THEORIA

Vol. 40/2 • May 2025

Established in 1952 by Miguel Sánchez-Mazas
Second Series

An International Journal for Theory, History and Foundations of Science

Ὁ Θεὸς ἀριθμετίζει

CALIJ
Centro de Análisis, Lógica e Informática Jurídica (CALIJ)

eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea

https://ojs.ehu.eus/index.php/THEORIA

THEORIA

REVISTA DE TEORÍA, HISTORIA Y FUNDAMENTOS DE LA CIENCIA

AN INTERNATIONAL JOURNAL FOR THEORY, HISTORY AND FOUNDATIONS OF SCIENCE

ESTABLISHED in 1952 by MIGUEL SÁNCHEZ-MAZAS Second Series

Mailing address: CALIJ-THEORIA, Elhuyar plaza 2 (20018 Donostia/San Sebastián, Spain)
Tel.: (+34) 943 017 447. Fax: (+34) 943 015 990. E-mail: editor.theoria@ehu.eus

THEORIA is a four-monthly journal (issues in January, May and September). The contents of this journal are covered, among others, by the following abstracting and indexing services: Arts & Humanities Citation Index®, ISI Alerting Services, Current Contents® / Arts & Humanities, Bulletin Signalétique 519, DICE, ICYT, ISOC and RESH of the Spanish research center CSIC, Mathematical Reviews, Current Mathematical Publications, MathSci, Philosopher's Index, Repertoire bibliographique de la Philosophie, and SCOPUS.

THEORIA's website

http://www.ehu.eus/theoria

Contents

Volume 40/2, May 2025, pp. 105-233

ISSN 0495-4548
e-ISSN 2171-679X

Articles

Mariela Rubin
Condicionales indicativos e intuicionistas / *Indicative conditionals and intuitionistic* 107

Dubian Cañas
Agencia, relativismo y conocimiento extendido / *Agency, relativism and extended knowledge* 130

Cristian López
The metaphysics of symmetries / *La metafísica de las simetrías: un mapa*.......... 154

José V. Hernández-Conde, María Caamaño-Alegre
Models of data and the representation of phenomena / *Modelos de datos y la representación de fenómenos: una inferencia basada en patrones* 172

Juan Guzmán, Jorge Manero
The role played by scientific realism in the measurement problem of quantum mechanics / *El rol que desempeña el realismo científico en el problema de la medición de la mecánica cuántica* 187

Vincenzo Crupi, Saúl Pérez-González
Mapping the communication of science / *Mapeando la comunicación científica* 210

Summary 232

Articles

Theoria, 2025, 40(2), 107-129
https://doi.org/10.1387/theoria.26370

THEORIA
ISSN 0495-4548 – eISSN 2171-679X

Condicionales indicativos e intuicionistas

(Indicative conditionals and intuitionistic)

Mariela Rubin*
IIF-SADAF/CONICET - Universidad de Buenos Aires
https://orcid.org/0000-0002-9392-3618

Palabras clave
Condicionales
Condicionales epistémicos
Condicionales Intuicionistas
Condicionales Indicativos

RESUMEN: En este trabajo voy a defender que el condicional intuicionista es tan buen modelo del condicional indicativo como cualquier otro condicional que se ha propuesto en la literatura. Para esto voy a presentar tres argumentos: (1) voy a argumentar que el Modus Ponens es una regla necesaria para comprender el condicional indicativo, (2) voy a defender que la regla de Introducción del condicional es la mejor regla de introducción que puede darse para un condicional, y (3) voy a mostrar que estas dos reglas no caracterizan solamente al condicional material sino que hay una cantidad infinita de condicionales que las validan. Y aún más, el condicional intuicionista en particular validará una serie de *desiderata* deseadas en la literatura sobre condicionales indicativos además de tener probabilidades que validan una versión de la Tesis de Stalnaker y una semántica que enriquece el condicional indicativo con una dimensión epistémica.

Keywords
Conditionals
Epistemic conditionals
Intuitionistic conditionals
Indicative conditionals

ABSTRACT: In this work, I will defend that the intuitionistic conditional is at least as good as a model of the indicative conditional as any other proposal in the literature. To do so I will present three main arguments: (1) I will argue that Modus Ponens is a necessary rule to understand indicative conditionals, (2) I will defend that the Introduction rule is the best rule to model a conditional, and (3) I will show that these two rules don't characterize necessarily the material conditional, rather there are infinitely many conditionals that satisfy them. Even more, the intuitionistic conditional in particular will validate several desiderata posed by the literature on conditionals, as well as having probabilities that match a version of Stalnaker's Thesis and a semantics that will add an epistemic flavor to the indicative conditional.

1. *Introducción*

En este trabajo voy a defender que el condicional intuicionista es un buen modelo del condicional indicativo. En particular, voy a argumentar que dada una postura inferencialista del significado (es decir, entender que el significado está determinado por el uso), tenemos buenas razones para defender que el par de reglas de Introducción y Eliminación del condicional representan el condicional indicativo (es decir, el condicional que utilizamos en el lenguaje natural). Dado un lenguaje proposicional o de primer orden y un cálculo de deducción natural, donde las letras griegas representan meta-variables y → representa un condicional indicativo, estas son las reglas usuales de Introducción y Eliminación:

* ***Correspondence to:*** Mariela Rubin. Bulnes 642, CP: 1176, Buenos Aires, Argentina – marubin@gmail.com – https://orcid.org/0000-0002-9392-3618

How to cite: Rubin, Mariela (2025). «Condicionales indicativos e intuicionistas»; *Theoria. An International Journal for Theory, History and Foundations of Science,* 40(2), 107-129. (https://doi.org/10.1387/theoria.26370).

Received: 30 May, 2024; Final version: 24 July, 2025.
ISSN 0495-4548 - eISSN 2171-679X

$$\dfrac{\begin{matrix}[A]^1\\ \vdots\\ B\end{matrix}}{A \to B}\,[1]\ \mathrm{I}\text{–}\!\to \qquad\qquad \dfrac{A \to B \quad A}{B}\ E\text{–}\!\to$$

La regla de Introducción nos dice que si suponemos el antecedente, A, (eso representan los corchetes, que A es una suposición) y llegamos mediante la utilización de las reglas permitidas por el cálculo a B, entonces podemos inferir $A \to B$. El [1] a la derecha de la raya de consecuencia lógica (la raya horizontal) nos indica que el supuesto 1, el supuesto de que A es el caso, cierra con la Introducción del condicional. La regla de Eliminación es la famosa regla de Modus Ponens, nos dice que si tenemos un condicional, $A \to B$, y tenemos el antecedente, A, podemos inferir el consecuente, B. Usualmente, suele decirse de forma apresurada que estas reglas representan el condicional material. El condicional material es el condicional de la lógica clásica que se puede representar mediante las siguientes tablas de verdad y que graficaré con el símbolo $\supset$:

Table 1

Condicional Material

A	$\supset$	B
1	1	1
1	0	0
0	1	1
0	1	0

Sin embargo, no es secreto que para que estas reglas representen el condicional material es necesario que las reglas que caracterizan a la negación sean las reglas de la negación clásica, y que la definición de prueba absorba las propiedades estructurales (es decir, reflexividad, monotonía y transitividad). Dicho de otro modo, el par de reglas I-$\to$ y E-$\to$ caracteriza una clase de condicionales, entre los que se encuentran el condicional material pero también el intuicionista y el minimal[1].

