sociales, aprendizaje cultural o economía conductual. Cada agente posee reglas relativamente simples, pero el conjunto exhibe fenómenos macroscópicos difíciles de deducir analíticamente (Gilbert, N. y Troitzsch, K. G., 2005); (2) Redes profundas (*deep learning*) y modelos conexionistas de nueva generación. Aunque inicialmente se enfocaron en reconocimiento perceptivo, arquitecturas como los *transformers* han ampliado la simulación de procesos lingüísticos y de razonamiento (Bengio, 2021); y (3) Simulaciones Monte Carlo y técnicas de inferencia aproximada, los cuales son útiles cuando el

espacio de estados es inabarcable; posibilitando la evaluación de la plausibilidad cognitiva bajo incertidumbre, al hacer un muestreo de escenarios contrafácticos.

Los tipos principales que existen de modelos cognitivos basados en simulaciones:

(i) *Emulation Theory.* Propuesta por Rick Grush (1995; 2004), postula que el cerebro construye internamente circuitos —denominados *emuladores*— que imitan los patrones entrada-salida de diversas operaciones cognitivas (acción, percepción, imaginación). Estos emuladores reciben copias efectoras (*efference copies*) de las órdenes motoras dirigidas al cuerpo y generan "señales sensoriales ficticias" con menor retardo que los propios sensores. Al funcionar en paralelo, permiten anticipar los resultados de nuestras acciones y, en ausencia de movimiento, sostener procesos de imaginación motora y visual. Existen tanto emuladores modales (específicos de cada sentido, por ejemplo, visión o audición) como emuladores amodales (más abstractos, que representan la geometría o la cinemática del entorno), lo que otorga flexibilidad y riqueza a la simulación interna. Aunque esta teoría fue formulada en términos neuronales para el control motor, recientes trabajos en robótica muestran su aplicación práctica. Nicolescu y colaboradores (2025) diseñan arquitecturas distribuidas para equipos humano-robot que usan la misma idea de "emuladores" internos: cada robot mantiene un modelo interno (emulator) que simula en tiempo real las acciones propias y ajenas, anticipando obstáculos o solapamientos de tareas en entornos dinámicos. De este modo, la emulación no es solo predicción sensoriomotora, sino también coordinación colaborativa en sistema multiagente.

(ii) *Internal Forward Models*. Estos son arquitecturas en las que un componente del sistema cognitivo, el *modelo interno*, simula el comportamiento del cuerpo o de agentes externos para predecir las consecuencias de posibles acciones. Estos modelos facilitan el control anticipatorio, la corrección en tiempo real y la planificación basada en las predicciones de los efectos de las acciones. En robótica —como se muestra en "Simulation-Based Internal Models for Safer Robots"— esta arquitectura incluye un "motor controller" que genera comandos de acción, un "forward model" que recibe estos comandos y anticipa la trayectoria resultante, y un "error estimator" que compara la trayectoria simulada con la trayectoria real del robot, permitiendo ajustar el modelo interno y la selección de acciones. Los *forward models* sensoriomotores se expanden hoy hacia modelos de inteligencia social. Rabinowitz et al. (2018) construyen un "ToMnet" que, mediante meta-aprendizaje, aprende a predecir el comportamiento de otros agentes en un *gridworld*, usando muy pocas observaciones previas y refinando "on-line" su modelo interno del otro. Esta estrategia extiende el concepto clásico al no solo anticipar señales sensoriales, sino también aprender a simular intenciones y creencias falsas de agentes heterogéneos.

(iii) Mental Models. En la teoría de Philip Johnson-Laird (1983) los modelos mentales son representaciones internas que el sujeto construye para comprender y razonar sobre situaciones del mundo real. Según esta perspectiva, la mente funciona como una "máquina de construir modelos". Al procesar información, se crea un modelo mental de las relaciones entre objetos y eventos. El razonamiento deductivo se lleva a cabo manipulando mentalmente estos modelos, en lugar de aplicar reglas formales de lógica. Estos modelos son dinámicos, incompletos y

susceptibles de revisión a medida que se recibe nueva información. Este enfoque ha sido especialmente influyente en psicología cognitiva y ciencias de la computación para diseñar sistemas que imiten la comprensión humana de textos, resolución de problemas y toma de decisiones. Lebiere et al. (2025) muestra cómo arquitecturas cognitivas (p. ej. ACT-R) pueden personalizar un modelo mental a un individuo con muy pocos datos, trazando su historial de decisiones y ajustando parámetros de razonamiento causal para predecir fallos de juicio o cargas cognitivas futuras. Así, los modelos mentales se convierten en agentes analíticos capaces de "leer" y guiar comportamientos reales en tiempo real.

(iv) Simulation Theory. La teoría de simulación en psicología del sentido común plantea que, para entender y predecir el comportamiento de otros, "nos ponemos en su lugar" simulando sus estados mentales internamente. Esta idea se contrapone a la "theory-theory" (teoría-teoría), que sostiene que usamos inferencias basadas en reglas o teorías sobre la mente ajena. En "Folk Psychology: Simulation or Tacit Theory?" (1992), Stich y Nichols discuten varios argumentos a favor y en contra de la simulación: a favor, reportan que a veces sentimos que imaginamos cómo actuaríamos en la situación de otro antes de predecir su conducta. Asimismo, estudios empíricos sobre empatía sugieren que la simulación (o "role-taking") puede explicar el contagio emocional y la imitación motora en ciertos contextos. Sin embargo, no toda predicción de comportamiento ajeno requiere simulación; a veces basta con aplicar teorías implícitas de psicología. Así, la simulación implica generar "inputs ficticios" (pretend inputs) para alimentar nuestros propios sistemas de procesamiento emocional o cognitivo. Una distinción clave es entre simulación "off-line" (interna y desvinculada de la acción real) y teorías informacionales basadas en la memoria y la inferencia. Stich y Nichols sitúan la explicación de nuestras prácticas de "folk psychology" en el seno de la ciencia cognitiva donde la estrategia dominante consiste en postular una estructura de conocimiento interna (un cuerpo de reglas, principios o proposiciones) que guía la atribución de estados mentales a otros. A esta posición la denominan "tacit theory" (Theory-Theory), pues ese conocimiento suele ser inconsciente o "tácito". En contraposición, la Simulation Theory (ST) sostiene que, para predecir o explicar la conducta ajena, simulamos internamente nuestro propio sistema de razonamiento, alimentándolo con "inputs ficticios" que representan creencias y deseos del otro; el resultado de esa simulación se transforma luego en la predicción atribuida al sujeto observado.

Aunque los *simulation-theorists* puedan demostrar que no poseemos un "módulo" interno de teoría tácita, ello no deslegitima la existencia de creencias, deseos u otros estados intencionales como entidades causales. La conclusión de Stich y Nichols es que la validez del método de simulación no afecta el estatus ontológico de los contenidos mentales, sino únicamente la arquitectura cognitiva subyacente. Asimismo, el debate sobre el eliminativismo quedaría intacto. Los eliminativistas sostienen que la "folk psychology" –interpretada como la teoría tácita que postula creencias y deseos– es un marco teórico radicalmente falso y, por tanto, debe eliminarse. Pero si nuestra habilidad interpretativa deriva de simulación en lugar de un *theory-theory*, simplemente desaparece el blanco (el *theory-theory*), no las creencias y deseos mismos. Así, la simulación no abandona cualquier tipo de práctica explicativa intencional, sino que redefine cómo la implementamos internamente. La validez de la Simulation Theory (ST) como explicación psicológica es un tema debatido. Stich y Nichols argumentan que la teoría-teoría (TT) explica la cognición penetrable: si nuestro conocimiento es

incompleto, nuestras predicciones serán erróneas, lo que apoya a TT sobre una simulación "off-line" sin acceso a información contextual. Para distinguir ST de TT, proponen experimentos variando el acceso a la información y observando la precisión predictiva, asegurando que el problema no sea la diferencia en los sistemas de razonamiento o la falta de inputs ficticios. Solo un fallo sistemático de predicción, en un escenario donde estos factores están controlados, podría refutar la ST.

Entre los enfoques de modelización, los modelos mentales de Johnson-Laird son cruciales para entender si modelar un proceso equivale a explicarlo. A diferencia de otras teorías centradas en la predicción o atribución de estados mentales, Johnson-Laird postula que los modelos mentales recrean la esencia de un escenario, permitiendo manipularlos internamente para "ejecutar" la dinámica del fenómeno. Un modelo mental exitoso no solo predice, sino que explica la causalidad subyacente. Así, la capacidad de modelar se vincula con la generación de explicaciones causales, la exploración de contrafactuales y la adaptación a nueva información, aspectos esenciales para una explicación psicológica genuina. En contraste, la *Simulation Theory* de Gallese y Goldman, o los modelos de emulación, ofrecen mecanismos predictivos pero no profundizan en la explicación de motivos o arquitectura de razonamiento. Finalmente, cabe apuntar que la IA, especialmente los LLMs, puede ser un instrumento de simulación con gran potencial. Lin (2025) propone ver los LLMs como simuladores psicológicos que construyen "personas" imaginarias con creencias y deseos coherentes. Más allá de sus usos prácticos, los LLMs ofrecen dos aplicaciones en investigación psicológica: simular roles para explorar diversas perspectivas y servir como modelos computacionales para investigar procesos cognitivos. Esta visión, que transforma los LLMs de meros asistentes a instrumentos experimentales para el modelaje cognitivo, abriría una prometedora vía para simular identidades, generar hipótesis sobre el comportamiento y explorar los principios algorítmicos de la mente.

3. EXPLICACIÓN PSICOLÓGICA

La explicación es primordial para saber acerca de cualquier cosa.

Gerald Bakker y Len Clark

3.1. De la explicación científica a la explicación psicológica

La ciencia, como se dice usualmente, no solo describe sino que también explica; no se preocupa solo del qué sino también del *porqué.* Considerar el carácter explicativo como algo específico del conocimiento científico convertiría en científicas afirmaciones que usualmente no tomamos por tales, pues el carácter explicativo, no solo no es exclusivo de la ciencia, sino que tampoco es propio de toda ella. Ni solo la ciencia es explicativa, ni toda la ciencia lo es. Ciertas disciplinas científicas, o partes de ellas, no son explicativas, al menos no lo son *prima facie* (zoológicas o botánicas). ¿Cuándo podemos decir que un fenómeno puede ser explicado científicamente? ¿Existe únicamente un tipo de explicación para todas las diversas ciencias? ¿Acaso no es válido un mismo tipo de explicación para todas? ¿Por qué esto es así? Una parte central del trabajo será la de dar una "explicación" del propio concepto de explicación. Esta tarea, aun siendo titánica y exceder con creces las pretensiones y objetivos de la presente investigación, tiene una gran relevancia si se quiere abordar la cuestión sobre

si los modelos cognitivos basados en simulaciones pueden arrojar luz sobre las cuestiones que han ido interesando a la filosofía de la mente.

Para entender si los modelos cognitivos basados en simulaciones pueden constituir una explicación psicológica, primero cabe desglosar cómo se concibe la explicación en un sentido general, luego aplicar esos modelos a la explicación psicológica, y finalmente ver dónde encajan dichos modelos. La "explicación" es un concepto central en la ciencia y la filosofía, pero no hay una única manera de entenderla. Hempel, en 1965, expuso su modelo de cobertura legal, el cual imperó durante mucho tiempo como modo de explicación científica, siendo esta consiguientemente la concepción estándar. Para Hempel, explicar un fenómeno es poder responder a la pregunta: ¿por qué? Esto es, por qué ocurrió dicho fenómeno. Habitualmente en física, una explicación consistiría en dar cuenta de las leyes de cobertura que de acuerdo con estas se dio el fenómeno. Esto es: $P_1 \ldots P_n$ (*explanans*) → C (*explanandum*). El *explanans* tiene que estar constituido al menos por una ley de cobertura, lo demás pueden ser condiciones iniciales. Asimismo, las leyes de cobertura pueden ser universales o estadísticas. Será, por un lado, un modelo nomológico-deductivo, cuando las leyes de cobertura sean universales, de modo que se deduciría el *explanandum* del *explanans* al ser este segundo consecuencia lógica del primero. Por otro lado, se trata de un modelo estadístico-inductivo cuando las leyes sean estadísticas, siendo el nexo entre el e*xplanans* y el explanandum no de implicación lógica —deductivo— como en el caso anterior, sino inductivo. En todo caso, para que tengamos propiamente una explicación en este segundo modelo, la probabilidad inductiva debe de ser alta. Con todo, para Hempel tenemos un explicación si: (1) el argumento inferencial está correctamente construido sintácticamente —y además, para que la explicación sea correcta, las premisas tienen que serlo—; (2) es posible establecer una simetría sustancial entre la explicación de un fenómeno y su predicción; (3) al ser un modelo general, con los ajustes pertinentes tiene que poder aplicarse satisfactoriamente a todos los casos de las ciencias, no solo naturales sino también sociales. Esto último es posible porque las explicaciones según este modelo son idealizaciones teóricas, son una reconstrucción racional de la práctica efectiva de los científicos. Pero estos requisitos son, según los casos, o muy laxos, o demasiado estrictos. Asimismo, una "buena explicación debe basarse en premisas que sean relevantes para la conclusión, cosa que sin embargo no exige el modelo de cobertura legal" (Amoretti, M. C. y Serpico, D., 2024. pp. 119-120). Debido a estas limitaciones se han propuesto otras posibles caracterizaciones de explicación científica. Uno de los primeros y más destacados críticos fue Wesley Salmon (1984; 1998), el cual argumentó que una explicación no puede guardar la forma de argumento inferencial, pues debe de poder dar cuenta del vínculo de relevancia, causalidad, etc., así como poder explicar eventos singulares o que son poco probables, lo cual el modelo de Hempel no podía. Cabe sin embargo definir el conjunto de enunciados capaces de especificar los factores estadísticamente relevantes para que tenga lugar el fenómeno. De este modo, una probabilidad baja puede constituir una explicación del *explanandum* si es estadísticamente relevante. No obstante, por descontado, una mera correlación estadística no puede constituir una explicación, sino que tiene que haber una relación causal entre *explanans* y *explanandum*.