Los condicionales indicativos son aquellos condicionales de la forma «Si se da A, entonces se da B» —con énfasis en que quien emite el condicional asume o cree o sabe que el antecedente es posible. Este énfasis se marca para contraponer esta clase de condicionales con los condicionales contrafácticos, es decir, aquellos condicionales de la forma «Si hubiese pasado A, habría pasado B»— que asumen que el antecedente del condicional es falso. A lo largo de la literatura sobre condicionales indicativos se han elaborado muchísimas propuestas para modelar este término[2]. Todas ellas iluminan algún aspecto del condicional, pero ninguna termina de presentar una propuesta completamente satisfactoria. Evidencia de ello es que al día de hoy nuevas teorías siguen proliferando. Algunos ejemplos de teorías recientes son las de Egré *et al.* (2024), Douven (2023) o Iacona (2023), entre otras. La mayor parte de las teorías en la literatura se concentran en explorar las condiciones de verdad de los condicionales. Quizás con las excepciones de (a) el inferencialismo de Douven (una postura psicologista de los condicionales, que debe diferenciarse del inferencialismo del que voy a hablar

[1] Este argumento puede reconstruirse de forma análoga si tomamos un cálculo de secuentes, si se toman las reglas usuales para el condicional y se restringe la cantidad de conclusiones que puede tener el secuente. En este caso se debe argumentar que las reglas del condicional material y las del intuicionista son las mismas reglas, pero que lo que cambia es el contexto de deducción que está determinado parcialmente por las estructura del secuente.

[2] Ver (Edgington, 2020) para un resumen bastante exhaustivo de las teorías existentes.

a continuación) o (b) las semánticas dinámicas al estilo de Gillies (2004) o Yalcin (2007) (que presentan otra forma de entender la consecuencia lógica, en términos de preservación de cuerpos de información). En este trabajo me voy a concentrar en el conjunto de reglas como método de caracterizar el significado de una conectiva, es decir, voy a sostener una tesis inferencialista del significado. Y voy a defender que el mejor conjunto de reglas para modelar el significado del condicional indicativo está dado por el par I-$\rightarrow$ y E-$\rightarrow$. Pero no en tanto reglas que caracterizan el condicional material sino en tanto reglas que caracterizan el condicional intuicionista. Mi intención es dar nuevos argumentos en favor del condicional intuicionista. Normalmente los argumentos en favor de este condicional vienen desde la literatura lógica. Es decir, desde la idea de que los condicionales que se utilizan en matemáticas son o deberían ser intuicionistas. El mayor exponente de estos argumentos es Dummett (1991). El objetivo de este texto es mostrar que hay buenos argumentos clásicos de la literatura sobre condicionales indicativos que también apoyan este condicional. Puesto que tiene una interpretación epistémica, tal y como muchas autoras sostienen respecto de los condicionales indicativos (Edgington, 1995; Égré *et al.*, 2021; Gillies, 2004), y que las reglas que lo caracterizan nos permiten derivar un conjunto de inferencias que la literatura suele desear que los condicionales indicativos satisfagan. A su vez, como veremos, también tiene la propiedad de poder hacer coincidir la probabilidad del condicional con su probabilidad condicional, una de las desideratas más comunes en la literatura, conocida a veces como la Tesis de Stalnaker.

La estructura del trabajo es la siguiente: en la sección 2 voy a presentar la propuesta inferencialista. Luego, procederé a argumentar primero en favor de la regla de Modus Ponens como una regla necesaria para la comprensión y la correcta caracterización de cualquier condicional. Luego, voy a argumentar en favor de la regla de Introducción como un buen modelo del Test de Ramsey. Finalmente, voy a dar razones en favor del condicional intuicionista, primero respondiendo posibles críticas y luego, mostrando cómo funcionan las probabilidades intuicionistas para finalmente hablar de los modelos de Kripke que le agregan una dimensión epistémica.

2. *Semántica Inferencialista*

En esta sección voy a explicar qué son las semánticas inferencialistas dado que a lo largo de este texto voy a asumir una postura inferencialista para defender la tesis propuesta. La literatura sobre condicionales indicativos es muy variada. Hay mucho escrito desde la lógica, la filosofía del lenguaje, la epistemología formal y también desde la psicología. En los textos de lógica sobre condicionales indicativos suele trabajarse desde posturas referencialistas, es decir, desde posturas que entienden que el significado de un término está dado por aquello que denota, que en el caso de las conectivas lógicas son sus condiciones de verdad. A lo largo de toda esta literatura es muy difícil encontrar consenso entre autoras, por fuera de la importancia que tienen los condicionales indicativos en nuestras prácticas inferenciales. En este contexto, propongo que nos centremos en otro tipo de semánticas, semánticas en las que el significado de los términos va de la mano de su rol inferencial. Esto no debe leerse necesariamente como un argumento en favor de la tesis inferencialista, sino más bien como una propuesta metodológica. Dado que las semánticas referencialistas no parecieran ponerse de acuerdo en cuál es el significado del condicional, quizás explorar la pregunta por su significado desde otro tipo de semántica ayude a esclarecer el problema.

La tesis inferencialista se opone a las semánticas referencialistas en tanto que entiende que las palabras adquieren su significado en función de los contextos en los que ocurren. Más específicamente, la tesis semántica que defiende el inferencialismo nos dice que:

Definición 1. Tesis Semántica Inferencialista: el significado de las palabras está dado por sus condiciones de uso.

Esta tesis puede ser una tesis general sobre todo el lenguaje, como bien se sugiere en Brandom (1994; 2009), que es quien acuña el término «inferencialismo», o bien puede ser una tesis sobre algunos términos específicos. Puede limitarse al vocabulario lógico, al mejor estilo de los textos de Belnap (1962), Dummett (1991) o Prawitz (2006), en cuyo caso será una tesis sobre inferencialismo en lógica.

Definición 2. Tesis Semántica Inferencialista en Lógica: el significado de las constantes lógicas está determinado por sus reglas de inferencia.

Antes de continuar, voy argumentar que quizás haya una versión intermedia entre el inferencialismo en lógica y el inferencialismo generalizado en el que entendamos que el significado de ciertos términos funcionales está dado por sus reglas. Es decir, el inferencialismo en este contexto no sería una tesis exclusiva de las conectivas lógicas, sino de ciertos términos cuyo significado está dado por la función que ellos cumplen en la oración.[3] Para poner un ejemplo, el término «mesa» tiene un referente icónico, una tabla con cuatro patas que suele estar rodeada de sillas que todas podemos imaginar cuando pensamos en una mesa. Palabras como «y», «no», «si, entonces», «a», «con», «pues»,[4] no tienen un significado que refiera a una realidad extralingüística. Se suele decir que lo que caracteriza a estas palabras es que su significado es funcional. Es decir, significan en tanto que cumplen una función en las oraciones en las que participan. Las semánticas referencialistas explican esta función a partir de la contribución que hacen estas expresiones a las condiciones de verdad de los enunciados. Lo que voy a proponer ahora es que pensemos la semántica de algunos términos funcionales a partir de su rol inferencial. Esto es:

Definición 3. Tesis Semántica Inferencialista Funcional: el significado de las expresiones funcionales está determinado por sus reglas de inferencia.

Esto lo planteo aquí porque hay dos discusiones transversales al inferencialismo relacionadas con qué conectivas son lógicas y cuáles no, y en particular, una discusión sobre si los condicionales indicativos son conectivas lógicas. No voy a indagar en estos problemas por cuestiones de espacio. Sin embargo, me voy a limitar a proponer que, sean lógicos o no, se da lo siguiente:

Definición 4. Tesis Semántica Inferencialista sobre los Condicionales Indicativos: el significado de los condicionales indicativos, como expresiones funcionales que son, está determinado por sus reglas de inferencia.

Las razones para adoptar una tesis inferencialista del significado vienen desde lugares diferentes. Sin embargo, creo que basta para motivar esta propuesta la mera idea de que es una tesis semántica con muchísimo desarrollo en la literatura, pero que sin embargo no ha sido explorada en lo que refiere a condicionales indicativos.

Aquí es importante remarcar dos cosas: en la literatura sobre condicionales hay cierto consenso *naïve* respecto de qué inferencias una querría que valide el condicional. En general se espera que valide Modus Ponens ($A \rightarrow B, A \vDash B$, donde $\vDash$ representa la noción de consecuencia lógica), Import-Export ($A \rightarrow (B \rightarrow C) \equiv (A \& B) \rightarrow C$, donde $\equiv$ representa la noción de equivalencia lógica) y que sea estrictamente más fuerte que el condicional material ($A \rightarrow B \vDash A \supset C$). Es decir, que haya valuaciones que hagan falso al condicional indicativo cuando el condicional material es verdadero. Las motivaciones para que valide Modus Ponens (MP, de aquí en adelante) las veremos más adelante, pero basta con decir que hay muchísimos estudios empíricos que demuestran que una de las pocas reglas de inferencia que la gente suele seguir correctamente es MP —ver (Oberauer & Wilhelm, 2003) para más información. Import-Export es la equivalencia que suele permitirnos traducir condicionales anidados en oraciones más sencillas, y es una equivalencia sumamente discutida en la literatura —ver (Égré *et al.,* 2023). En tanto que la idea de ser estrictamente más fuerte que el condicional material surge de mirar su tabla de verdad. La idea principal es que muchísimos condicionales indicativos son falsos a pesar de tener antecedente falso.

Esto es un punto de partida importante, dado que tal consenso *naïve* no existe entre las teorías referencialistas, en las que ni siquiera hay acuerdo respecto de si los condicionales tienen condiciones de verdad —ver (Edgington, 1995)

[3] Para una buena explicación de qué es un término funcional ver (Fries, 1952; Hallebeek, 1986)

[4] Y quizás, de forma más controvertida podríamos incluir a palabras como «verdad» en este conjunto.

para más información. En ese sentido, creo que proponer una semántica inferencialista mueve la falta de consenso a un nivel metodológico. La idea es que, dado que no hay acuerdo sobre las condiciones de verdad, pensemos el significado en términos de inferencias válidas, sobre las que tenemos al menos alguna base común.

En términos generales, la propuesta inferencialista entabla una relación entre el significado y las reglas de inferencia. Esta relación puede ser de constitución o una relación más epistémica de entendimiento, en la que las reglas expresan un significado que está afuera. Siguiendo a Buacar (2015) podemos distinguir dos caras de la tesis inferencialista:

Definición 5. Tesis Semántica: el significado de las expresiones se agota en sus reglas. Esto es, las reglas de inferencia definen el significado de las expresiones lógicas.

Definición 6. Tesis Epistémica: conocer el significado de una expresión se agota en saber realizar las inferencias correspondientes.

Dependiendo del tipo de inferencialismo que se defienda distintos conjuntos de reglas serán aptos para determinar el significado de una conectiva. Podemos distinguir tres tipos generales de inferencialismo: el atomismo, el molecularismo y el holismo. El atomismo va a defender que hay un par de reglas de Introducción y Eliminación que determinan el significado de cada conectiva, y solo ese par de reglas basta para determinar su significado. Esta es la postura más estándar dentro del inferencialismo en lógica, entre sus defensores podemos citar a (Dicher, 2020; Read, 2010; Steinberger, 2011), por nombrar algunos. Por ejemplo, ya conocemos las reglas del condicional minimal, intuicionista y material, es decir aquellas que presentamos en la Introducción. Para estos autores, este par de reglas basta para determinar el significado del condicional. En tanto que el molecularismo va a defender que el significado de una conectiva está dado, no solo por el par de reglas que determinan cómo usar esa conectiva, sino por el conjunto más grande de reglas que determinan el total del lenguaje. Así el significado del condicional, como defendía Dummett en (1991), va a estar expresado por las reglas de Introducción y Eliminación del condicional, pero también por las reglas de la negación, de la conjunción y de la disyunción. También podemos pensar que no son solo las reglas de las conectivas las que determinan el significado de los términos, sino que como veremos que defiende Belnap en (1962) las reglas estructurales también influyen. En este sentido, como bien señala unx referi anónimo, una puede defender un tipo de atomismo o molecularismo en el que el significado de las conectivas dependa solo del par de reglas que presenta la conectiva o también dependa de las reglas estructurales válidas en el sistema.

Por último, el holismo defiende que el significado de una expresión está dado por todas las inferencias válidas que la involucran. Esta postura es más usual en las tesis inferencialistas generales como las famosas posturas de Brandom (1994) y Sellars (1953), pero no tanto en el ámbito de la lógica. Puesto que para conocer el significado de una conectiva lógica, una debería conocer todas las inferencias válidas, así como todas sus instancias, cosa que es imposible para cualquier ser humano.

En la sección que aquí continúa, me voy a detener en un desafío que se le presentó a esta posición filosófica, dado que nos va a ayudar a entender qué es lo que queremos defender.

2.1. Tonk

En su famoso artículo «The Runabout Inference Ticket», Prior (1960) elabora una objeción muy famosa. Nos presenta una conectiva llamada tonk, $\oplus$, cuyas reglas son las siguientes:

$$\frac{A}{A \oplus B}\,\mathrm{I}-\oplus \qquad\qquad \frac{A \oplus B}{B}\,\mathrm{E}-\oplus$$

Como se puede ver, esta conectiva inmediatamente trivializa cualquier sistema (que pueda probar al menos un teorema y que sea transitivo):

$$\dfrac{\dfrac{A}{A \oplus B}\,\mathrm{I}-\oplus}{B}\,\mathrm{E}-\oplus$$

Ahora, ¿cuál es el problema de estas reglas? Gran parte de la literatura alrededor de tonk sigue a Stevenson (2018)[5] y argumenta que el problema de estas reglas es que no pueden caracterizarse en función de la contribución de tonk a las condiciones de verdad de los enunciados (dado que esta conectiva no puede caracterizarse con tablas de verdad deterministas). El significado para Stevenson está dado por las condiciones de verdad de una conectiva, que nos permiten recortar un conjunto correcto de reglas. Y que en el mejor de los casos nos ayudará a entender su significado. Otra respuesta, muy distinta, la propone Belnap (1962). En este texto nos dice que hay un significado anterior sobre el que se montan las reglas, pero que ese significado no son las condiciones de verdad de las conectivas, sino las propiedades estructurales de la noción de consecuencia lógica —que en el caso de la lógica clásica y la intuicionista son la reflexividad, la monotonía, la contracción, la conmutatividad y la transitividad. Las reglas estructurales son reglas que no nos hablan sobre un lenguaje específico sino sobre cómo funciona la consecuencia misma. La idea de Belnap es que un par de reglas en sí mismo no hace al significado, sino que cobra significado cuando se lo introduce en el contexto de una lógica que ya tiene una noción de consecuencia específica. Las propiedades de la noción de deducción son el contexto sobre el que se agregan las conectivas y este contexto impone un criterio de admisión: las reglas deben ser conservativas respecto del contexto de deducibilidad. Donde conservativo quiere decir que, si tenemos un lenguaje $\mathcal{L}$ y lo expandimos con una conectiva nueva, no podamos probar cosas sobre el lenguaje viejo que no podíamos probar antes. Para ver esto en relación con cosas que ya conocemos, la negación clásica presentada con las reglas de deducción natural[6] no es conservativa respecto del fragmento condicional, dado que al agregar las reglas de la negación podemos probar la Ley de Peirce, $\vdash \big((A \supset B) \supset A\big) \supset A$, una fórmula que no contiene negaciones, solo condicionales.[7]

El problema de tonk entonces, va a decirnos Belnap, es que no es conservativa. Supongamos que partimos de un lenguaje vacío, en el que las únicas reglas que tiene la lógica son reglas estructurales, digamos reflexividad, monotonía y transitividad. En aquella lógica, agregar tonk nos permite probar $A \vdash B$, para cualesquiera dos fórmulas y esta es una regla estructural nueva. Otra forma de verlo es que si tenemos lógica clásica y le agregamos tonk, podemos probar $A \vdash B$ para cualesquiera par de fórmulas. En particular, podemos probar que cualquier fórmula B se sigue de cualquier fórmula A, incluso para las fórmulas que no contienen tonk. Es decir, tonk no es conservativa sobre el lenguaje, pero tampoco es conservativa respecto de la noción de consecuencia.