Explicar es identificar los mecanismos causales conforme a los cuales tiene lugar el fenómeno que se quiere explicar. Si en el modelo de Hempel teníamos una identidad entre explicación y predicción, con Salmon, aunque sea suficiente para la predicción el colegir una red de correlaciones estadísticas con un fenómeno, no lo es para la explicación de este. Asimismo, la idea de Salmon ha

sido retomada por los llamados neomecanicistas (véase Peter Machamer, Darden y Craver, 2000) para los cuales explicar un fenómeno no es solo identificar las causas, sino saber cómo, a partir de dichas causas, se produjo. El *explanans* ya no es una ley de cobertura, condiciones iniciales o conjunto de enunciados, cuanto un mecanismo. La explicación causal-mecanicista sostiene que explicar un fenómeno es identificar las causas y los mecanismos que lo producen. No se trata solo de leyes, sino de desentrañar el cómo funciona un sistema, detallando las interacciones entre sus componentes que dan lugar al fenómeno. Filósofos como Wesley Salmon, Peter Machamer, Lindley Darden y Carl Craver son prominentes en este enfoque. Explicar el funcionamiento de un reloj no es solo decir que siempre marca la hora (ley), sino mostrar cómo los engranajes, muelles y palancas interactúan para mover las manecillas. Mediante la intervención y/o manipulación de los experimentos mediante simulación, podemos representar, llegar a un entendimiento, más o menos certero, de los procesos que están teniendo lugar; la idea subyacente es que, si entendemos el mecanismo, podemos intervenir en él. Otra de las principales formas de explicación, que posteriormente se expondrá más extendidamente, es la funcional, la cual estriba en identificar el papel que tiene el fenómeno aislado puesto en perspectiva con el sistema en su conjunto en el cual se inscribe, y que explicaría su naturaleza, así como, en un sentido genético-histórico, la razón de su surgimiento o emergencia.

Por descontado, existen otros tipos de explicaciones de las ya mencionadas. La explicación unificadora, propuesta por Philip Kitcher, argumenta que una buena explicación es aquella que unifica una amplia gama de fenómenos bajo un pequeño número de patrones o esquemas explicativos. La explicación en este caso no reside en una ley o un mecanismo específico, sino en la capacidad de una teoría para simplificar y organizar nuestro conocimiento. Por ejemplo, la teoría de la evolución explica una vasta diversidad de fenómenos biológicos (adaptación, diversidad de especies, fósiles) con un conjunto limitado de principios (selección natural, herencia). Asimismo, cabe mencionar la explicación "intervencionista" definida por Woodward en *Making things happen: a theory of causal explanation* (2003) en términos de variables manipulables. X explica Y si intervenir sobre X produce cambios en Y de modo estable y predecible. (Cabe señalar que este tipo de explicación encajaría con el tipo de idealización minimalista caracterizada por Weisberg que vimos en el apartado segundo). Asimismo, en esta misma línea encontramos la explicación de Lipton en *Inference to the Best Explanation* (1991), para el cual una buena explicación identifica el factor que, de entre muchos posibles, hace una diferencia en el evento; siendo entonces una razón verdadera relevante para el fenómeno, haciéndose una inferencia a la mejor explicación teniendo en cuenta que elegimos la mejor explicación según una elección que es a su vez condicionada por varios factores científicos o paracientíficos. Esto nos lleva a explicaciones cuyo carácter es más pragmatista o contextualista. Autores como Bas van Fraassen señalan que las explicaciones no son verdades absolutas, sino que son sensibles al contexto, a los intereses de quien pregunta y al conocimiento disponible; esto es: lo que cuenta como una "buena explicación" puede variar. Este tipo de explicación tiene la ventaja de poner el foco en la relevancia y la utilidad en un contexto específico, lo cual nos pone ya sobre la pista del tipo de explicación que requiere la psicología y que puede ser distinta a la que es relevante en otras ciencias y disciplinas e, incluso, en cómo para diferentes fenómenos psicológicos puede ser más adecuado un tipo de explicación u otra, sin que tenga, a priori, que imponerse una sobre otra. También puede uno encontrarse con la explicación multinivel –y multimetodológica– que considera que los fenómenos complejos (como el cambio climático) requieren modelos acoplados en diferentes niveles (atmósfera, océano, biosfera, etc.), así como de métodos (ecuaciones diferenciales, simulaciones por

ensemble) para conseguir capturar la complejidad no lineal, también resulta pertinente de considerar para el estudio de sistemas complejos como la mente humana. Por último, cabe mencionar la caracterización de explicación que realiza Daniel Little —inspirándose en Mackie— en su libro *Microfoundations, Method and Causation* (1998). Little adopta el modelo de causalidad INUS de Mackie (a saber: *Insufficient but Necessary part of a condition which is itself Unnecessary but Sufficient*). Según este modelo, una causa no es una condición única y suficiente para un efecto, sino que es parte de un conjunto de condiciones que, en su totalidad, sí es suficiente para producirlo. Sin embargo, no es el único conjunto posible que puede causarlo —en cierto modo, es innecesario como conjunto, pues pueden existir otros, luego no es la única vía posible para que el efecto ocurra. Es decir, una causa es una parte necesaria de una condición suficiente, aunque esa condición suficiente no sea la única forma de producir el resultado. El problema de esta propuesta es que, aun sugerente, ya no es que exija un "cierre" de todas las posibles causas, sino más bien la dificultad que presenta es identificar exhaustivamente todas las condiciones necesarias y suficientes en sistemas sociales complejos, lo cual, al menos por ahora, resulta inviable. A pesar de estas dificultades, es un tipo de explicación que encajaría en varios puntos con el tipo de explicación que puede darse en los modelos cognitivos basados en simulaciones, pues, como se verá en el apartado cuarto, lo que nos muestran dichas simulaciones es la capacidad de llegar a un mismo resultado mediante caminos distintos —defendiéndose, consiguientemente, el funcionalismo y su tesis de la realizabilidad múltiple.

Una vez hecho un repaso por las principales caracterizaciones que han tenido lugar de explicaciones científicas, cabe aplicarlas ahora al dominio de la psicología. Para el caso de la explicación hempeliana, una explicación psicológica implicaría la identificación de leyes psicológicas generales que conectan estímulos con respuestas, estados mentales con acciones, o unas propiedades mentales con otras. Por ejemplo, "si una persona tiene el deseo X y la creencia Y, entonces realizará la acción Z". Asimismo, la explicación psicológica como identificación de mecanismos cognitivos/neurales, visión dominante en la ciencia cognitiva contemporánea, se centra, no en decir qué sucede, sino en cómo sucede. Este tipo de explicación implica desentrañar los mecanismos cognitivos (procesos de procesamiento de información, algoritmos) y/o los mecanismos neurales (redes neuronales, actividad cerebral) que subyacen a un comportamiento o una experiencia mental. Por ejemplo, explicando el reconocimiento de rostros detallando las etapas del procesamiento visual, las regiones cerebrales implicadas y las interconexiones entre ellas. Además, existe la visión de la explicación psicológica como atribución de estados mentales (*folk psychology*), siendo esta la forma más cotidiana y fundamental de explicación psicológica. Explicamos el comportamiento de alguien atribuyéndole creencias, deseos, intenciones, miedos, etc. (p. ej.: "Daniel fue a la nevera porque quería una cerveza y creía que había una"). El terreno de la psicología popular involucraría la comprensión interpersonal y la predicción del comportamiento en la vida diaria. La Teoría de la Simulación y la Theory-Theory, que se expusieron en el apartado anterior, comprenden que el ser humano es, ante todo, una máquina inferencial. Nuestra forma de vivir con los demás, comunicarnos con ellos, comprenderlos, etc., es mediante la atribución de estados mentales a terceros. Lo importante no es tanto acertar en dicha atribución, cuando en simplemente llevarla a cabo. Una de las líneas más interesantes a este respecto sería la de analizar la necesidad de dicha inferencia y atribución de estados mentales en la pragmática del lenguaje, así como en el estudio de la mente humana.

¿Cuál es la naturaleza de la explicación en psicología? A diferencia de las ciencias físicas, donde la explicación suele fundarse en leyes generales o en mecanismos causales observables, la psicología trabaja con fenómenos intencionales, muchas veces no directamente accesibles, y con procesos internos complejos. Esta situación plantea interrogantes particulares cuando se trata de evaluar el estatus explicativo de los modelos cognitivos basados en simulaciones: ¿pueden estas simulaciones ser consideradas explicaciones en sentido fuerte, o se trata meramente de descripciones funcionales sin correlato causal? En el ámbito de la psicología científica, se reconoce que una explicación válida debe cumplir con ciertos requisitos mínimos. Entre ellos se encuentran la generalidad (la capacidad de un modelo para explicar múltiples fenómenos similares), la capacidad predictiva, la consistencia interna, y, fundamentalmente, la referencia a mecanismos causales. Sin embargo, estos criterios deben matizarse debido a la especificidad del objeto de estudio: la mente humana. La psicología opera en varios niveles explicativos, desde el nivel neurobiológico hasta el funcional y el fenomenológico. Esta diversidad hace que la explicación psicológica se aleje del ideal clásico de las ciencias naturales, y se aproxime a un modelo más pluralista y pragmático (véase el llevado a cabo por McCauley en 1986 sobre el ajuste psicosocial de las mujeres adultas con síndrome de Turner. En este contexto, una explicación puede ser aceptable aun cuando no se refiera directamente a estructuras físicas observables, siempre que articule de forma coherente las funciones cognitivas implicadas y permita alguna forma de contraste empírico. No obstante, esta apertura no elimina la exigencia de que las explicaciones psicológicas aporten algo más que correlaciones: se espera que proporcionen una *comprensión* de los procesos subyacentes que generan el comportamiento o los estados mentales observados. Esta exigencia remite a dos grandes tradiciones explicativas: la explicación funcional y la explicación mecanicista. Implica la necesidad de dar cuenta de procesos internos no directamente observables, reconociendo tanto la importancia del nivel funcional —lo que hace el sistema— como la del nivel mecanicista —cómo lo hace.

¿Explicamos *cómo funciona* la mente o *por qué* funciona como lo hace? Uno de los enfoques dominantes en psicología cognitiva es el análisis funcional, propuesto por Robert Cummins (1983). Según esta perspectiva, explicar un fenómeno psicológico equivale a identificar la función que desempeña dentro del sistema. Por ejemplo, explicar la memoria de trabajo implica describir cómo contribuye al procesamiento general de la información, sin necesidad de especificar los mecanismos neurobiológicos subyacentes. El análisis funcional se centra en lo que hace el sistema, más que en cómo lo hace. Este enfoque es particularmente compatible con la modelización computacional y con los modelos basados en simulaciones, ya que permite representar funciones cognitivas sin necesidad de compromiso con estructuras físicas específicas. Una simulación que reproduce correctamente una tarea cognitiva como resolver un problema lógico, reconocer patrones, o aprender asociaciones, puede ser considerada, en este marco, una forma válida de explicación funcional. Mientras que la explicación mecanicista antes mencionada considera que una explicación científica debe identificar los mecanismos causales que producen un fenómeno; exigiendo, consiguientemente, que se especifiquen las entidades y actividades (organizadas de un modo determinado) que dan lugar al comportamiento observado. Una simulación solo podría ser explicativa si reproduce los mecanismos reales que operan en el sistema cognitivo humano, pero como se verá en el subapartado 4.1, los problemas de validez experimental de las simulaciones como su "infradeterminación" dificultan que mediante estos se pueda realizar una explicación mecanicista, si bien una explicación funcional podría darse, pero en detrimento de que uno pueda con dicha explicación funcional desentrañar la naturaleza de la mente

humana en su completitud y riqueza. Aunque analizar un sistema según las funciones que realiza (input → output) puede resultar ventajoso y no muy problemático —por ejemplo, explicar la memoria como función de almacenamiento y recuperación–, no es tan fácil para fenómenos complejos como la cognición, más aún si se trata de fenómenos psicológicos subjetivos, aquellos que son vividos en primera persona. Además, como la explicación funcional es independiente del sustrato físico, es compatible con la simulación, lo que la hace resultar demasiado vaga, precisamente por general (es susceptible de aplicación independientemente del sustrato físico del sistema), pudiendo ser necesario su conexión con mecanismos específicos si la explicación psicológica pretende explicar la psique humana (y no solo replicar un resultado).

Algunos autores, como Marr, abogan por una explicación de niveles, seguramente motivado por la insuficiencia de una mono-explicación —esto es: una explicación que se centra en solo un aspecto—. Marr se basa en tres niveles: computacional, algorítmico/representacional e implementacional. En el artículo "Vision: A Computational Investigation into the Human Representation and Processing of Visual Information" (2010), Marr sostenía que para entender un sistema complejo como la visión, uno debe especificar primero la computación (el problema que se resuelve), luego el algoritmo (los pasos para resolverlo), y finalmente la implementación (cómo se realiza en el hardware neural). Las simulaciones operan fuertemente en los dos primeros niveles, ofreciendo explicaciones mecanicistas de "cómo" se realiza una tarea cognitiva. De igual forma, el tercer presupuesto que encontramos es el de la generabilidad y falsabilidad (proveniente de la metodología científica). Como cualquier teoría científica, un modelo de simulación busca ser generalizable (aplicable a diversas situaciones) y falsable (capaz de ser refutado por la evidencia empírica). Las simulaciones no solo producen datos; generan predicciones sobre cómo se comportaría el sistema cognitivo bajo nuevas condiciones. Un modelo computacional es una teoría que puede *ser ejecutada*, pudiéndose revelar sus implicaciones y predicciones de una manera que las teorías abstractas no pueden, por lo que aquello que se afirma en dichos modelos puede constituirse, con una mayor legitimidad, en una explicación. En este sentido, la historia de la caracterización de la explicación científica guarda, ciertamente, semejanza con la del criterio de demarcación científica. No por menos, en ambas se juega la capacidad de distinción entre la ciencia y la llamada "pseudociencia". En el presente parece haberse abandonado en este segundo caso la pretensión de encontrar aquella fórmula única que permitiera determinar qué es científico y qué no lo es. Pasando de criterios únicos hasta propuestas más plurales como la que hace Mario Bunge. Este trabajo pretende insertarse en esa tendencia que existe para el caso de la demarcación científica; realizando una comprensión más plural de la explicación, en este caso, en la psicología.