5 Originalmente publicado en 1960.

6

$$\dfrac{\begin{array}{c}[A]^1\\ \vdots\\ \bot\end{array}}{\neg A}\,[1]\,\mathrm{I}-\neg \qquad\qquad \dfrac{A \qquad \neg A}{\bot}\,\mathrm{E}-\neg$$

$$\dfrac{\neg\neg A}{A}\,\mathrm{DN}$$

7 En una presentación en secuentes, dada una inferencia con múltiples premisas y múltiples conclusiones, es posible formular reglas independientes para cada una de las conectivas. En estos cálculos la posición de Belnap se vuelve aún más relevante: no basta el par de reglas aislada, en la medida que liberamos o restringimos nuestros secuentes (que vienen a representar la noción de consecuencia) el significado de las conectivas también va a variar. Volveré a esto en breve.

A las conectivas con la propiedad de ser conservativas muchas filósofas también las llaman *armónicas*, siguiendo la terminología de (Dummett, 1991). La propiedad de armonía, sin embargo, es un poco más difusa que la conservatividad. Dummett llama armónico a un par de reglas en la medida que las reglas de Eliminación estén balanceadas respecto de las reglas de Introducción. Esta idea puede determinarse de distintas maneras. Puede ser en términos de conservatividad y unicidad como hace Belnap[8]. También en términos de normalización, como propone Prawitz en (2006). O en términos de eliminación de corte, como propone Dicher en (2016), o mismo que satisfagan el proceso de eliminación general que define Read en (2010). El concepto de armonía es central en la discusión sobre inferencialismo, dado que es la vara para decidir qué conjuntos de reglas son aptas a la hora de determinar el significado de una conectiva.

Si volvemos al objetivo de este trabajo —defender el condicional intuicionista como modelo del condicional indicativo— es posible argumentar en favor del condicional intuicionista en tanto que sus reglas son armónicas (entendiendo armonía bajo cualquiera de estas acepciones). Sin embargo, esta es una propiedad específica de las conectivas lógicas, y en este trabajo estoy dejando abierta la posibilidad de que los condicionales indicativos puedan no ser conectivas lógicas (como menciono cuando introduzco la Definición 3). Por esta razón y por cuestiones de espacio, elijo no desarrollar esta línea argumental.

En relación a tonk, Prior explica en un segundo texto sobre tonk publicado en (1964) que esto no es un problema específico del inferencialismo sino de la lógica en general. Prior nos dice que es importante tener en mente la distinción entre un lenguaje formal significativo y un mero juego simbólico. En un juego simbólico no hay ningún significado que capturar y por lo tanto las reglas pueden agotar el significado. En ese sentido, ni dar tablas de verdad ni un contexto de deducción van a dotar de significado a aquellas expresiones, puesto que no hay nada detrás de ellas más que el mero juego formal propuesto. En el mejor de los casos, se presentarán propiedades de un *signo formador* de aquella conectiva que estemos caracterizando por tablas o reglas. En tanto que si hablamos de un lenguaje significativo el problema es aún peor, puesto que ni las reglas ni las tablas van a determinar el significado, todo lo contrario, lo van a presuponer. Dice Prior al respecto hablando de la conjunción:

> Cada uno de ellos nos dice algo que podría significar al decir que «y», por ejemplo, o «&», es un signo formador de conjunción. Pero ninguno de ellos nos dice lo que significa «y» o «&» en sí. (Prior, 1964, pp. 191-192).[9]

Y al respecto explica Buacar en su tesis doctoral:

> Las definiciones inferenciales definen efectivamente una clase de expresiones en tanto signo formador de conjunciones, pero «y» es algo diferente a un signo formador de conjunción, ese algo más al parecer no puede ser dado en términos de reglas o condiciones de verdad (Buacar, 2015, pp. 193)

Tonk entonces no tiene un problema dentro de un juego simbólico, excepto quizás que aquel juego es poco interesante. El problema con tonk es que la clase de símbolos formadores de tonk es vacía. Es decir, no refiere a nada. El punto de Prior en este texto es que solo lo que ya tiene significado puede describirse en términos formales, ya sea mediante tablas o mediante reglas. En el mejor de los casos, dice Prior, aquellas tablas o aquellas reglas pueden ayudarnos a comprender su significado. Es decir, las reglas van a tener un rol epistémico. De esta forma, las reglas o las tablas para $\wedge$ o & van a determinar una clase que nos ayudará a delimitar ciertos bordes del significado, pero no nos

[8] Unicidad es una propiedad que puede entenderse de varias formas, pero de manera muy breve puede decirse que una conectiva, $*$, está únicamente determinada si y solo si, si una agrega una conectiva nueva, $+$ que tiene las mismas reglas que $*$, entonces ambas son interderivables. Es decir, $A * B \dashv\vdash A + B$.

[9] «Each of these tells us something that could be meant by saying that "and", for instance, or "&", is a conjunction-forming sign. But neither of them tells us what is meant by "and" or by "&" itself.»

van a decir qué significan. Para Prior, tonk es una nota negativa: no hay reglas ni tablas que definan el significado, puesto que lo presuponen. Sin embargo, voy a seguir a Buacar en su forma de entender este problema:

> …considero que la dicotomía que plantea Prior no es tal, de lo que se trata es de explorar una tercera opción, intermedia, sobre el lenguaje simbólico. Desde esta nueva perspectiva, si bien las expresiones lógicas se formulan en un sistema formal, ellas pretenden capturar el sentido de las expresiones lógicas del lenguaje natural (o al menos un fragmento de éste). (Buacar, 2015, pp. 194)

Lo que vamos a estar buscando en este trabajo, entonces, es el conjunto de reglas que capturen el significado del condicional indicativo que utilizamos en el lenguaje natural. En particular, en las secciones siguientes voy a defender que quien no infiere de acuerdo a la regla de Eliminación no comprende cómo utilizar los condicionales. También defenderé que la regla de Introducción es la mejor y la más sencilla a la hora de determinar el comportamiento hipotético que caracteriza a esta conectiva. Pero como ya hemos visto, no basta con defender el par de reglas para defender el condicional intuicionista. Por eso, a continuación voy a defender que la lógica intuicionista captura al menos dos dimensiones epistémicas importantes para los condicionales indicativos.

Para resumir lo que se expuso hasta aquí, vamos a entender que las reglas permiten delimitar el significado de nuestro condicional, que es un término cuyo significado conocemos en un sentido vago y a través de nuestras prácticas lingüísticas cotidianas. En tanto que las reglas que defendamos van a estar capturando aquel significado que, en el espíritu de (Williamson, 2020), está inmerso en una práctica caótica sobre la que tenemos que normar para poder distinguir los buenos usos de los usos errados.

En lo que sigue voy a defender que tanto la regla de Eliminación como la regla de Introducción son buenas (si no las mejores) candidatas para modelar el comportamiento del condicional indicativo. Luego, argumentaré en favor del condicional intuicionista en general. Comienzo por la regla de Eliminación ya que es quizás la regla más famosa que dio la lógica, el Modus Ponens.

3. *Una defensa del Modus Ponens*

En su tesis doctoral, Birman (2015), que sigue el problema presentado por Kripke en (2023), nos presenta a un personaje curioso. Harry es una persona que nunca hizo una inferencia siguiendo MP y no conoce sus principios en absoluto.[10] La pregunta que se hace Birman es ¿qué pasa si tratamos de enseñarle a Harry a usar MP? Es decir, intentamos que *adopte* la regla. En este contexto adoptar una regla de inferencia, nos dicen Kripke y también Birman, significa inferir de acuerdo a MP tras aceptar su validez.[11]

Supongamos que le explicamos a Harry lo siguiente: «los enunciados condicionales junto con el antecedente del condicional siempre implican el consecuente» y supongamos que Harry nos dice que entiende lo que le decimos. Luego, procedemos a testear si Harry adoptó el MP. Le preguntamos: «¿Sabías que si hoy llueve comeremos tortas fritas?», «No», nos dice, «pero les creo.» «¿Y sabías que hoy llueve?», «No», nos dice, «pero les creo.» Y luego le preguntamos «¿Comeremos tortas fritas?» A lo que Harry nos responde: «No lo sé, debería fijarme qué compraron para comer.» El escenario es coherente con la premisa, puesto que si Harry no puede inferir de acuerdo a la regla de MP, presentarle la regla de forma explicita no va a servir de nada, puesto que la regla de MP involucra ejecutar un MP.

---

10 En el texto original el problema se presenta tanto con MP como con Eliminación del cuantificador Universal, es decir, que de «Para todo x Px», se sigue que «Pa.» Elijo presentar el argumento solo en términos de MP por cuestiones de espacio.

11 Para una buena explicación del concepto de «de acuerdo a» en el contexto del Problema de la Adopción ver (Fiore, 2022).

Otra forma de entender lo mismo es que una de las premisas implícitas del MP es la regla misma como se explica claramente en (Fiore, 2022).

La conclusión[12] que extraen Kripke y Birman de esta historia es la siguiente:

> Hay ciertas reglas que simplemente no podrías adoptar: no podrías decírte- las a tí mismo, porque si te las dijeras a vos mismo sin haberlas utilizado ya, serían inútiles; así que o no te ayudan o, de todos modos, eran superfluas. (Kripke (1974) citado en Birman, 2015, pp. 112)[13]

La idea principal del problema que nos proponen Kripke y Birman es que el MP (y también la regla de Eliminación del Universal) son reglas inadoptables porque la forma condicional de las reglas involucra utilizar MP para poder inferir algo de ella. De hecho, se podría decir que la realización exitosa de cualquier regla involucra ejecutar un MP (lo mismo sucede con la Eliminación del Universal). Por lo que, sigue Kripke, resulta razonable pensar que estas reglas son fundamentales e inadoptables. En el sentido de que o bien ya eran parte de nuestro arsenal de reglas o bien, si no las entendemos, no las vamos a poder entender al agregarlas explícitamente. Y aún más, si no las entendemos, no deberíamos poder aplicar ninguna regla de inferencia exitosamente.

Este argumento guarda una relación directa con el argumento de Aquiles y la Tortuga de Lewis Carroll (1895), en el que se plantea un desafío a la hora de justificar una inferencia que tiene la forma de MP. En la historia que nos presenta Lewis Carroll la Tortuga desafía a Aquiles a que la obligue a aceptar la conclusión del siguiente argumento:

(A) Las cosas que son iguales a una tercera son iguales entre sí.

(B) Los dos lados de este triángulo son iguales a un tercero.

(Z) Los dos lados de este triángulo son iguales entre sí.

Dado que (Z) se sigue lógicamente de (A) y (B), quien comprende este argumento tiene la siguiente obligación: si acepta la verdad de (A) y (B) debe aceptar la verdad de (Z). Sin embargo, la Tortuga le dice a Aquiles que acepta las premisas pero no acepta la conclusión. Por lo que Aquiles le pide a la Tortuga que acepte el siguiente enunciado hipotético:

> (C) «Si (A) y (B) son verdaderas, (Z) también lo es.»

La Tortuga lo acepta, pero le dice a Aquiles que aún así sigue sin aceptar la verdad de (Z). La historia sigue al infinito con Aquiles pidiéndole a la Tortuga que acepte más y más enunciados hipotéticos de la forma de (C), y la Tortuga que los acepta pero se niega a aceptar la conclusión del argumento.

Una forma de interpretar lo que está pasando en la interacción de estos dos personajes es que la Tortuga no entiende cómo utilizar la regla de Modus Ponens. Dice Brown: «En todo caso, la presunción es que por ignorancia, error o excentricidad no está haciendo un uso normal de sus palabras.» (Brown, 1954, pp. 175)[14]. Por lo que agregar enunciados condicionales que explicitan la regla solo intensifica el problema. En la medida que la Tortuga no acepte que de un condicional y su antecedente se sigue el consecuente, no habrá premisas que puedan agregarse que la ayuden a aceptarlo, dado que todas las premisas que se agreguen son instancias de esta regla de inferencia que la Tortuga está fallando en comprender. El argumento de Lewis Carroll apunta a la pregunta por la justificación del MP, y el énfa-

12 Una de las conclusiones ...

13 «There are certain rules which you just couldn't adopt: you couldn't tell them to yourself, because if you told them to yourself without already using them, they would be useless; so they either don't help you or they were superfluous anyway.»

14 «At any rate, the presumption is that through ignorance, mistake, or eccentricity he is not malting a standard use of his words.»

https://doi.org/10.1387/theoria.26370

sis de su argumento es que cualquier justificación posible es circular. En tanto que el argumento de Kripke y Birman se centra en la pregunta de si es posible hacer alguna inferencia si una no sabe usar MP. Sin embargo, la lectura de Brown apoya nuestro argumento: quien se niega a inferir el consecuente de un condicional aceptando el antecedente y el condicional es una persona que no entiende cómo usar el condicional.

El punto central de esta sección es que, independientemente de cuáles sean las otras reglas con las que definamos el condicional indicativo, las hablantes necesitan inferir de acuerdo con MP. En palabras de Finn (2019) esto se debe a la naturaleza auto-gobernada del MP. La idea es que cualquier regla de inferencia involucra ejecutar un MP. De lo que se sigue que inferir de acuerdo con MP involucra ejecutar un MP. Por lo que, de no poder inferir utilizando las reglas de MP, la hablante no podría utilizar correctamente las otras reglas del condicional —cualesquiera sean—. Si esta discusión la enmarcamos en la discusión sobre condicionales indicativos, no importa cuáles sean las reglas que gobiernen el comportamiento del condicional indicativo una de ellas tiene que ser MP.

Es cierto que alguien podría contrargumentar diciendo que el condicional involucrado en la formulación de las reglas de inferencia no es un condicional indicativo, sino material. Y respecto del condicional material no tenemos ninguna duda de cuáles son las reglas que gobiernan su significado. Sin embargo, y si bien es cierto que cuando hacemos lógica (y meta-lógica) pensamos qué condicional con el que están formuladas las reglas es el material, también es cierto que estas reglas que determinan el comportamiento de las conectivas lógicas tienen un rol —o una intención— normativa respecto de la forma en la que inferimos en nuestra vida cotidiana. Recordemos que el problema comienza porque nos damos cuenta de que Harry podría tener una vida mucho mejor si entendiese cómo usar MP. El problema de Harry no es que no conoce el condicional material; el problema de este personaje es que le estamos presentando enunciados con forma condicional y no puede realizar un MP, sea cual sea la naturaleza de ese condicional.

Pero incluso si negásemos el rol normativo de las reglas y asumiésemos que las reglas correctas son aquellas que describen la práctica de las hablantes, podríamos argumentar, siguiendo el espíritu de Brown, que Harry es una reducción al absurdo de personas que no entienden el condicional. El punto que quiero enfatizar es que nadie le pide a Harry que entienda el condicional material; lo que se le pide es que entienda *algún condicional*. Porque el MP es la propiedad más básica que caracteriza al condicional. Sin esta regla lo que tenemos son sujetos que solo existen en el universo escéptico, en el que alguien que hace como que no entiende nos pide que justifiquemos aquello que jamás dudaríamos que es verdad. Evidencia de esto hay de sobra en estudios empíricos que muestran que una de las pocas reglas lógicas que las hablantes saben hacer correctamente es MP —ver (Rips, 1994 pp. 177-178).

Hasta aquí argumenté que la regla de Eliminación del condicional determina su significado en tanto que no podemos entender un condicional si no sabemos usar MP. A su vez, argumenté que es una regla necesaria para entender cómo utilizar cualquier otra regla. En la próxima sección voy a argumentar que la regla de Introducción es la mejor representación del pensamiento hipotético, una propiedad sustantiva del condicional. Luego, voy a explicar por qué este par de reglas en el contexto de la lógica intuicionista nos da una doble lectura epistémica del condicional que lo acerca a los condicionales indicativos. Argumentaré que su interpretación nos habla de aquello que sabemos y aquello que podemos saber y que nos permite capturar nuestros grados de creencia en un condicional a partir de su probabilidad condicional.