3.2. Jerry Fodor, la lógica de la simulación y la explicación funcional

Jerry Fodor aborda la relación entre la simulación computacional y la explicación psicológica (Fodor, 1968) argumentando que la simulación de la conducta no equivale, por sí misma, a una explicación de esta. Para que una simulación sea explicativa, debe cumplir con dos condiciones fundamentales. La primera es la equivalencia débil, que postula que el repertorio de comportamientos posibles de la máquina debe ser idéntico, en los aspectos teóricamente relevantes, al del organismo que se simula. Este repertorio de comportamientos no se limita a la conducta observable, sino que incluye el conjunto de todas las acciones que el organismo podría realizar en diferentes circunstancias, lo que requiere que

la simulación sea capaz de dar cuenta de contrafactuales. La segunda condición es la equivalencia fuerte, que exige que los procesos internos de la máquina que subyacen a su comportamiento sean del mismo tipo que los procesos psicológicos del organismo. Para Fodor, la mera correspondencia entre la conducta observable de una máquina y un organismo es insuficiente. El autor pone de ejemplo un fonógrafo que reproduce la voz humana, el cual, aunque simula exitosamente el habla, no ofrece una explicación de la producción del lenguaje porque no realiza las operaciones mentales que subyacen a la capacidad de hablar; rechazando de este modo pruebas como el juego de Turing, que solo se basan en la capacidad de la máquina para simular la conversación, por considerarlas insuficientes para demostrar una verdadera capacidad de pensamiento.

La simulación debe, por tanto, ser capaz de dar cuenta de contrafactuales; si el organismo se comportase de cierta manera en una situación no observada, la máquina también debería hacerlo. Esta capacidad de proyectar a partir de datos observados a un espacio de comportamientos posibles es lo que dota a la simulación de un estatus similar al de una teoría científica. Una simulación logra la equivalencia fuerte cuando los procesos internos de la máquina que generan el comportamiento son del mismo tipo que los procesos psicológicos que subyacen al comportamiento del organismo. Esto significa que la explicación no solo debe acertar en lo que el organismo hace y puede hacer, sino también en cómo lo hace. Fodor lo ilustra con el ejemplo de una máquina que prueba teoremas del cálculo proposicional utilizando tablas de verdad. Si un matemático humano usa heurísticas distintas para probar los mismos teoremas, la máquina, aunque logra la equivalencia débil (ambos pueden probar los mismos teoremas), falla en la equivalencia fuerte porque los procesos subyacentes son diferentes.

La equivalencia fuerte depende de la noción de equivalencia funcional, que es la correspondencia entre los estados internos de la máquina y los estados psicológicos del organismo. Esta equivalencia no se define por la similitud física de los componentes (por ejemplo, carbono frente a silicio), sino por el papel que cada estado desempeña en el sistema computacional o psicológico. La elucidación de estas relaciones, así como la individuación de los procesos psicológicos, no puede hacerse a priori, sino que surge del desarrollo de teorías psicológicas empíricamente exitosas. Fodor concluye que la cuestión de si una máquina podría pensar, sentir o tener otros estados mentales es, en parte, una cuestión de hecho empírica (si se cumple la equivalencia fuerte) y, en parte, una cuestión lingüística. Si la simulación cumple las exigencias metodológicas de una teoría científica, la atribución de estados mentales a la máquina sería justificable, aunque no obligatoria, ya que podríamos decidir, por razones pragmáticas, reservar dichos términos para los seres humanos. Fodor aclara que la similitud requerida no es de naturaleza física (un cerebro no tiene que ser de silicio ni un ordenador de carne y hueso), sino de tipo funcional. Dos procesos son funcionalmente equivalentes si tienen el mismo papel en el sistema; es decir, si sus relaciones causales con otros estados internos y con la conducta observable son idénticas. El autor subraya que esta equivalencia funcional es el objetivo de la psicología teórica. Las teorías psicológicas, al postular entidades y procesos, determinan implícitamente las condiciones para la equivalencia funcional de esos estados. Por lo tanto, el éxito de una simulación como explicación funcional depende del grado en que la teoría psicológica que implementa el programa de la máquina logre una correspondencia fina y detallada con la estructura causal de la mente del organismo. Una simulación puede ser más o menos explicativa dependiendo de

la precisión con la que refleje las relaciones funcionales entre los estados mentales y la conducta, lo que implica que la explicación funcional es un concepto que admite grados.

Que dos máquinas puedan resolver un problema de la misma manera aparente, pero la secuencia de operaciones lógicas o los "costos" computacionales (como la complejidad o el tiempo) puedan diferir, revelando que los procesos internos no son idénticos (véase el caso de la IA no explicable), sugiere que la noción de equivalencia funcional puede ser, en la práctica, un ideal regulativo más que una condición plenamente verificable. Esta limitación se acentúa al considerar la noción de función misma, pues en biología y filosofía de la biología la función de una característica o de un proceso es típicamente entendida en términos de su papel causal en el mantenimiento o la reproducción de un sistema. Existen dos enfoques principales: el etiológico y el de la teoría del rol causal. El enfoque etiológico, o histórico, sostiene que la función de una característica es lo que esta hizo en el pasado para ser seleccionada (por la selección natural, por ejemplo). El enfoque del rol causal define la función por el papel que la característica desempeña en el sistema actual. Ambas visiones presentan problemas cuando se aplican a la mente. Una explicación psicológica meramente funcional podría describir el rol causal de un proceso (p. ej., "el proceso P tiene la función de integrar información visual para la acción"), pero no diría nada sobre su origen histórico o evolutivo, que podría ser crucial para una comprensión completa. El "porqué" de la existencia de un proceso, no solo su "qué hace", podría ser una parte esencial de la explicación.

Finalmente, la visión funcionalista corre el riesgo de caer en un reduccionismo que ignora la importancia de los niveles de descripción. Aunque Fodor reconoce la importancia de los niveles de descripción, la crítica posfodoriana a menudo señala que la explicación psicológica puede requerir una articulación compleja entre los niveles funcional, neurobiológico y social. Una explicación que se centre únicamente en el isomorfismo funcional podría ignorar los detalles neurofisiológicos que son causales en sí mismos, o el papel de la cultura y la interacción social en la formación de los procesos psicológicos. Por lo tanto, una explicación meramente funcional podría ser insuficiente porque no logra dar cuenta de la totalidad de las causas que dan lugar a la conducta, desde las bases biológicas hasta el contexto social y cultural. En este sentido, la "equivalencia fuerte" de Fodor, si bien es un paso necesario para una explicación, podría no ser el último paso, ya que la noción de función que utiliza podría no ser lo suficientemente robusta para capturar la complejidad de la mente en su totalidad.

Los problemas que presenta una explicación psicológica meramente funcional, como la incapacidad para dar cuenta de la intencionalidad, la conciencia o la etimología de los procesos, pueden ser abordados de manera más efectiva por una explicación psicológica de carácter normativo. A diferencia de la aproximación puramente funcional que se limita a describir cómo un proceso opera (su rol causal), una explicación normativa se enfoca en por qué un proceso debe operar de una manera específica, evaluando su funcionamiento a la luz de estándares de corrección, racionalidad, verdad o éxito.

3.3. Predicción, explicación y comprensión

¿Los modelos simulativos pueden capturar lo que en última instancia es un fenómeno vivido, subjetivo e intencional? Esto es, ¿pueden explicar la mente humana en sentido completo? Como se vio, la noción

de explicación científica en Hempel presupone la asimetría entre predicción y explicación. Esta asimetría resulta problemática. Realizar predicciones que resultan ser confirmadas empíricamente no constituye, por sí solo, una explicación científica, si acaso la teoría o modelo que lo apoya no se acepta si quiera como científica (véase el caso histórico de Velikovski). Las predicciones exitosas suelen resultar en la aceptación de la teoría expuesta, pero no todas las predicciones poseen el mismo grado de especificidad e, igualmente, no en todas las disciplinas resulta posible comprobar empírica o experimentalmente dichas predicciones. La fórmula según la cual nos dice que ante una predicción correcta nos encontramos ante una teoría correcta no es tan sencilla, ni se cumple en todos los casos.

¿Puede la ciencia explicar todo? ¿puede explicar cualquier cosa? Una de las vías de explicación psicológica es la de la *reducción.* El filósofo de la ciencia antes mencionado, Hempel, comprendía que la psicología tenía que poderse reducir, esto es, traducir sus proposiciones, a proposiciones sobre la conducta física de los seres humanos. Su objetivo, compartido con el conductismo lógico, es hacer de la psicología una disciplina científica, por lo que tenía que eliminarse todo elemento de carácter subjetivo, "en primera persona" como diría Dennet, todo aquello que no pudiera ser observado públicamente. No podemos tener acceso a los estados mentales internos de los otros, pero igualmente, como se veía cuando se expuso la teoría de la simulación y la "teoría-teoría", esto es, la teoría de la teoría de la mente (ToM, por sus siglas en inglés, *Theory of Mind*), podemos hacer atribuciones de estados mentales sirviéndonos únicamente de la conducta que observamos de ellos.

Wilhelm Dilthey, autor de la *Introducción a las Ciencias del Espíritu* y de la que es considerada la primera carta anti-cientificista, considera que solo es posible acceder mediante la introspección al carácter subjetivo, privado y experiencial de los fenómenos psíquicos. Al ser los estados mentales, intrínsecamente, significativos, no es posible un acceso científico a ellos, sino meramente por medio de la comprensión, de la *verstehen*. Como otros fenomenólogos teorizarían después –véase Edmund Husserl o Edith Stein–, el concepto principal que opera aquí y que lleva a una comprensión del otro (de su dolor, por ejemplo) es la empatía. El hiato que se cierne entre nosotros y ellos se recorta mediante el salto empático de la imaginación (la comprensión). Para que la pretensión de hacer de la psicología una ciencia, del modo de las ciencias naturales, estas teorías y sus proposiciones tienen que considerase un sin sentido. Pero, desgraciadamente, que sean tildadas de sinsentido porque se considere, como Hempel, que el significado de un enunciado queda establecido por las condiciones de su verificación, no elimina que tengan un sentido –ya sea histórico, cultural, personal, etc.–. Quizá las cuestiones a las que hacen alusión los fenomenólogos no debería de formar parte de la disciplina de la psicología científica, pero si se considera que la psicología engloba el estudio de toda la vida psíquica, aquellos fenómenos psicológicos que pueden ser explicados científicamente, mediante el punto de vista de la tercera persona, no serán todos los realmente existentes. La fenomenología rechaza que el problema de las otras mentes pueda resolverse simplemente por vía de inferencias a partir de lo observable (conducta, neurobiología, lenguaje), pues ninguna de esas mediaciones captura plenamente la vivencia inmediata. No obstante, la fenomenología no se resigna a una incomunicabilidad radical entre conciencias. En lugar de asumir una brecha ontológica o epistémica insalvable, Edmund Husserl y la tradición fenomenológica sostienen que el otro aparece en el mundo como un cuerpo viviente (*Leib*), portador de expresividad y sentido. Así, el conocimiento del otro no es inferido desde fuera como una hipótesis, sino que es co-constituido fenomenológicamente en la experiencia perceptiva y empática. En obras como la Crisis de las ciencias europeas o Meditaciones cartesianas, Husserl plantea que la

experiencia del otro supone una forma de accesibilidad verificable de lo que es originalmente inaccesible: no accedemos a las vivencias del otro como accedemos a las propias, pero tampoco nos son opacas por completo. El comportamiento corporal, las expresiones faciales, la entonación no son simples datos físicos, sino indicadores encarnados de una vida subjetiva. Esta comprensión es posible gracias a un proceso de empatía (*Einfühlung*), que no es una inferencia ni una proyección, sino una forma primaria de aprehensión intersubjetiva. Sin embargo, el enfoque fenomenológico clásico presenta varios problemas. En este sentido hereda de la tradición cartesiana la idea de que los contenidos de la conciencia son dados con certeza, incuestionables desde la primera persona. Kripke, en su interpretación de la gramática del dolor y del seguimiento de reglas de Wittgenstein, considera que este, más que abogar por una perspectiva del conductismo lógico, como se le suele atribuir, que solo atendería a la mera expresión del dolor, parece recuperar esa noción de *empatía* que es clave en las perspectivas fenomenológicas:

> El método de Wittgenstein en su debate del problema de las otras mentes es paralelo a su método en el debate de las reglas y el lenguaje privado [...] propone una paradoja escéptica. Aquí la paradoja es el solipsismo: la mera noción de que podría haber mentes distintas de la mía, con sus propias sensaciones y pensamientos, parece carecer de sentido [...] Wittgenstein da una solución escéptica, arguyendo que cuando la gente usa realmente expresiones que atribuyen sensaciones a otros no pretende realmente hacer ninguna aserción cuya inteligibilidad sea socavada por el escéptico (solipsista). [...] Una vez más, la interpretación correcta de nuestro discurso normal envuelve una cierta inversión: no nos compadecemos de otros porque les atribuyamos dolor, atribuimos dolor a otros porque nos compadecemos de ellos. (Más exactamente, se revela que nuestra actitud es una actitud hacia otras mentes en virtud de nuestra compasión y actitudes relacionadas). (Kripke, 2006, p. 151).

La investigación psicológica contemporánea se encuentra en una encrucijada epistemológica: por un lado, busca ofrecer explicaciones rigurosas, formales y empíricamente contrastables de los procesos mentales; por otro, enfrenta la dificultad de capturar la dimensión subjetiva, intencional y vivida de la mente humana. Esta tensión se manifiesta especialmente al evaluar el papel de las simulaciones como modelos explicativos: ¿pueden realmente explicar la mente, o solo reproducir ciertos aspectos funcionales o comportamentales? ¿Qué diferencia hay entre *explicar* una mente y *comprender* su experiencia? La tradición filosófica ha diferenciado clásicamente entre *Erklären* (explicación) y *Verstehen* (comprensión); si la explicación busca identificar leyes causales o mecanismos observables, la comprensión se orienta a captar el sentido o significado de una acción o experiencia desde el punto de vista del agente. Si defendemos, como señala Thomas Nagel en su conocido ensayo *What Is It Like to Be a Bat?* (1974), que hay un "cómo es" que constituye la esencia de la experiencia consciente que no puede capturarse en descripciones objetivas externas, esto plantea una limitación estructural para la simulación, pues esta no puede simular el punto de vista en primera persona, al menos no sin caer en una atribución proyectiva no suficientemente fundamentada.

Este debate o conflicto de modos distintos de aproximación a lo mental parece ser una constante en la historia del pensamiento. Esta divergencia, así como el motivo principal del presente trabajo, se puede vislumbrar claramente con el siguiente caso. En 1965 hubo una correspondencia entre Giuliano y Simmons, en donde se puede ver la diferencia de oposiciones, de ferviente actualidad una vez se ha

materializado el sueño de una inteligencia artificial capaz de entender el lenguaje natural e interactuar con nosotros a través de este. Giuliano tenía la impresión de que "con la prisa por demostrar que se puede llevar a cabo una comunicación interactiva con el ordenador, se ha perdido de vista […] que aún queda demasiado que aprender sobre el lenguaje" (McCorduck, 1991, p. 257). Para él, cualquier tipo de demostración resulta estéril si lo es también el conocimiento que tenemos del lenguaje que está entrando en juego. Simmons, por el contrario, tenía una idea diferente de cómo funciona la ciencia. Como recoge Pamela McCorduck en su libro *Máquinas que piensan* (1991):

> Nosotros aplicamos impacientes la poca teoría de la que se dispone en el campo de la lingüística y lógica, escribió Simmons, pero a menudo la teoría se queda atrasada con respecto a la construcción de modelos y a veces se deriva de ella. No somos alquimistas en busca de un elixir de la vida o de la piedra filosofal; somos científicos que acumulan conocimiento mediante la más estricta experimentación: construir modelos pequeños y muy complejos y poner a prueba sus límites. (*Ib.*)

4. EVALUACIÓN DE LA VALIDEZ DE LOS MODELOS COGNITIVOS BASADOS EN SIMULACIONES COMO EXPLICACIÓN PSICOLÓGICA

Para comprender el modelo de Turing del cerebro, fue crucial darse cuenta de que [...] todo lo que hacía el cerebro, lo hacía en virtud de su estructura como sistema lógico, y no porque estuviera dentro de la cabeza de una persona, o porque fuera un tejido esponjoso compuesto por un determinado tipo de formación biológica de células. Y si esto fuera así, entonces su estructura lógica podría igualmente venir representada por algún otro medio, incorporada mediante alguna otra maquinaria física. Era una visión materialista de la mente, pero una visión que no confundía los patrones lógicos y las relaciones con las sustancias físicas y con las cosas, como solía hacer la gente tan frecuentemente.

Johnson-Laid

4.1. Conductismo lógico y validez experimental de las simulaciones

La teoría del conductismo lógico sostiene que hallarse en un estado mental es hallarse en un estado de conducta o de comportamiento (Priest, 1994, p. 55). Así, todos los estados mentales para esta teoría, o bien son conductas, o bien disposiciones para llevarlas a cabo. El objeto de la psicología sería, entonces, meramente la conducta, comprendida como acción corporal públicamente observable. Incluso aquellos estados mentales que no parecen tener una acción corporal asociada pueden traducirse en –o, incluso, retrotraerse deductivamente a– una conducta. Así, puede verse que los presupuestos epistemológicos del conductismo lógico parecen colegir con los del empirismo más radical, pues solo poseen significado aquellos enunciados que son derivados de aquellos observacionales autoevidentes. Todo el lenguaje de corte mentalista que no tuviera propiamente un correlato observacional se rechaza por no tener significado, carecer de sentido alguno. Esta postura, a pesar de tener más matices de los

que aquí se ha expuesto, ha sido objeto de varias críticas, como aquella que realizase Jawoski respecto de que la propia teoría del conductismo lógico, para funcionar, ser operativa, se apoya implícitamente en conceptos como mente, representación mental, estado mental, intención, etc. Más aún, puede añadírsele la crítica de que si el lenguaje es comprendido como conducta verbal, la conducta verbal del lenguaje de corte mentalista, o bien no puede ser tildado de sin sentido, o bien se acepta que el lenguaje, así como otros procesos cognitivos, no pueden definirse meramente como conducta observable, al menos no sin una gran reducción por el camino del propio fenómeno que pretende estudiarse. Dentro de los debates entre los conductistas lógicos, existen desacuerdos sobre si acaso pudiera medirse la verdad o falsedad de las proposiciones psicológicas si estas no fueran sobre la conducta; que los conceptos psicológicos, si no dispusiéramos de criterios públicos –de corrección–, no tendrían propiamente significado. Pero, de lo que no existe debate dentro de la teoría del conductismo lógico es que, si nuestro vocabulario psicológico no refiere a la conducta, entonces no refería a *nada*. El problema de esta tesis compartida estriba en qué parte del mundo psíquico refiere a la conducta, y si es suficiente para explicar, precisamente, la conducta humana.

¿Cómo puedo saber que otras personas tienen mente? Mientras que cada uno experimenta su propia mente de forma directa, el acceso a las mentes ajenas se da solo a través de la observación del comportamiento, produciéndose la llamada asimetría de conocimiento. Sin embargo, este comportamiento podría estar causado por factores distintos a estados mentales, lo que hace que la mente del otro quede siempre, en cierto modo, velada. Dos formulaciones escépticas refuerzan esta duda. Por un lado, el problema del error: si no hay forma de distinguir, en términos de experiencia, entre un caso bueno (una creencia verdadera) y uno malo (una creencia falsa), entonces no hay base suficiente para saber cuál de los dos se da. El "problema del error" se suele ejemplificar con el caso de Verity, quien descubre que su pareja ha estado profundamente infeliz a pesar de años de aparente armonía. Esto ilustraría con crudeza esa distancia infranqueable entre lo que parece y lo que realmente ocurre en la mente ajena. Si fenoménicamente la experiencia de Verity, tanto si su pareja hubiera sido feliz como infeliz, son indistinguibles, la evidencia que tenía Verity para ambos casos es la misma. Por otro lado, el problema de las fuentes *–the problem of sources–* sostiene que, si puedo conocer algo, debe haber un medio para ello, y no existe tal medio para acceder directamente a los estados mentales de los otros. Una respuesta tradicional ha sido el argumento por analogía: se observa que en uno mismo ciertos estados corporales causan estados mentales, y estos, a su vez, dan lugar a ciertas conductas. Al ver en los demás cuerpos y conductas similares, se infiere la existencia de estados mentales análogos. Pero este argumento enfrenta múltiples objeciones. Primero, se trata de una generalización apresurada desde un único caso: el mío. ¿Acaso soy yo un ejemplo paradigmático? ¿O relevante para realizar una inferencia de solo un único caso? Segundo, incluso si se aceptara esta inferencia, la correlación entre conducta y mente no siempre se sostiene: existen conductas complejas sin mente (véase inteligencias artificiales como los LLMs), y estados mentales sin conducta (como ocurre en los sueños mientras dormimos –y no somos sonámbulos). Además, la autoobservación de nuestra conducta es mediada a su vez por modelos teóricos, lo que debilita la idea de una observación directa de la propia mente que pueda luego servir de base para analogías. Así, el problema de otras mentes remite de nuevo al problema de nuestra mente, del acceso a ella. Descartes comprendía que de lo que no podía dudar era de que él pensaba –y que, como pensaba, existía–, pero algunos filósofos han considerado que esto ya es mucho suponer. ¿Quién es ese "yo" que piensa? ¿No tendríamos que hablar más bien de "lo pensado", del mismo modo que no podemos, según Wittgenstein, hablar de mi dolor, sino de doler?

Para que un término tenga significado, debe poder ser usado conforme a reglas públicas, reglas que puedan ser comprendidas, compartidas y corregidas por otros hablantes. Si el uso de un término dependiera únicamente de experiencias privadas, inaccesibles a cualquier otro, entonces no habría manera de verificar su uso ni de enseñarlo, corregirlo o acordar su sentido. En este marco, los términos mentales —como "dolor", "creencia", "deseo" o "felicidad"— no pueden depender de una supuesta experiencia interna y privada, sino que adquieren su significado a partir de la conducta observable que habitualmente los acompaña. Así, no es que la conducta simplemente "exprese" el estado mental, sino que constituye el criterio de su posesión: estar "en dolor" es comportarse de una cierta forma (como gritar, llorar, frotarse una herida o retirarse de un estímulo). Véase la famosa analogía de la "caja del escarabajo" de Wittgenstein. Supongamos que cada persona tiene una caja y llama "escarabajo" a lo que hay dentro, pero nadie puede mirar en la caja de otro. Si eso es así, entonces el contenido real de la caja se vuelve irrelevante para el uso del término "escarabajo". Lo mismo ocurre con los términos mentales: si su sentido dependiera de un contenido mental interno, accesible solo al sujeto, entonces ese contenido sería irrelevante para la función del término en el lenguaje. Así, hablar de "dolor" o "felicidad" no implica referirse a algo oculto en una mente privada, sino a patrones observables y públicamente reconocibles de comportamiento. Desde esta perspectiva, la pregunta por si los otros tienen mente se vuelve sin sentido. Si decir que alguien tiene mente significa que manifiesta ciertos tipos de conducta inteligible en términos mentales, y si efectivamente los demás se comportan así, entonces podemos decir que tienen mente. El problema de otras mentes no se plantea porque no hay una distancia metafísica que salvar entre la conducta y el estado mental: el estado mental simplemente es una forma de describir y organizar cierto comportamiento.

Ahora bien, esta postura conductista enfrenta críticas importantes. La más contundente es que parece negar lo obvio: que hay algo más en el dolor que el mero "comportamiento de dolor". La experiencia subjetiva, la cualidad vivida del dolor, parece no agotarse en la conducta asociada. Yo no necesito mirarme llorando para saber que estoy sufriendo: el dolor se me impone de forma inmediata, como una vivencia directa, interna. Esto parece darme un acceso privilegiado a mis propios estados mentales, un tipo de conocimiento de primera persona que no se obtiene por observación externa. No obstante, para responder al problema de otras mentes, quizá no haga falta una teoría completa de lo mental. Basta con mostrar que algunos estados mentales pueden ser entendidos en términos conductuales. Si podemos establecer que ciertos patrones de conducta permiten inferir ciertos estados mentales con suficiente fiabilidad, entonces el escepticismo radical sobre las otras mentes pierde fuerza. El conductismo, en este sentido, podría no ser una teoría satisfactoria del todo de la mente, pero sí una vía útil para enfrentar el problema específico del conocimiento de otras mentes. En parte resulta insatisfactoria pues nuestra experiencia del dolor parece ir más allá de la conducta que lo expresa. Por ello, otra alternativa consiste en entender los estados mentales como entidades teóricas que nos permiten explicar y predecir la conducta de los demás. Así como en ciencia postulamos entidades no observables —como electrones o campos gravitatorios— porque mejoran nuestras teorías también postulamos creencias, deseos o emociones porque forman parte de nuestras mejores explicaciones del comportamiento humano. Desde esta perspectiva, hablar de mentes ajenas no es referirse a un misterio inaccesible, sino a un conjunto de herramientas explicativas que nos permiten vivir y actuar en un mundo social inteligible.

Respecto de la validación de una simulación, esta consiste en determinar en qué medida los resultados que produce el modelo son representativos de los fenómenos psicológicos reales. Existen distintos tipos de validez que pueden aplicarse a este propósito, una lista rapsódica de la misma sería: (i) Interna: refiere a la consistencia lógica y matemática del modelo. Un modelo con alta validez interna no contiene contradicciones ni errores de implementación; (ii) Externa: se refiere a la capacidad del modelo para generalizar sus resultados más allá de las condiciones de entrenamiento; (iii) Constructivista: implica que los elementos del modelo (variables, estructuras, procesos) representan constructos teóricos reales del dominio considerado (psicológico, en este caso); y (iv) Predictiva: evalúa si el modelo puede anticipar correctamente fenómenos no incluidos originalmente en su diseño. Aunque existan de hecho diversas estrategias metodológicas de sobra conocidas para la evaluación de estas formas de validez –confrontar los *outputs* del modelo con datos experimentales humanos; evaluar si distintos investigadores pueden implementar el modelo obteniendo resultados coherentes; si, a pesar de sus limitaciones, el modelo proporciona la mejor explicación disponible de un conjunto de datos– estas enfrentan varias limitaciones estructurales en el contexto psicológico. Uno de los principales retos es la falsabilidad: dado que muchos modelos tienen múltiples parámetros ajustables, pueden adaptarse fácilmente a distintos patrones de datos, lo que reduce su capacidad de ser refutados en sentido estricto. Este fenómeno se conoce como *overfitting*, y es especialmente problemático en modelos conexionistas o en redes neuronales profundas, donde la complejidad del sistema puede ocultar la falta de adecuación teórica. Otro desafío relevante es el de la validación empírica indirecta. Muchas simulaciones no se validan directamente con datos comportamentales, sino con otras simulaciones, lo que puede generar una forma de circularidad epistemológica.