4. *Una defensa de la Introducción hipotética*

Para comenzar esta sección quizás sirva pensar un poco en cómo se piensan los condicionales cuando hacemos lógica. A la literatura sobre condicionales podemos dividirla en tres grandes grupos: por un lado tenemos las teorías clásicas que defienden el condicional material, sus tres grandes defensores son Grice (1989), Jackson (1987) y Williamson (2020). Por otro lado están las teorías suposicionales que buscan describir al condicional indicativo pensándolo como una conectiva que reconstruye el proceso de hipotetizar, aquí se ubica la mayor cantidad de literatura, cito a Adams (1975), Edgington 1995), Stalnaker (1976), de Finetti (1936), Belnap (1970) por nombrar solo algunas autoras relativamente famosas, entre cientos de teorías. Y por último están las teorías de la *restrictor-view*, cuya defensora más famosa es Kratzer (1979) o Gillies (2004) y que suele tener mucho más desarrollo en el ámbito de la lingüística.

Todas las teorías suposicionales de los condicionales parten del Test de Ramsey como inspiración para pensar su funcionamiento (o sus condiciones de verdad). Volvamos a esta famosa cita que inaugura la gran mayoría de los textos sobre condicionales:

> Si dos personas discuten «Si *p*, entonces *q*...» y ambas tienen dudas sobre *p*, están añadiendo hipotéticamente *p* a su conjunto de conocimientos y argumentando sobre esa base acerca de *q*... Podemos decir que están fijando sus grados de creencia en *q* dado *p*. (Ramsey, 2016 pp. 143)[15]

Noten que agregar hipotéticamente *A* a aquello que sabemos y ver si de eso se sigue *B* es una forma de describir la regla de Introducción del condicional en deducción natural: podemos aseverar un condicional si de suponer el antecedente llegamos al consecuente a partir de los otros supuestos que tenemos. La regla de Introducción del condicional de deducción natural cristaliza el espíritu suposicional de la práctica de aseverar condicionales.

Una posible objeción es que esta regla nos permite derivar paradojas monotónicas como $A \vdash B \rightarrow A$, por lo que no termina de capturar el espíritu del Test de Ramsey. Pero esta paradoja es un caso límite que todas las teorías suposicionales validan en algún sentido. Por ejemplo, hay una versión de la paradoja de monotonía que tienen las teorías probabilísticas de Adams (1975) y Edgington (1995) y Egré *et al.* (2024): estas teorías nos dicen que la probabilidad de un condicional indicativo $Pr(A \rightarrow B)$ es igual a la probabilidad condicional de *B* dado *A* (esto se desarrollará con más detalle en la sección 5.2; a esto se lo suele llamar la Tesis de Stalnaker). Para Adams y Edgington esta función de probabilidad es una especie de función parcial que solo permite evaluar condicionales que tienen un único condicional como conectiva principal. En tanto que la función de probabilidad de Egré *et al.* funciona igual que la de Adams y Edgington para las fórmulas que solo contienen un condicional como conectiva principal y arroja alguna cosa distinta cuando las fórmulas tienen condicionales anidados a la derecha, o contienen algún condicional que no es la conectiva principal. Para estas propuestas si algo es verdadero, entonces su probabilidad condicional va a ser 1 independientemente de cuál sea el antecedente sobre el que condicionalicemos. Incluso en las teorías de *Restrictor View*, como las semánticas dinámicas de Gillies (2004) también arrastran otra versión de esta paradoja por cómo está formulada la cláusula de verdad del condicioanl. Y de forma más moderada también para la semántica de Stalnaker (1976), que en la medida que el consecuente sea verdadero en todos los mundos más cercanos, el condicional será verdadero independientemente del antecedente. Es decir, si bien no lo valida, hace verdaderos a todas las instancias de esta paradoja que tienen como consecuente un condicional verdadero. Por lo que la paradoja de monotonía no creo que sea un problema demoledor para una teoría, en la medida que casi todas las teorías conviven con esta paradoja de alguna forma u otra.

Algo interesante para tener en cuenta es que gran parte de las semánticas de los condicionales proponen reglas muy parecidas a la regla de Introducción. Por ejemplo, Gillies (2004) propone una semántica dinámica en la que tenemos un estado epistémico K, un contexto C que se actualiza a medida que aseveramos oraciones y el significado de las conectivas está dado por el potencial que tienen de modificar el contexto. En este marco, nos propone la siguiente definición para un condicional indicativo y epistémico: si actualizamos hipotéticamente nuestro contexto con el antecedente y no llegamos a un absurdo, entonces podemos aseverar el condicional. O dicho de otro modo, si suponer el antecedente no modifica nuestro estado epistémico entonces podemos aseverar el condicional. De esto se sigue que el consecuente debía estar incluido ya en nuestro arsenal de conocimientos previos (suponiendo que somos agentes epistémicamente ideales tal y como supone Gillies). Noten que esta cláusula captura algo muy parecido a lo que nos permite hacer la regla de Introducción. Si dados

15 «If two people are arguing "If p will q" and both are in doubt as to p, they are adding p hypothetically to their stock of knowledge and arguing on that basis about q (...) we can say that they are fixing their degrees of belief in q given p.»

https://doi.org/10.1387/theoria.26370

nuestros supuestos previos podemos inferir el consecuente al suponer el antecedente, entonces o bien el consecuente estaba incluido en el producto de cerrar bajo consecuencia el conjunto que contiene a nuestros supuestos previos junto con el antecedente, o bien el antecedente era incompatible con nuestros supuestos previos.

También es interesante notar que la regla de Introducción es prácticamente análoga a un lado de la cláusula que propone Gärdenfors en (1984), que dice que $A \rightarrow B$ si y solo si B pertenece a K^*_A, donde K^*_A es el producto de revisar K por A, donde la revisión devuelve algo distinto de K si y solo si A es una contradicción. Esto es prácticamente lo mismo que lo que proponen las semánticas dinámicas.

Ahora, la gran pregunta que surge es por qué si la regla de Introducción captura el espíritu del Test de Ramsey igual de bien que lo hacen los enfoques suposicionales, el condicional que resulta del par de reglas de Introducción y Eliminación de deducción natural no es considerado un enfoque suposicional. La respuesta es que la literatura suele tener un sesgo clásico.

La razón por la que las teorías del condicional material no entran en la categoría suposicional, no es que la regla de Introducción del condicional no capture el espíritu de Ramsey, sino que el par de reglas de Introducción y Eliminación en el contexto de la lógica clásica nos devuelve un condicional que se caracteriza por una cantidad infinita de paradojas. Estas paradojas nos alejan demasiado de una idea suposicional. Pero no es que el condicional material no nos diga nada sobre el funcionamiento suposicional o hipotético, sino que nos dice demasiado.

Pero si volvemos al argumento de Belnap sobre tonk, podemos ver que esto se debe a que estamos mirando el par de reglas y asumiendo de forma sesgada el contexto de deducción. Es decir, ese par de reglas en el aire no basta para describir el condicional material, de hecho no basta para describir ningún condicional. Y una vez que notamos que $\models$ o $\vdash$ no tienen por qué representar la transmisión de verdad como lo hace el martillo de la lógica clásica, podemos ver que este par de reglas es completamente compatible con una teoría suposicional del condicional.

Hasta aquí entonces tenemos lo siguiente: la regla de Eliminación, como ya vimos en la sección anterior, es una regla innegociable a la hora de describir el funcionamiento de los condicionales. A tal punto que podemos decir que si no sabemos usar MP, no sabemos usar un condicional. En tanto que la regla de Introducción es la que captura el funcionamiento suposicional de los condicionales. Por otro lado, y como veremos a continuación, en la medida en que la noción de consecuencia refiera a la lógica intuicionista vamos a quitarnos de encima la mitad de las paradojas del condicional material. Mientras que si la noción de consecuencia refiere a la lógica minimal vamos a quitarnos de encima todas, excepto por la paradoja de la monotonía del condicional (pero que ya vimos que todas las teorías suposicionales tienen en alguna medida). Por lo que podríamos decir que, en el marco de un sistema de deducción natural, el problema no son las reglas del condicional sino las reglas para la negación. Es decir, las reglas que convierten al condicional minimal e intuicionista en el condicional material.