Incluso en los casos donde una simulación reproduce fielmente los datos empíricos conocidos, persiste el problema de la ambigüedad interpretativa. Este problema ha sido ampliamente discutido en filosofía de la ciencia, especialmente en el marco del debate sobre la infradeterminación de las teorías (Quine, 1953; Stanford, 2009, 2023), la cual afirma que para cualquier conjunto de datos observacionales, siempre existirán múltiples teorías científicas lógicamente consistentes con esos datos. En el contexto de la simulación cognitiva, el problema se manifiesta en que múltiples modelos con supuestos teóricos diferentes pueden producir los mismos *outputs* observables. Este fenómeno es conocido como equifinalidad, que refiere a la capacidad de un sistema, sobre todo en el caso especial de los sistemas abiertos, en alcanzar el mismo estado final a partir de diferentes condiciones iniciales y diferentes procesos. Así, pone en cuestión la capacidad de las simulaciones para ofrecer explicaciones causales sólidas, pues, si varios modelos explican lo mismo con mecanismos distintos, ¿cuál representa mejor la realidad? La situación se agrava si consideramos que algunos modelos son ajustados específicamente para encajar con los datos disponibles, lo que puede llevar a una degeneración explicativa. En tales casos, el modelo pierde poder generalizador y se convierte en una herramienta meramente descriptiva. Además, (Romero, 2018) las teorías se prueban mediante la comparación de predicciones (enunciados) de modelos con datos. Un dato empírico es una proposición simple referida a un estado fáctico que se adquiere con la ayuda de operaciones empíricas (experimentos u observaciones). Luego un dato empírico no es un hecho, sino una proposición que informa sobre un hecho, de modo que siempre comparamos proposiciones con proposiciones. Y dado que las proposiciones son objetos conceptuales, están cargadas de teoría (y el hecho en sí mismo, por supuesto, es independiente de la teoría).

Así, toda simulación parte de supuestos teóricos previos, que condicionan tanto la estructura del modelo como su interpretación. Por ejemplo, una simulación basada en el supuesto de que la mente opera como un sistema bayesiano priorizará ciertas formas de razonamiento y descartará otras, no porque estas últimas sean empíricamente incorrectas, sino porque no encajan en el marco elegido. Por estas razones, algunos autores han cuestionado que las simulaciones puedan constituir explicaciones científicas en sentido fuerte. Winsberg (2010) ha señalado que muchas simulaciones actúan como "cajas negras epistémicas", cuya utilidad depende menos de su veracidad ontológica que de su eficacia práctica. Desde esta perspectiva, las simulaciones serían instrumentos de descubrimiento, pero no necesariamente representaciones fieles del sistema mental. Finalmente, hay que considerar que, en psicología, los datos empíricos no siempre son tan robustos o reproducibles como en otras ciencias. Esto plantea una dificultad adicional para validar simulaciones, ya que el criterio de correspondencia empírica es, a menudo, ambiguo o controvertido. Las limitaciones inherentes a la infradeterminación, la carga teórica y la ambigüedad interpretativa muestran que no basta con simular para explicar. Para que una simulación tenga validez explicativa experimental, debe estar cuidadosamente diseñada, contrastada con datos independientes y anclada en un marco teórico sólido que dé sentido a sus resultados. En ausencia de estas condiciones, el riesgo es confundir representación con apariencia, y explicación con imitación.

La frase de Von Neumann subraya el presupuesto epistemológico clave de la simulación, según el cual no es suficiente con describir un fenómeno; para realmente entenderlo, uno debe ser capaz de construir un modelo (que a menudo puede ser simulado) que replique su comportamiento. Si el modelo no produce los resultados esperados, revela una falta de comprensión. Que la simulación sea una suerte de prueba de comprensión, implica que la capacidad de simular un proceso es la prueba definitiva de que se ha comprendido dicho proceso. Según esta idea, si puedes programar las reglas y condiciones de un sistema y ves cómo evoluciona, has desentrañado sus mecanismos subyacentes. Von Neumann, además de trabajar en el contexto de problemas físicos y matemáticos complejos (como la bomba atómica), es uno de los padres de la computación moderna, y responsable de la arquitectura de los modernos ordenadores, y su célebre principio bien se puede extender a la cognición. Si no podemos simular cómo una mente (o un modelo de ella) percibe, aprende o decide, entonces nuestra comprensión de esos procesos cognitivos sigue siendo incompleta o superficial. Así, la simulación es una herramienta de comprensión científica que trascendería la mera observación y en la construcción y prueba de modelos explicativos. Como dice Daniel Dennet:

> La idea del programa de investigación de la IA [Inteligencia Artificial] es que podemos aprender sobre la mente construyendo y probando modelos que sean, en efecto, simulaciones del proceso de pensamiento, o simulaciones de la memoria, o simulaciones de la percepción, o lo que sea que estemos investigando. Pero, como ocurre con cualquier modelo, siempre es necesario recordar que es un modelo de la cosa, no la cosa misma. (Dennet, 1996, p. 51).

Si la simulación es capaz de realizar una misma tarea que nosotros hacemos, ¿es lícito apelar a nociones abstractas para explicar cómo el ser humano es capaz de realizar esas mismas tareas? Veamos el caso expuesto por Steven Pinker en *El lenguaje del pensamiento* en 1985. En el capítulo dedicado al mentalés este expone cómo un ordenador computacionalmente podría reconstruir un silogismo aristotélico, describiéndolo en términos meramente lógicos y computacionales. ¿Qué disponíamos

antes que explicase cómo somos capaces de llevar a cabo razonamientos deductivos como los silogismos? Ciertamente, nada concreto. En cambio, si tenemos una forma computacional de dar cuenta de dicho proceso ¿no es acaso razonable que sea así cómo lo llevamos a cabo también los seres humanos? O, si se quiere, ¿al menos de forma análoga a la forma computacional? ¿Cuál es la razón de oponerse ante ese tipo de explicaciones cuando la alternativa es tan poco concreta y quimérica como la explicación antaño en física de ciertos fenómenos por la hipótesis de una supuesta "acción a distancia"? Puede considerarse que los modelos cognitivos basados en simulaciones se sitúan generalmente en el ámbito de la explicación funcional. Muchos de ellos, como los basados en arquitecturas cognitivas (p. ej., ACT-R; Anderson et al., 2004), redes neuronales artificiales o sistemas multi-agente, están diseñados para replicar conductas cognitivas observadas, esto es: tiempos de reacción, tasas de error, decisiones probabilísticas, etc. En este sentido, cumplen razonablemente con los criterios de generalidad (pueden aplicarse a múltiples tareas); consistencia interna (siguen reglas formales precisas); y, en ciertos casos, capacidad predictiva. Además, las simulaciones tienen la ventaja de ofrecer una formalización computacional clara y ejecutable de los supuestos funcionales. Sin embargo, es más discutible el valor explicativo de las simulaciones desde el punto de vista de la explicación mecanicista. En muchos casos, los modelos basados en simulación no están diseñados para reflejar los mecanismos neuronales reales ni las propiedades físicas del cerebro. Incluso en simulaciones conexionistas, las unidades y pesos no tienen una correspondencia directa con estructuras biológicas identificables. Por ello, aunque estos modelos puedan replicar la conducta, no permiten inferencias realmente sólidas sobre los mecanismos causales implicados. Como han argumentado Bechtel et al. (2009), una simulación puede reproducir el comportamiento de un sistema sin revelar por qué o cómo ese comportamiento ocurre; la simulación en ese caso tomaría la forma de una *ilusión explicativa*, pero, por el contrario, nuestro empleo de dichos modelos simulativos no tiene como objetivo *parecer* que explican, sino explicar realmente.

4.2. Emergentismo, brecha explicativa y pluralismo científico

Desde, sobre todo, Descartes se hace una división entre la mente y el cuerpo, pues de que se tiene mente no se puede dudar —pienso, luego existo—, en cambio del cuerpo sí; no se trata una verdad apodíctica, a partir de lo cual pueda construirse el resto de conocimiento, que lo será en función de que remita en último término a ese ser pensante. El pensamiento dominante negaba cualquier tipo de ciencia de la mente, pues esta no era una máquina, a diferencia del cuerpo, del resto de animales, etc. El "dilema del teórico" al que alude el conductismo plantea una disyuntiva: si la psicología introduce estados internos (creencias, representaciones, procesos mentales) para explicar la conducta, abandona el ideal de objetividad empírica; pero si se limita a registrar correlaciones entre estímulos y respuestas observables, renuncia a ofrecer explicaciones genuinas y se queda en una mera *lista de lavandería* de regularidades. Pero, de igual modo, al introducir procesos mentales hipotéticos para explicar la conducta, arrastraría —según sus proponentes— a la conjetura metafísica. El conductismo psicológico sostiene que, para preservar la objetividad, la psicología debe limitarse a vincular estímulos y respuestas observables. Jonathan Laid acierta en mostrar que esta disyuntiva se erige sobre un supuesto falso: creer que el único cometido de la ciencia es formular leyes descriptivas (y, a lo sumo, usarlas para predecir y controlar el comportamiento). En realidad, la labor científica no solo quiere describir los fenómenos mediante leyes, sino también explicarlos.

La naturalización de la mente es el proyecto filosófico y científico de explicar los fenómenos mentales en términos de procesos naturales, físicos y biológicos. Las simulaciones, como herramientas computacionales, se sitúan dentro de este proyecto: buscan mostrar cómo fenómenos complejos —como la percepción o la memoria— pueden surgir de sistemas artificiales implementados en máquinas o programas. Sin embargo, este intento tropieza con el problema de la emergencia. Algunos autores han señalado que ciertos aspectos de la mente, como la consciencia, las emociones o la comprensión simbólica, son propiedades emergentes que no pueden reducirse a sus componentes más simples. En otras palabras, no hay una correspondencia directa entre la estructura del modelo y la experiencia mental resultante. Esta dificultad se refleja en lo que Joseph Levine (1983) denominó la brecha explicativa: el salto que existe entre los datos *físicos* o computacionales y la vivencia subjetiva. Según Levine, incluso si pudiéramos especificar todas las correlaciones neurofisiológicas de un estado consciente, seguiría siendo un misterio por qué ese estado *se siente* de una determinada manera. Esta brecha también afecta a las simulaciones. Aunque un modelo pueda replicar el comportamiento de un agente que reconoce emociones, no implica que ese sistema *experimente* emociones. Simular la conducta asociada a la ira, por ejemplo, no es lo mismo que *sentir* ira. Diversos autores han propuesto marcos filosóficos que reconocen estas limitaciones sin abandonar por completo el ideal de naturalización. Uno de ellos es Donald Davidson, con su tesis del monismo anómalo (1970). Según Davidson, los eventos mentales son idénticos a eventos físicos, pero no existen leyes psicofísicas estrictas que permitan reducir lo mental a lo físico de manera sistemática. Se da una irreductibilidad de los conceptos mentales, de modo que realizar la traducción a conceptos físicos, es una tarea imposible e, incluso, absurda, pues los conceptos de las teorías físicas tienen sentido dentro, precisamente, de dichas teorías, no fuera, las cuales además tienen por campo de estudio un dominio distinto al de la psicología. Así, en este sentido, la mente es irreductiblemente anómala al no dejarse atrapar en regularidades estrictas, como las que exigen los modelos simulativos.

Al presupuesto epistemológico que parece subyacer a los modelos cognitivos basados en simulaciones se contraponen críticas como que la reducción de la cognición a procesos simbólicos o algorítmicos omite aspectos fundamentales de la experiencia humana, como la conciencia fenoménica, la subjetividad, la intencionalidad, las emociones, etc. Estas argumentarían que la mente no es simplemente un ordenador que procesa información de forma desencarnada —*disembodied*—. Desde las perspectivas de la cognición encarnada (*embodied cognition*) y el enactivismo, la mente no es una entidad separada que procesa información del mundo, sino que emerge de la interacción dinámica y recíproca entre un cuerpo situado, un cerebro y su entorno. Los modelos de simulación, al aislar a menudo el proceso cognitivo de su contexto físico y social, pierden esta dimensión fundamental. En *The Embodied Mind* (1992), Francisco Varela, Evan Thompson y Eleanor Rosch argumentan que la cognición no es solo una representación interna del mundo, sino una enacción o construcción activa del sentido a través de la interacción sensoriomotora. Desde esta perspectiva, simular procesos cognitivos sin tener en cuenta el cuerpo y la experiencia directa es una caracterización fundamentalmente incompleta y no puede ser una explicación psicológica adecuada. Ese carácter de lo mental que emergería de la interacción con el entorno nos pone sobre la pista del problema ontológico de lo nuevo, que es uno de los problemas metafísicos fundamentales clásicos. La emergencia implica una novedad ontológica de carácter cualitativa; en oposición al reduccionismo descendente presente en neurociencia, etc. En el surgimiento de toda propiedad emergente, por supuestamente humilde que sea, se desmiente el reduccionismo descendente. La reducción

descendente solo es legítima cuando no se da una novedad cualitativa real. ¿Cómo explicar la aparición de novedades cualitativas irreductibles? ¿Se puede negar su existencia, de modo que pueda postularse un reduccionismo descendente?

> En Bunge —como en muchos otros autores emergentistas— la novedad cualitativa se concibe como resultado de las interacciones entre la composición, la estructura, los mecanismos y el entorno de un sistema. Es cierto que no se conoce ningún mecanismo universal de emergencia. Hay novedades cualitativas de las que se entiende su mecanismo de emergencia y otras de las que no. Pero sabemos que cambiando la composición, estructura, mecanismo o entorno de un sistema sus propiedades emergentes se alteran o desvanecen. (Pérez-Jara y Camprubí, 2025, p. 114).

Parecen existir varios ejemplos de novedades cualitativas: una sociedad tiene propiedades de las que sus individuos carecen; un individuo tiene propiedades de las que sus células carecen; una célula tiene propiedades que no se encuentran en sus componentes moleculares; una estrella tiene luminosidad, pero no la tienen sus componentes atómicos, etc. También hay muchos ejemplos de cómo, al aparecer novedades cualitativas, otras propiedades desaparecen o se desvanecen. Un electrón tiene espín y número leptónico, pero una célula no; un individuo está vivo, pero no una familia, etc. A pesar de ello, hay quien pudiera decir que, del mismo modo que el estado agregado de la materia se puede explicar por las moléculas, en un futuro podemos hallar un modo de explicar mediante lo más simple lo más complejo sin aludir al concepto, para muchos intuitivo, pero ciertamente especulativo, de emergencia. Mario Bunge (1997; 2003) comprende que, en los procesos emergentes, unas propiedades se ganan y otras se pierden sin que se conozca ningún tipo de mecanismo universal que explique esta dialéctica de surgimiento de novedades cualitativas y desvanecimiento de los mismos.

Un ejemplo paradigmático de los límites de las explicaciones naturalistas y simulativas es el de la sociobiología, que ha intentado explicar comportamientos complejos —como el altruismo, la agresividad, el amor, o la moralidad— en términos de estrategias de supervivencia y reproducción genética. Aunque estas explicaciones pueden tener valor evolutivo o adaptativo, reducen los fenómenos humanos a meros epifenómenos biológicos, perdiendo de vista su dimensión significativa, cultural y experiencial. En psicología se tiende a producir un riesgo es similar: una simulación que modela la toma de decisiones como maximización de utilidad puede describir ciertas conductas, pero no explica el sentido que el agente atribuye a sus decisiones, ni las razones normativas que lo guían. Esto es debido, no solo a la brecha explicativa y la emergencia de sistemas complejos que se comportan de modo distinto y, por ende, requieren un tipo de explicación diferente, sino también por el pluralismo científico que algunos filósofos, como Suppes, Nancy Cartwright o Helen Longino, por mencionar a unos pocos, defienden —y que se aboga por él en el presente trabajo. De forma clásica fue propuesto del siguiente modo:

> La idea general es que algunos fenómenos naturales no pueden explicarse completamente con una sola teoría ni investigarse a fondo con un único enfoque. En consecuencia, se requieren múltiples enfoques para la explicación e investigación de dichos fenómenos. En algunos casos, el interés en el pluralismo está motivado por el análisis de cuestiones específicas dentro de una ciencia, y en otros, por el análisis de cuestiones filosóficas y metodológicas generales. La forma

> en que se entiende el pluralismo —ya sea, por ejemplo, afirmando una heterogeneidad ontológica o epistemológica radical o simplemente la diversidad de mecanismos en la naturaleza— varía según el pensador y el tema. (Kellert, S. H. et al., 2006/1956. *vii*).

Más actualmente y en el panorama de la filosofía española, incluso en entornos no analíticos, parece tenerse la misma intuición:

> Nosotros no podemos dictar la estructura de la realidad, tan solo operamos, vamos transformando cosas y en el propio proceso de las transformaciones se van organizando esos cúmulos operatorios, que interpretamos como categorías ontológicas porque nos proporcionan las junturas naturales por las que se divide la realidad cuando se transforma. Las ciencias cierran de acuerdo con la estructura operatoria y con la estructura de los resultados de lo que se está operando: por eso el cierre operatorio puede interpretarse como un cierre categorial, es decir, el cierre de las ciencias nos da la pista de cómo está estructurado el mundo en categorías. Las categorías son algo así como el «mapa» de la estructura del mundo: las relaciones entre las ciencias, entre sus fronteras y entre sus cierres operatorios son las que nos informan de que la legalidad biológica es distinta de la legalidad física, de la química, de la histórica, de la psicológica, de modo que esas categorías no se reducen unas a otras. Por mucho que los físicos pretendan hacer teorías del todo y reducirlo todo a física, la terca realidad es que el mundo no tiene esa estructura unificada. Es necesario reconocer un pluralismo gnoseológico y ontológico, lo que significa reconocer que unas áreas de la realidad son irreductibles a otras. (Alvargonzález, 2024, p. 206).

En esta misma línea, la pregunta que hiciese Jackson respecto de si Mary la neurocientífica que sabe todo de las características físicas del color, sabe algo más del color rojo cuando lo experimenta fenoménicamente por primera vez, puede responderse haciéndose una defensa materialista, sin incurrir en un reduccionismo científico, aludiendo a que el acceso que tiene la propia materia —que constituye a Mary— respecto del proceso que tiene lugar cuando la luz incide en la retina no tiene que ser el mismo que el acceso que se tiene desde un modelo matematizado de aquella parcela o parcelas de la realidad (no de toda ella, sino de aquello que es susceptible de ser parametrizado de ella). Desde las coordenadas, no de un monismo fisicalista, sino de un pluralismo, se puede comprender que existen diversos accesos a un objeto, sin que exista propiamente un acceso privilegiado; pretendiéndose que dicho acceso privilegiado no sea un modo de acceso fenoménico, sino nuoménico (utilizando terminología kantiana). Dicha brecha es solo epistémica, pues si se considera también ontológica se estaría añadiendo una propiedad no reductible a la materia, sin tampoco saber dar cuenta de cómo esta puede existir o de qué modo no se puede, en último término, colapsar en una física.

Asimismo, cabe argumentar que si defendemos un pluralismo científico, así como una postura cientificista débil, de modo que aunque se considere que el mejor modo de conocer que tenemos es el científico no se considere este el único existente, cabe también tener una visión plural de la explicación, pues los diferentes modos de explicación que se veían en el apartado anterior están más o menos presentes o resultan más o menos pertinentes según sea la naturaleza del fenómeno que se pretende comprender. Diversas metodologías de investigación (cuantitativas, cualitativas, simulaciones, experimentos, observaciones, etc.) son apropiadas y necesarias para investigar diferentes aspectos de

la realidad. Y, asimismo, la realidad misma está estratificada, esto es: existen diferentes "niveles" de realidad, cada uno con sus propias propiedades y leyes emergentes, que requieren explicaciones específicas. Por ejemplo, el nivel microfísico, el nivel químico, el nivel biológico, el nivel social, etc. Los sistemas complejos exhiben propiedades y comportamientos a nivel macro que no pueden ser simplemente reducidos o explicados exhaustivamente por las propiedades de sus componentes individuales a nivel micro. Rolando García, uno de los mayores colaboradores con el psicólogo, Piaget, en su libro sobre los *Sistemas complejos* (2006), expresa esta idea del modo que sigue:

> En nuestra concepción de los sistemas complejos, lo que está en juego es la relación entre el objeto de estudio y las disciplinas a partir de las cuales realizamos el estudio. En dicha relación, la complejidad está asociada con la imposibilidad de considerar aspectos particulares de un fenómeno, proceso o situación a partir de una disciplina específica (García, R. p. 21).

Extendiendo lo dicho por R. García, no solo a partir de una disciplina específica, pues, a diferencia de lo que podría sugerir la etiqueta unificadora de la psicología, en esta disciplina existen fenómenos que se explican mejor funcional, causal, pragmática o normativamente, como se defenderá en el siguiente apartado. La existencia de diferentes niveles de explicación se debe a que abordamos aspectos de objetos (y sujetos) desde diversas ciencias y disciplinas (lo que nos es accesible por una, no lo es por otra, y al revés). Si la psicología misma no tiene un único objeto de estudio —no se limita solo a la conducta públicamente observable, como sostiene el conductismo lógico—, tampoco puede pretender que la explicación psicológica sea única; esta también debe ser plural. Y, para el caso de una explicación psicológica, como se defenderá en el apartado siguiente, es necesario, aunque no suficiente, que contenga una explicación normativa. Pues, como se veía, una explicación enteramente funcional puede llevarnos a creer que alguien posee una habilidad de la que en verdad carece. Lo que también mostraba la insuficiencia de aquellas teorías (TS, TT, etc.) que basan el conocimiento de otras mentes a la mera inducción a partir de su conducta visible. A este respecto se pueden poner ejemplos más o menos sencillos de la psicología, desde la inferencia del significado de un gesto que no conoces su significado por su uso en un contexto, a la naturaleza de una conducta, como por ejemplo, levantar el dedo índice de la mano. ¿Se trata de la misma acción levantar un dedo "voluntariamente" a hacerlo por estimulación eléctrica por medio de un EEG? Cualquiera que haya realizado ambas cosas, sabrá que no *aparecen* del mismo modo al sujeto, en este segundo caso se siente más como un espasmo nervioso, pero a pesar de cómo se vive en primera persona, la conducta observable es la misma: levantar el dedo índice de la mano. ¿Diríamos entonces en este caso que es la misma acción? En este sentido podemos encontrar, no solo iguales conductas observables con distintas causas, sino también con distinto significado social o normativo. A pesar de ello, algunos autores como David Davidson en su pequeño artículo *La mente material* han criticado la idea de que haya "muchas acciones para una misma conducta" o "diferentes conductas para una misma acción", alegando que el escenario donde levantar el puño es mal visto y el escenario en donde es bien visto son, de hecho, diferentes en sus componentes físicos. De modo que siempre se van a poder diferenciar físicamente una conducta que representa una acción, de otra que no lo hace; aunque, por descontado, será mediante un análisis de los procesos y mecanismos mediante experimentación, no por la mera conducta observable en el sentido más ramplón del término, a saber, aquella con la cual podemos acceder mediante los sentidos humanos. De todas formas, existen ejemplos dentro de la psicología, como el caso de los delirios, en donde estos pueden ser causados por diferentes factores, aunque el contenido de los delirios, así como la conducta

asociada de las personas delirantes, sea la misma. En algunos casos pueden ser explicadas funcionalmente, como suele hacerse en el análisis funcional de la conducta, dando cuenta de que surgieron en un momento de gran estrés ambiental de la vida del sujeto y cumplían los delirios una función de autorregulado. Pero asimismo existen delirios claramente causados por enfermedades, tanto físicas como psiquiátricas, y que estas, por descontado, no pueden ser explicadas funcionalmente.

4.3. Explicación normativa de la mente

La postura que se pretende defender es una explicación psicológica debe dar cuenta de la normatividad de lo mental —pensamiento, lenguaje, creencias, intenciones, etc.—. Sellars, en su libro *Empiricism and the Philosophy of Mind* (1997), critica la idea empirista de que el conocimiento comienza con datos sensoriales brutos y no conceptuales ("lo dado"). Argumenta que toda experiencia significativa y todo conocimiento están ya imbuidos de conceptos, y los conceptos son inherentemente normativos. La capacidad de pensar, creer y justificar no sería un proceso meramente causal (como el "reino de la ley" de la ciencia natural), sino que opera en un "espacio de las razones" donde las proposiciones se justifican unas a otras inferencialmente. Este espacio es constitutivamente normativo: las inferencias son correctas o incorrectas, las creencias están justificadas o no. La mente, en tanto participa en el razonamiento, se encuentra en este espacio. Nuestros conceptos de estados mentales internos (pensamientos, sensaciones, etc.) son, en parte, postulados teóricos que nos permiten explicar y predecir el comportamiento de los demás y el nuestro propio, pero estos postulados se entienden dentro del marco de nuestro lenguaje público y sus normas. En esta misma línea, McDowell (1994) desarrolla la idea de Sellars del "espacio de las razones" para argumentar que la mente humana está inherentemente constituida por la racionalidad y la normatividad. La mente no es una esfera separada del mundo, sino que su contenido es inherentemente objetivo y está sujeto a los mismos estándares de corrección que las cosas en el mundo. No hay una "mente sin mundo" o un "mundo sin mente" cuando se trata de la experiencia significativa humana. Del mismo modo, Brandom (1994) sostiene que el contenido conceptual de las creencias y los pensamientos se explica por su rol en las inferencias. Lo que hace que una creencia sea la creencia que es, no es una representación interna, sino las relaciones inferenciales normativas en las que se involucra (lo que implica, lo que es implicado por ella, lo que la justifica, lo que la excluye). El significado —y por extensión, el contenido mental— no es primariamente semántico (representacional), sino pragmático y normativo. El significado surge del "juego de dar y pedir razones" en una comunidad lingüística.

La breve exposición del pensamiento de los llamados filósofos de Pittsburgh muestran el reto que es la explicación normativa de lo mental. El enactivismo, una rama de la ciencia cognitiva sostiene que la cognición no es simplemente una representación interna del mundo, sino una acción dinámica y activa de "dar sentido" al mundo a través de la interacción de un organismo encarnado con su entorno. La mente y el mundo se configuran mutuamente., a través asimismo del lenguaje, el cual está intrínsecamente ligado a una forma de vida, y las reglas que lo rigen son internas a esa práctica, no algo externo que luego se aplica a ella. La explicación psicológica debe capturar cómo los fenómenos mentales no solo se procesan, sino que se generan activamente y se construyen a través de la relación entre el agente y su entorno. No es un modelo pasivo de entrada-salida, sino un proceso de creación de significado mediante la acción. Si la explicación psicológica se centra en la enacción dinámica y

generativa de los fenómenos por un organismo encarnado en su entorno, entonces la cuestión de la normatividad (es decir, lo que debería ser, las reglas que gobiernan el "juego" del lenguaje y el pensamiento, y el seguimiento de esas reglas) emerge como un elemento crucial que, a menudo, parece ausente en los modelos puramente computacionales de IA. Los fenómenos psicológicos son generados activamente por un agente encarnado en su interacción con el entorno. Pero esta interacción y generación no ocurren en un vacío, sino que están intrínsecamente ligadas a prácticas normativas y a la comprensión de "cómo se hacen las cosas correctamente".

Con respecto a la definición, no de la mente, sino del arte, han existido históricamente diferentes aproximaciones —trascendental, intencional, funcionalista, institucional, histórica, simbólica, etc.—. Resulta pertinente para la caracterización de la explicación normativa de lo mental la aproximación institucional, en concreto la teoría institucional de arte de George Dickie. En ambas, lo que se considera, ora obra de arte, ora "lo mental", difiere según tiempo y lugar. Pero, sendas filosofías, del arte y de la mente, intentan hacer de la obra de arte y de la mente la base de su disciplina, por lo que la definición de cada una de ellas cobra especial relevancia. Para George Dickie su teoría institucional del arte no es un instrumento de identificación de obras de arte; más bien comprende el autor que las teorías del arte son una explicación de por qué una obra de arte *es* arte. Explicar por qué alguien hizo X (un acto psicológico como "pedir disculpas") no es solo trazar la actividad neural o los pasos algorítmicos. Es entender el juego lingüístico de las disculpas, las normas que rigen cuándo y cómo se ofrecen, y el rol que juegan dentro de una forma de vida. La "enacción" de la disculpa está, por ejemplo, impregnada de esa normatividad. Una explicación institucional considera que las obras de arte son tales por la posición que ocupan en un contexto institucional, en concreto, esto es: "el mundo del arte", en tanto que institución social. La comprensión de un concepto como "justicia" o "belleza" no se reduce a cómo el cerebro procesa estímulos; se explica por cómo participamos en los juegos lingüísticos y las prácticas sociales donde esos conceptos adquieren su sentido normativo.