En lo que sigue opto por defender el condicional intuicionista y no el minimal por dos razones: primero, porque en algún sentido estos dos condicionales son el mismo condicional, en tanto que para el fragmento del lenguaje que no contiene negaciones nos permiten probar exactamente las mismas cosas. Segundo, pero más importante, porque la lógica minimal no nos da información sobre cómo inferir cosas con negaciones. Es una lógica que nos dice qué hacer en casos de contradicciones, pero nada más. Y esa es la razón principal por la que nos quitamos de encima todas las paradojas de la implicación material: porque todas involucran negaciones y la lógica minimal nos da normas muy mínimas sobre cómo operar con fórmulas negadas. En ese sentido, no nos dice nada sobre qué hacer con un condicional negado. Con esto no quiero argumentar que sea una mala lógica de los condicionales, sencillamente creo que si queremos buscar normas sobre cómo inferir en un lenguaje que contiene tanto un condicional como una negación, la lógica intuicionista va a ser más útil.

Por otro lado, y como veremos en la sección siguiente, las semánticas intuicionistas nos proporcionan interpretaciones epistémicas que permiten una buena interpretación del condicional indicativo. Es interesante ver que en la medida en la que una crea que las reglas de deducción natural son aquellas que determinan el significado de las conectivas (a diferencia de quienes defienden que las reglas de secuentes son mejores o igual de buenas para este trabajo (Negri, 2002)), entonces es necesario sostener una postura molecularista: dado que el significado del condicional va a estar atado a las reglas que caractericen la negación. Si una mantiene una postura atomista, es necesario comprometerse con que el significado de las reglas puede darse en secuentes, en tanto que estos cálculos permiten que el mismo par de reglas determine dos condicionales distintos restringiendo la cantidad de conclusiones. La discusión por qué tipo de reglas son mejores para capturar el significado de una conectiva es muy extensa y por cuestiones de espacio no se desarrollará aquí. Sin embargo, no está demás mencionar esto: si queremos reglas de deducción natural como las que se presentan al principio del artículo es necesario comprometerse con una posición molecularista, pero es posible defender la tesis que aquí presento sobre los condicionales desde una postura atomista del significado en la medida que el cálculo sea otro.

Recapitulo el argumento central hasta aquí: a la hora de capturar el espíritu del Test de Ramsey no hay razones para rechazar la regla de Introducción excepto por la creencia de que el par de reglas (Introducción y Eliminación) nos conduce al condicional material. Pero ya hemos visto que no. A su vez, tiene a su favor que es la regla más sencilla capaz de capturar lo que estas teorías suposicionales intentan capturar.

Es cierto que alguien podría argumentar, y con verdad, que la regla de Introducción permite aceptar condicionales sin considerar el antecedente, siempre que tengamos razones para aceptar el consecuente. En tanto que considerar el antecedente es una parte sustantiva del Test de Ramsey. Sin embargo, creo que esto depende de la lectura que hagamos del Test de Ramsey. Una forma razonable de leerlo, tal y como hacen las semánticas dinámicas, es preguntarse si de aceptar el antecedente, el consecuente sigue siendo parte de nuestro conjunto de creencias. Esta pregunta no nos obliga a establecer una relación relevante entre antecedente y consecuente. Y de hecho, en general las clausulas de verdad de los condicionales que sí se suelen considerar ramsianos no son relevantes en este sentido. Por otro lado, la segunda parte de la cita de Ramsey puede tomarse como la vara de relevancia: cómo ajustamos nuestro grado de creencia en el consecuente de suponer que el antecedente es verdadero, determinará que estemos más o menos dispuestas a aceptar el condicional. Para eso necesitamos una teoría de la probabilidad que nos diga cómo hacer esa cuenta. El condicional material es incapaz de hacer eso, dado que la probabilidad de $A \supset B$ es igual a la probabilidad de $\neg A \vee B$. En tanto que la probabilidad de B dado A no puede coincidir con la probabilidad de un condicional indicativo que satisfaga ciertos requisitos muy básicos, como explica Lewis (1976). En lo que sigue, voy a argumentar en favor del condicional intuicionista y mostrar que a diferencia del condicional material, este condicional sí satisface este requisito, también conocido como la Tesis de Stalnaker.

Una cosa que no deberíamos dejar de nombrar es que este par de reglas es el mínimo necesario para validar el metateorema de la deducción ($\Gamma \cup \{A\} \vDash B$ si y solo si $\Gamma \vDash A \rightarrow B$). Es decir, a partir de estas reglas vamos a poder tener un condicional que internalice la noción de consecuencia con la que estamos trabajando. En ese sentido, el condicional intuicionista es un buen candidato para representar un condicional que exprese la noción de deducción que utilizamos cuando hablamos.

Por último, algo que no quiero dejar de remarcar es que si creemos que, siguiendo el consenso que hay en la literatura un condicional debe validar MP, Import-Export y ser estrictamente más fuerte que el condicional material, entonces, como muestra Fitelson en (2013), el condicional más débil capaz de satisfacer estas tres inferencias[16] es el condicional intuicionista.

5. *Una defensa del condicional intuicionista*

En esta sección quiero defender el condicional intuicionista como un buen modelo del condicional indicativo y también defenderlo de posibles críticas. En las secciones anteriores argumenté que las reglas que caracterizan al condicional intuicionista son las reglas que determinan el significado del condicional indicativo. A su vez, remarqué que hay un conjunto de inferencias que de forma *naïve* la literatura desearía que valide un condicional indicativo: MP, Import-Export y ser es-

[16] Junto con algunos supuestos muy básicos extra.

https://doi.org/10.1387/theoria.26370

trictamente más fuerte que el condicional material. Una característica fundamental del condicional intuicionista es que es el condicional más débil[17] que satisface estos tres requisitos. Es cierto, sin embargo, que hay una cantidad infinita de condicionales entre el intuicionista y el material que también lo satisfacen. No voy a hacer un argumento exhaustivo de por qué el intuicionista es mejor que una cantidad infinita de otros condicionales. En lo que sigue argumentaré, de forma modesta, que al menos hay buenas razones para tomar al condicional intuicionista como un buen modelo del condicional indicativo. Comenzaré por mostrar que invalida la mitad de las paradojas del condicional material y luego, procederé a mostrar que tiene un semántica que captura no solo el espíritu ramsiano sino también la intuición epistémica detrás de los condicionales. Y, principalmente, voy a argumentar que a diferencia de muchos otros condicionales (clásicos y no clásicos), el condicional intuicionista nos proporciona una norma pragmática-epistémica clara de cuándo aseverar un condicional, dado que la probabilidad del condicional es equivalente a su probabilidad condicional.

5.1. Las paradojas del condicional material

Comencemos mirando el siguiente ejemplo que tomo con leves modificaciones de (Gillies, 2004):

Ejemplo 1. Ha habido un asesinato en la mansión y nos quedan varios sospechosos, entre ellos dos hombres, el coronel Mostaza y el reverendo Verde, y una mujer, la Señorita Escarlata. Tenemos buenas razones pero no concluyentes para creer que si fue un hombre entonces fue el Coronel Mostaza. Mi amiga, con muchísima seguridad, me dice:

— Si lo hizo un hombre, entonces fue el Coronel Mostaza. Pero yo estoy en desacuerdo, así que le respondo:

— No es el caso que si fue uno de los dos entonces fue el Coronel Mostaza. Después de todo, el Reverendo Verde sigue siendo un buen sospechoso.