En *Investigaciones Filosóficas*, Wittgenstein argumentó que el significado y el pensamiento no residen en una entidad mental interna o en una representación fija, sino en el uso del lenguaje en el contexto de "juegos lingüísticos". Estos juegos lingüísticos tienen lugar en formas de vida, y son prácticas sociales que están gobernadas por reglas. Para Wittgenstein, seguir una regla no es un proceso psicológico misterioso; es una práctica social. La validez de una regla no reside en un imperativo lógico abstracto, sino en que es aceptada y seguida por una comunidad de hablantes. Si alguien sigue una regla, no hay una instancia privada que determine si la está siguiendo correctamente; es la comunidad la que lo juzga. Así, entender una palabra o un concepto es saber cómo usarla correctamente en diferentes contextos. Pensar y hablar no son solo procesos computacionales; son actividades que implican juicios de corrección e incorrección, de lo que "debería" o "no debería" hacerse. Estos juicios son inherentemente normativos. Los fenómenos psicológicos son intrínsecamente normativos, por lo que la explicación de un fenómeno psicológico no puede limitarse a describir los mecanismos o procesos que lo generaron. También debe dar cuenta de su posición dentro de un espacio de razones y reglas que le otorgan su significado y su carácter de "correcto" o "incorrecto" en un determinado contexto social y lingüístico. La "enacción" de un pensamiento o de una acción (como dar un argumento, hacer una promesa, o sentir vergüenza) no es meramente un proceso físico o computacional. Implica la participación en un juego con reglas implícitas o explícitas. Un agente no solo "hace", sino que "hace correctamente" (o incorrectamente) según las normas de su comunidad o

el contexto de la actividad. Consiguientemente, la explicación de un fenómeno psicológico no puede limitarse a describir los mecanismos o procesos que lo generaron. También debe dar cuenta de su posición dentro de un espacio de razones y reglas que le otorgan su significado y su carácter de "correcto" o "incorrecto" en un determinado contexto social y lingüístico. Por ejemplo, para entender el acto de habla de pedir disculpas no solo hace falta trazar la actividad neuronal que tiene lugar cuando se pide disculpas o el número de pasos algoritmos necesarios para llevarlo a cabo, sino que hay que conocer las normas que rigen cuándo y cómo se ofrecen, y el rol que juegan dentro de una forma de vida.

> Y por eso "seguir la regla" es una práctica. Y creer que se sigue la regla no es seguir la regla. Y por eso no se puede seguir una regla 'privadamente', pues de otro modo creer seguir la regla sería lo mismo que seguir la regla. (Wittgenstein, *Investigaciones Filosóficas*, §202).

Una regla no es una fórmula que se pueda aplicar mecánicamente ni una interpretación mental que se impone a los casos particulares. Esto se ve claramente en su crítica a la idea de que toda aplicación de una regla es en realidad una interpretación (IF §198). Si eso fuera así, no habría nunca una base firme para distinguir entre seguir correctamente una regla y simplemente creer que se la está siguiendo. Esto abriría la puerta al escepticismo: cualquier uso podría considerarse conforme a alguna regla bajo una interpretación adecuada, y por tanto, ninguna regla nos orientaría realmente. Para evitar esta paradoja, Wittgenstein subraya que seguir una regla es una práctica inserta en un marco intersubjetivo: se aprende en comunidad, se valida en comunidad, y su corrección se mide en relación con una forma de vida compartida. "'Seguir la regla' es una práctica" significa que el acto de seguir una regla no es un estado mental interno o un mecanismo neurofisiológico aislado, sino una actividad encarnada en un contexto social. No es una representación mental de la regla lo que la valida, sino el acto recurrente y reconocible de hacer algo que es juzgado como "seguir la regla" por otros en una comunidad. El núcleo del argumento estriba en la idea de que creer que se sigue la regla no es seguir la regla. Si el pensamiento implica la aplicación de conceptos (y los conceptos son reglas de uso), y si las reglas no pueden seguirse privadamente, entonces el pensamiento mismo está intrínsecamente ligado a la normatividad pública. "El carácter distintivo de la mente humana no es que sea una máquina de calcular. Es que es una máquina de dar sentido, de poner en el espacio de las razones" (McDowell, 1994, p. 119). Sin normas públicas que rijan el uso de conceptos y el seguimiento de reglas, no hay significado. Si lo mental fuera solo un proceso causal, no habría lugar para el error genuino, solo para un resultado diferente. El pensamiento y el lenguaje no son fenómenos privados, sino actividades arraigadas en nuestras formas de vida, y es la normatividad lo que permite la comunicación, la comprensión mutua y la constitución de un mundo compartido.

Esta defensa de una caracterización normativa —aunque no solo normativa— de lo mental es que esta no implica necesariamente abandonar los intentos por naturalizar la mente. Por ejemplo, Dretske en su libro *Explaining behavior* (1988) busca naturalizar la explicación por razones, pero su distinción entre tipos de causas y el rol del contenido mental inevitablemente introduce una forma de normatividad. Si no hay normatividad (es decir, condiciones de corrección para el contenido representacional), la distinción que Dretske hace entre, por un lado, un mero movimiento corporal (que puede ser explicado puramente por causas físicas) y, por otro, un comportamiento intencional (que es explicado por razones) se desmorona. Sin la idea de que una creencia representa correctamente (o

incorrectamente) algo, o que un deseo apunta a un estado de cosas deseado, los estados mentales perderían su carácter distintivo como causas estructurantes. Simplemente serían más "causas disparadoras" a nivel neurofisiológico. Para Dretske, el contenido de una creencia o deseo es lo que realmente explica el comportamiento. Y el contenido es inherentemente normativo en el sentido de que tiene condiciones de verdad o satisfacción. Si ignoramos la normatividad, reducimos el contenido a una mera propiedad física del cerebro, perdiendo su poder explicativo sobre el tipo de comportamiento que produce. No podríamos explicar por qué la creencia de que "el agua está en el vaso" causó que el agente bebiera, a menos que el contenido "agua en el vaso" sea relevante para la acción de beber, y esa relevancia implique condiciones de corrección (que la creencia sea verdadera, que el deseo sea satisfecho). ¿Por qué considera Dretske que las razones pueden ser causales? Sus "razones" son creencias y deseos, y la eficacia causal de estas reside en su contenido. Toda explicación psicológica debe de poder ser normativa, pues si eliminamos la normatividad, las "razones" se convierten en meros eventos neuronales, y la explicación de la acción por razones —la "psicología" en contraste con la "biología" de los movimientos— perdería su fundamento. Las razones no son meras ilusiones o narrativas, pudiéndose explicar el comportamiento mediante causas que podemos establecer desde la física, sino que, como afirma Dretske, el propio contenido de dichas razones tienen un poder causal que solo puede ser explicado desde el propio nivel normativo. Estas "condiciones" a las que el contenido de una creencia o deseo se refiere son, en esencia, sus condiciones de corrección o satisfacción. Si creo que "hay agua en el vaso", mi creencia es correcta si el agua está allí. Es este aspecto normativo del contenido (su función de representar el mundo, y la posibilidad de que lo haga bien o mal) lo que le permite ser una causa estructurante que explica por qué ocurre ese comportamiento y no otro. Si los estados mentales no tuvieran este tipo de contenido (con sus condiciones de corrección), serían indistinguibles de cualquier otro evento físico que simplemente desencadena una respuesta, perdiendo así su especificidad como estados mentales con poder explicativo sobre la acción intencional. Así, la explicación normativa propuesta propone que la explicación de un fenómeno observable (F) se logra al demostrar cómo F se constituye y se hace inteligible dentro de un sistema de prácticas normativas donde su ocurrencia y su forma son evaluadas y adquieren significado a través de la adhesión a un conjunto de principios constitutivos y regulativos.

¿Qué consecuencias tiene esto para la pregunta que guía este trabajo respecto de si los modelos cognitivos basados en simulaciones pueden constituir una explicación psicológica? Considérese, por ejemplo, un programa de IA. Uno puede preguntarse si este, además de poder seguir reglas, las entiende en el sentido normativo de Wittgenstein. ¿No sabe por qué esas reglas son correctas o qué significa equivocarse? Un algoritmo puede generar una respuesta lingüísticamente "correcta", pero parece que no la ha generado porque entienda la regla de corrección, sino porque ha sido programado para ello o ha aprendido patrones estadísticos. Pero ¿las tiene que *entender* —entendiéndose por "entender" de la misma forma que nosotros— en el sentido de la equivalencia fuerte que defendía Fodor? Si consideramos que seguir una regla no es lo mismo que creer que se sigue una regla, parece que la creencia sobre seguir una regla o no carece de relevancia para determinar si se sigue una regla. Luego no parece teóricamente que haya un obstáculo insoslayable, según esta perspectiva, en que un modelo basado en simulaciones pueda capturar dichas condiciones normativas de la mente y reproducirlas.

Cabe pensar, dentro de este marco, qué puede decirse respecto a los modelos de lenguaje. La creatividad que puede emerger de LLMs –Large Language Models– como ChatGPT proviene,

precisamente, de la capacidad que tienen para errar. Explicaré con un poco más de detalle esta idea. La creatividad de modelos como ChatGPT no surge del cumplimiento inflexible de reglas gramaticales o de una comprensión conceptual profunda, sino precisamente de su capacidad para generar secuencias que, aunque muchas veces se ajustan a patrones de uso conocidos, en otras ocasiones se desvían, combinan inesperadamente, reorganizan fragmentos en formas nuevas. Lo que en un sentido clásico se podría describir como "error" —una formulación inesperada, un uso inusual de una expresión, una inferencia no canónica— puede, en otro contexto, abrir una vía de sentido inédita. En este sentido, los LLMs participan de una forma de comportamiento que recuerda al modo en que los humanos seguimos reglas: no por ejecución mecánica, sino por incorporación estadística a un espacio de usos probables, modelado por contextos pasados pero abierto a combinaciones futuras. Y si bien los LLMs no poseen intenciones ni comprenden las reglas en el sentido humano —porque no habitan una forma de vida ni forman parte de una comunidad de hablantes en sentido estricto—, su producción discursiva está sujeta a un campo normativo construido a partir del lenguaje humano, cuyas tensiones y flexibilidades se filtran en sus salidas. Es allí donde puede emerger lo "creativo": no de un cálculo cerrado, sino de una red de regularidades que admite su propia transgresión. Del mismo modo, resulta análogo lo que Kahneman menciona en su último libro Ruido (2023). Uno de los problemas respecto de la objetividad de los juicios es que muchos de estos no son comprobables. Donde hay cálculo, no hay propiamente ruido, solo cuando entra en juego el juicio humano. Kahneman lo expresa del modo siguiente:

> Las normas compartidas dan a los profesionales una idea de los elementos que deben tener en cuenta y de cómo emitir y justificar sus juicios finales. En la compañía de seguros, por ejemplo, los tasadores de reclamaciones no tenían ninguna dificultad para ponerse de acuerdo y describir las consideraciones pertinentes que debían incluirse en la lista de control para evaluar una reclamación. Por supuesto, este acuerdo no impidió que los tasadores de reclamaciones variasen ampliamente en sus evaluaciones, porque la doctrina no especifica con claridad cómo proceder. No hay una receta que pueda aplicarse mecánicamente. En cambio, la doctrina deja espacio a la interpretación. Los expertos siguen haciendo juicios, no cálculos. De ahí que el ruido sea inevitable. Incluso profesionales con idéntica formación que están de acuerdo con la doctrina que aplican se distanciarán unos de otros en su aplicación. (Kanheman, p. 253).

Sea como fuere, lo que sí está claro es que si un modelo cognitivo basado en simulación aspira a ser una explicación psicológica completa no solo debe simular la dinámica interna de un proceso –pudiendo ofrecer, así, una explicación causal-mecánica o funcional de lo mental–, también debe simular la interacción del agente con un entorno normativo; es decir, el entorno en el que el agente opera debe tener reglas implícitas o explícitas que gobiernen el "juego" en el que participa la cognición. La simulación debería poder demostrar cómo el agente aprende, internaliza o se ajusta a las expectativas de corrección. Así como abordar la cuestión de la comprensión de la regla como una práctica social, no solo como una computación interna. Desde luego, este es el mayor desafío, ya que requeriría que la simulación refleje cómo la participación en una comunidad de "juegos lingüísticos" dota de significado normativo a la conducta. Pero, aunque sea un reto, no significa que sea teóricamente imposible, e, igualmente, que los modelos basados en simulaciones puedan verse más enriquecidos de incorporarse esta normatividad en su modelaje.

Con todo, para que estos modelos constituyan una explicación psicológica completa, debe de considerarse también el tipo de explicación normativa, la cual nace del supuesto de que explicar un fenómeno no es solo mostrar sus causas o su generación, ni atener meramente a cuestiones "internas", sino situarlo dentro de un marco de inteligibilidad definido por reglas y principios normativos que dictan su "corrección" o su "validez" dentro de un dominio específico. Si este trabajo sirve para señalar la pertinencia teórica del uso de modelos basados en simulaciones, en concreto para modelar cuestiones normativas —intrínsecas— de lo mental, su pretensión de aplicación práctica deberá de ser puesta en marcha en un futuro próximo y evaluar, así, su viabilidad real.

5. CONCLUSIONES

A lo largo de este trabajo se ha intentado abordar, desde una perspectiva filosófica y epistemológica, el valor explicativo de los modelos de la cognición humana basados en simulaciones. Frente a la creciente presencia de sistemas simulativos en el discurso científico contemporáneo —y especialmente en el campo de la psicología cognitiva y la inteligencia artificial—, el objetivo ha sido someter a examen su capacidad real para representar y explicar los procesos mentales humanos. El punto de partida ha sido una constatación: las simulaciones, en sus distintas formas, constituyen una herramienta cada vez más utilizada en la investigación psicológica, no solo para visualizar datos o explorar hipótesis, sino también, y cada vez más, para formular explicaciones del comportamiento cognitivo. Sin embargo, esta expansión técnica y discursiva no ha ido siempre acompañada de una clarificación teórica sobre qué tipo de explicación ofrecen tales modelos, ni sobre qué condiciones deben cumplirse para que puedan ser considerados explicaciones en sentido pleno.

El recorrido desarrollado ha mostrado, en primer lugar, que el concepto de modelo científico es profundamente plural y depende del contexto disciplinar y del marco teórico en que se inscribe. Mientras en algunas ciencias los modelos tienden a ser interpretados como representaciones estructurales del mundo, en otras se los considera instrumentos funcionales o dispositivos heurísticos. En el caso específico de la psicología, esta ambivalencia se hace aún más aguda debido a la complejidad de su objeto: la mente humana no se deja reducir fácilmente a mecanismos observables, y su carácter intencional, subjetivo y experiencial plantea desafíos específicos para la modelización. Por ello, se ha sostenido que cualquier evaluación del valor explicativo de una simulación debe tener en cuenta no solo su eficacia técnica, sino también su adecuación a los niveles de explicación propios de la psicología.

En segundo lugar, se ha argumentado que los modelos cognitivos basados en simulaciones, si bien poseen una gran capacidad para reproducir conductas, explorar escenarios contrafácticos y formalizar hipótesis complejas, no pueden ser considerados explicaciones psicológicas completas sin una serie de condiciones adicionales. Simular un comportamiento no equivale, por sí mismo, a explicarlo. Para que una simulación tenga valor explicativo debe cumplir criterios de validez interna, externa, constructiva y predictiva; debe estar integrada en un marco teórico sólido que le proporcione significado; y debe establecer alguna forma de correspondencia, ya sea estructural o funcional, con los procesos reales que ocurren en la mente humana. En ausencia de estas condiciones, el modelo corre el

riesgo de convertirse en una reconstrucción artificial que, aunque útil, no proporciona comprensión genuina del fenómeno que reproduce.

Este trabajo no está exento de limitaciones. En primer lugar, la discusión se ha desarrollado fundamentalmente en el plano teórico y filosófico, sin entrar en el análisis detallado de simulaciones concretas más allá de referencias generales a modelos conocidos. Un enfoque más empírico o comparativo, que examinara casos específicos de éxito y fracaso, podría haber complementado el análisis filosófico con datos de primera mano. En segundo lugar, se ha centrado en la psicología cognitiva y en la modelización computacional, dejando al margen otras formas de explicación psicológica —como la fenomenología, el enactivismo, etc.— que también podrían aportar perspectivas relevantes al problema de la comprensión de la mente. Aunque estas omisiones se justifican por la necesidad de acotar el objeto de estudio, señalan vías posibles para investigaciones futuras, como en una tesis doctoral, en donde se puedan abordar preocupaciones que, aun relevantes, no pueden ser investigadas en el presente trabajo fin de máster, como de igual modo la posibilidad de un pluralismo explicativo en psicología que no incurra en un eclecticismo estéril.

Asimismo, respecto de lo dicho en este trabajo, es necesario evaluar en un futuro la viabilidad de utilizar múltiples tipos de explicaciones de manera conjunta. Si bien la combinación y complementariedad de distintos enfoques explicativos es una práctica habitual en la investigación científica que puede enriquecer la comprensión, también genera importantes problemas epistemológicos. Estos problemas incluyen la posibilidad de incompatibilidades e incoherencias entre las explicaciones, así como la dificultad de determinar si un tipo de explicación posee mayor relevancia que otros. Por ello, es crucial un análisis filosófico que aborde la validez y coherencia de tales prácticas, explorando cómo se pueden combinar y complementar de forma rigurosa diferentes modos de explicación sin comprometer la solidez del conocimiento. Vías futuras de investigación estarían dirigidas a analizar estas cuestiones, así como al desarrollo de criterios más precisos y consensuados para la validación epistémica de simulaciones cognitivas, que integren elementos empíricos, formales y teóricos de diversa índole. La exploración de modelos híbridos, que combinen simulación computacional con enfoques interpretativos, especialmente en contextos clínicos o educativos donde la dimensión subjetiva del agente es esencial en la actualidad. De modo que cabe asimismo hacer un examen crítico de las implicaciones éticas, sociales y políticas de utilizar simulaciones como base para la toma de decisiones, pues estas afectan a personas reales, en ámbitos como la salud mental, la justicia, la selección de personal, etc.

BIBLIOGRAFÍA

Alvargonzález, D. (2024), *La filosofía de Gustavo Bueno. Comentarios críticos*. Universidad de Oviedo.

Amoretti, M. C., y Serpico, D. (2024). Filosofía de la ciencia: palabras clave. Alianza Editorial.

Bechtel, W., & Wright, C. D. (2009). What is psychological explanation? En J. Symons & P. Calvo (Eds.), *The Routledge Companion to Philosophy of Psychology* (pp. 113-130). Routledge.

Bengio, E., Jain, M., Korablyov, M., Precup, D., y and Bengio, Y. (2021). Network based Generative Models for Non-Iterative Diverse Candidate Generation. *ArXiv* (*Cornell University*), *34* https://arxiv.org/abs/2106.04399

Blum, C., Winfield, A. F. T., & Hafner, V. V. (2018). Simulation-Based Internal Models for Safer Robots. *Frontiers In Robotics And AI*, 4: 74.

Brandom, R. B. (1994). *Making It Explicit: Reasoning, Representing, and Discursive Commitment.* Harvard University Press.

Bunge, M. (1977). Emergence and the mind. *Neuroscience.* 2(4): 501-509.

Bunge, M. (2003), *Emergence and Convergence: Qualitative Novelty and the Unity of Knowledge*. University of Toronto Press.

Cartwright, N. (2021). Rigour versus the need for evidential diversity. *Synthese*, 199: 13095-13119.

Cassini, A., & Redmond, J. (2021). *Models and Idealizations in Science: Artifactual and Fictional Approaches*. Springer Nature.

Chirimuuta, M. (2024). *The Brain Abstracted: Simplification in the History and Philosophy of Neuroscience*. MIT Press.

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

Davies, M., & Stone, T. (Eds.). (2001). *Mental Simulation*. Blackwell.

Dennet, D. (1996). *Kinds of Minds: Toward an Understanding of Consciousness*. Penguin Random House.

Dennett, D. (2006). Personal and subpersonal levels of explanation. En J. L. Bermúdez (Ed.). *Philosophy of Psychology: Contemporary Readings* (pp. 17-21). Routledge.

Dretske, F. (1988). *Explaining behavior: Reasons in a World of Causes*. MIT Press.

Frigg, R. (2022). *Models and Theories: A Philosophical Inquiry*. Taylor & Francis.

Frigg, R., y Hartmann, S. (2021). "Models in Science", *The Stanford Encyclopedia of Philosophy* (Summer 2025 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = <https://plato.stanford.edu/archives/sum2025/entries/models-science/>.

Frigg, R., y Nguyen, J. (2020). *Modelling Nature: An Opinionated Introduction to Scientific Representation*. Springer Nature.

Fodor, J. A. (1968). *Psychological explanation.* Random House.

García, R. (2006). *Sistemas complejos: Conceptos, métodos y fundamentación epistemológica de la investigación interdisciplinarian.* Editorial Gedisa.

Garnham, A. (2013). *Mental Models and the Interpretation of Anaphora*. Psychology Press.

Garnham, A., & Oakhill, J. (2013). *Mental Models in Cognitive Science: Essays in Honour of Phil Johnson-Laird*. Psychology Press.

Gershman, S. J., & Goodman, N. D. (2014). Amortized inference in probabilistic reasoning. *Proceedings of the Cognitive Science Society*, 36: 1556-1561.

Giere, R. N. (2004). How models are used to represent reality. Philosophy of Science 71 (5):742-752

Gilbert, N. y Troitzsch, K. G. (2005). *Simulation for the social scientist.* Second Edition. Open University Press.

Grimm, V. *et al.* Pattern-Oriented Modeling of Agent-Based Complex Systems: Lessons from Ecology. *Science* 310,987-991(2005). DOI:10.1126/science.1116681

Goldman, A. I. (2006). *Simulating minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. Oxford University Press.

Grush, R. (1995). *Emulation and Cognition.* Dissertation, University of California, San Diego.

Grush, R. (2004). The emulation theory of representation: Motor control, imagery, and perception. *Behavioral and Brain Sciences*, 27(3): 377-396.

Hempel, C. G. (2005). *La explicación científica: Estudios sobre la filosofía de la ciencia*. Grupo Planeta (GBS).

Hochstein, E. (2016). One mechanism, many models: A distributed theory of mechanistic explanation. *Synthese*, 193(5): 1387-1407.

Jaworski, W. (2011). *Philosophy of Mind: A Comprehensive Introduction*. John Wiley & Sons.

Johnson-Laird, P. N. (1983). *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Harvard University Press.

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). *Ruido: Un fallo en el juicio humano*. Debate.

Kellert, S. H., Longino, H., y Waters, C. K. (eds.) (2006). *Scientific Pluralism.* University Minnesota Press.

Kripke, S. A. (2006). Wittgenstein a propósito de reglas y lenguaje privado: una exposición elemental. Tecnos Editorial S A.

Lebiere, C., Anderson, J. R., & Gonzalez, C. (2025). Cognitive Models for Machine Theory of Mind. *Cognitive Systems Research*, 65: 101-115.

Levine, J. (1983). "Materialism and qualia: The explanatory gap". *Pacific Philosophical Quarterly* 64. pp. 354-361.

Lin, Z. (2025). Large Language Models as Psychological Simulators: A Methodological Guide. *Journal of Artificial Intelligence Research*, 72: 1-28.

Lipton, Peter (1991), *Inference to the Best Explanation*, London, Routledge.

Little, D. (1998) *Microfoundations, Method and Causation*: On the Philossophy of the Social Science. Editorial Routledge.

Machamer, et al. (2000), "Thinking about mechanisms", *Philosophy of Science,* 67, 1: 1-25.

McDowell, J. (1994). *Mind and World.* Harvard University Press.

Michlmayr, L. (2002). *Simulation Theory versus Theory Theory*. En R. Kuznar (Ed.), Folk Psychology and Beyond (pp. 45-62). Routledge.

Marr, D. (2010). "Vision: A Computational Investigation into the Human Representation and Processing of Visual Information". The MIT Press.

Morgan, M. S., & Morrison, M. (1999). *Models as Mediators: Perspectives on Natural and Social Science*. Cambridge University Press.

Nagel, T. (1974). What Is It Like to Be a Bat? *The Philosophical Review*, *83*(4), 435–450. https://doi.org/10.2307/2183914.

Nicolescu, M., Blankenburg, J., Anima, B. A., Zagainova, M., Hoseini, P., Nicolescu, M., & Feil-Seifer, D. (2025). Simulation theory of mind for heterogeneous human-robot teams. *Frontiers In Robotics And AI*, *12*. https://doi.org/10.3389/frobt.2025.1533054

Papineau, D. (2011). "What exactly is the explanatory gap?". *Philosophia* 39. pp. 5-19.

Pérez-Jara, J., & Camprubí Bueno, L. (2025). Materialismo, correlacionismo y emergencia: Una evaluación filosófica de la ontología y epistemología de Gustavo Bueno. *Eikasía Revista De Filosofía*, (128): 37–122.

Pinker, S. (2010). *The Language Instinct: How The Mind Creates Language*. Harper Collins.

Priest, S. (1994). *Teorías y filosofías de la mente.* Cátedra.

Putnam, H. (1975). *Mind, Language and Reality: Philosophical Papers*. Vol. 2. Cambridge University Press.

Quine, W. V. O. (1953). *From a Logical Point of View.* Harvard University Press.

Rabinowitz, N. C., Perbet, F., Song, F., Zhang, C., Eslami, S. M. A., & Botvinick, M. (2018). Machine Theory of Mind. En *Proceedings of the 35th International Conference on Machine Learning* (ICML 2018), 70: 4218-4227.

Romero, G. E. (2018). *Scientific Philosophy*. Springer.

Salmon, W. C. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton University Press.

Salmon, W. C. (1998). *Causality and Explanation*. Oxford University Press.

Salmon, W. C. (2006). *Four Decades of Scientific Explanation*. University of Pittsburgh Press.

Searle, J. (1980). Minds, Brains and Programs. *Behavioral and Brain Sciences*, 3(3): 417-424.

Sellars, W. (1997). *Empiricism and the Philosophy of Mind*. Harvard University Press.

Stanford, K. (2009, 2023). "Underdetermination of Scientific Theory". *The Stanford Encyclopedia of Philosophy* (Summer 2023 Edition). Edward N. Zalta & Uri Nodelman (eds.). URL = <https://plato.stanford.edu/archives/sum2023/entries/scientific-underdetermination/>.

Stich, S., & Nichols, S. (1992). "Folk Psychology: Simulation or Tacit Theory?". En M. Davies & T. Stone (Eds.), *Mental Simulation*. Blackwell.

Suárez, M. (2024). *Inference and Representation: A Study in Modeling Science*. University of Chicago Press.

Sun, R. (2005). Cognition and Multi-Agent Interaction: From Cognitive Modeling to Social Simulation. *The Cambridge handbook of computational psychology* (pp. 530–548). Cambridge University Press. https://doi.org/10.1017/CBO9780511816772.023

Turing, A. M., Putnam, H. & Davidson, D. (1985). *Mentes y máquinas*. Editorial Tecnos.

von Neumann, J., & Burks, A. W. (Eds.). (1966). *Theory of Self-Reproducing Automata*. University of Illinois Press.

Varela, F. J., Rosch, E., & Thompson, E. (1992). *The embodied mind: Cognitive Science and Human Experience*. MIT Press.

Winsberg, E. (2010). *Science in the Age of Computer Simulation*. University of Chicago Press.

Wittgenstein, L. (2021). *Investigaciones filosóficas*. Trotta.

Wolpert, D. M., Ghahramani, Z., & Jordan, M. I. (1995). An internal model for sensorimotor integration. *Science*, 269(5232): 1880–1882.

Wood, M. D., Thorne, S., Kovacs, D., Butte, G., & Linkov, I. (2016). *Mental Modeling Approach: Risk Management Application Case Studies*. Springer.