Sabemos que si este condicional se leyera materialmente, negar «Si fue uno de los dos quien lo hizo, entonces fue el Coronel Mostaza» implicaría que el Reverendo Verde es el culpable ($\neg(A \rightarrow B) \vDash_{CL} A$). Pero esa claramente no es la información que yo quiero transmitir. Lo que quiero decir es que no hay pruebas concluyentes que incriminen a ninguno de los dos. Esto podemos decirlo con el condicional intuicionista, puesto que que negar aquel condicional no implica que el Reverendo Verde sea el asesino. Este ejemplo es un caso de un condicional epistémico porque habla sobre nuestros estados de información e indicativo porque es un condicional que utilizamos en la vida cotidiana. Y creo que este es un buen caso para ver por qué el condicional intuicionista funciona en tanto modelo de los condicionales indicativos con carga epistémica.

Voy a comenzar presentando la lectura que propone Dummett del condicional intuicionista que resulta muy natural a la hora de pensar en condicionales epistémicos:

> La estipulación estándar de lo que debe contar como prueba de un enunciado condicional es que debe ser una operación de la que podamos reconocer que, aplicada a cualquier prueba del antecedente, producirá una prueba del consecuente. (Dummett, 1991, pp. 227)[18]

Noten que, naturalmente, la descripción de cuándo podemos aseverar un condicional (es decir, cuándo tenemos una prueba del condicional) coincide con la forma que tiene el intuicionismo de determinar qué argumentos son válidos. Podemos aseverar que un argumento que va de un conjunto de premisas a cierta conclusión es válido cuando podemos producir una prueba de la conclusión a partir del supuesto de que hay una prueba de las premisas. Por otro

[17] En el contexto de una lógica estructural.

[18] «The standard stipulation of what is to count as a proof of a conditional statement is that it must be an operation of which we can recognise that, applied to any proof of the antecedent, it will yield a proof of the consequent.»

lado, dado que el condicional intuicionista es estrictamente más fuerte que el condicional material, este invalida una gran cantidad de paradojas que motivaban la búsqueda de una semántica diferente. Si volvemos al ejemplo, de negar «si fue uno de los dos quien lo hizo, entonces fue el Coronel Mostaza», no se sigue que el culpable sea el Reverendo Verde. De hecho, es sencillo probar que invalida muchísimas paradojas del condicional material. Por enumerar algunas miremos los siguientes ejemplos que son inválidos en la lógica intuicionista:

1. $(A \to B) \vee (A \to \neg B)$
2. $(A \to B) \vee (\neg A \to B)$.
3. $\big(\neg A \to (\neg B \to A)\big) \wedge \neg B$. Por lo tanto, A

La lógica intuicionista invalida todas las paradojas del condicional material que involucran el uso insustituible de la regla de doble negación en su derivación: aquella misma regla que convierte al condicional intuicionista en el clásico en el marco de un cálculo de deducción natural. Por otro lado, nos permite tener una lógica mucho más fuerte e informativa que muchas otras propuestas, como es el caso de la lógica minimal que antes mencionaba, o el caso de algunas lógicas dinámicas que al no ser reflexivas ni monotónicas son inferencialmente muy débiles. En ese sentido, la lógica intuicionista tiene mayor poder normativo, puesto que valida muchísimas inferencias. Naturalmente, en esta fuerza deductiva también se cuelan infinitas paradojas del condicional que desearíamos que sean inválidas. Por ejemplo, he aquí algunas:

4. A Por lo tanto, $B \to A$
5. A Por lo tanto, $\neg A \to B$

Mi propuesta es entender este tipo de inferencias como válidas pero no aseverables —en el mejor espíritu de Jackson (1987) y de Stalnaker (1976). Esto es, si bien estas inferencias están autorizadas por la lógica, no hay ningún contexto en el que tengamos razones para utilizarlas. Una de las bondades de pensar el significado en términos de reglas es que nos permite pensar a las inferencias válidas como *tickets de inferencia*, como diría Ryle (1950). Es decir, como transformaciones que se nos permite ejecutar, pero que no estamos obligadas a hacer. De hecho, podemos guardar nuestro ticket para toda la vida y nunca usarlo.

Esto habilita que nuestras creencias no estén cerradas bajo consecuencia. Si bien es cierto que yo debería aceptar la conclusión de un argumento, si yo asevero un conjunto de premisas de las que se sigue lógicamente cierta oración y alguien me pregunta y yo me doy cuenta de aquella relación lógica. Pero fuera de aquel contexto, no tengo por qué utilizar aquella conclusión en mis razonamientos, porque estar permitido no implica que tenga razones para aseverarlo.

Una de las principales bondades de esta propuesta es que, a diferencia de muchas otras teorías del condicional indicativo, aquí podemos establecer una vara de aseveración o *aseverabilidad* que haga uso de la Tesis de Stalnaker (una de las propiedades mas deseadas en la literatura sobre condicionales indicativos). A continuación voy a explicar brevemente cómo funcionan las probabilidades de los condicionales intuicionistas y la probabilidad condicional para luego poder explicar cuándo la conclusión de una inferencia válida es aseverable en esta propuesta.

5.2. La probabilidad del condicional intuicionista

Como se puede ver en Adams (1975), De Finetti (1936), Edgington (1995) o Égré *et al.* (2024), entre muchos otros, es común pensar que las creencias parciales pueden modelarse mediante funciones de probabilidad. Y en particular estas autoras (entre muchas otras) han tratado de cuadrar la idea de que el grado de creencia en un condicional indicativo debe ser igual al de la probabilidad condicional. A esta idea se la llama Tesis de Stalnaker.

Tesis de Stalnaker: $Pr(A \to B) = Pr(B|A) = \dfrac{Pr(A \wedge B)}{Pr(A)}$, si $Pr(A) > 0$.

Sin embargo, por razones que nos llevan muy por fuera de este trabajo, (Lewis, 1976), y muchas otras autoras luego, han mostrado que si una tiene una función de probabilidad clásica y asume la Tesis de Stalnaker, entonces la función de probabilidad trivializa.

En (2003), Weatherson muestra cómo construir una función de probabilidad intuicionista. Es decir, es posible definir una función de probabilidad Pr para un lenguaje proposicional $\mathcal{L}$, que contenga una conjunción, $\wedge$, una disyunción, $\vee$, una negación, $\neg$, una constante de verdad, $\top$, (para facilitarnos la notación) y un condicional, $\longrightarrow$, tal que $Pr : For(\mathcal{L}) \mapsto [0, 1]$ y que satisfaga las siguientes propiedades:

P1 Si $\vdash_I A$, entonces $Pr(A)=1$

P2 Si $A \vdash_I$, entonces $Pr(A)=0$.

P3 Si A y B son lógicamente equivalentes, entonces $Pr(A) = Pr(B)$

P4 Si $A \vdash_I B$, entonces $Pr(A) \leq Pr(B)$.

P5 $Pr(A \vee B) = Pr(A) + Pr(B) - Pr(A \wedge B)$

Donde $\vdash_I$ se lee como la noción de consecuencia de la lógica intuicionista, $\vdash_I A$ se lee como ser un teorema de la lógica intuicionista y $A \vdash_I$ ser un anti-teorema de la lógica intuicionista, es decir, una contradicción. Quien esté familiarizada con las probabilidades podrá notar que aquí (a diferencia de las probabilidades clásicas) la negación no es composicional. Dado que $\neg A \vee A$ no es un teorema de la lógica intuicionista, es posible asignarle $Pr(A) = Pr(\neg A) = 0$. Pero puesto que $A \wedge \neg A$ es una contradicción de la lógica intuicionista, y dada la propiedad [P5], $Pr(A \wedge \neg A) = Pr(A) + Pr(\neg A) - Pr(A \vee \neg A) = 0$. Por otro lado, para quienes no estén familiarizadas con estas funciones, estos 5 requisitos suelen considerarse requisitos básicos para que una función se pueda considerar una función probabilística (a pesar de que P5 suele debilitarse para otras lógicas no clásicas).[19]

La pregunta natural que surge a esta altura es ¿cuál es la probabilidad del condicional intuicionista? La respuesta, quizás para nuestra sorpresa es la siguiente: Morgan y Mares (1995) muestran que es posible construir una relación binaria $Pr(A|\Gamma) : For(\mathcal{L})x\mathcal{P}\ (For\mathcal{L}) \mapsto [0, 1]$ donde A es una fórmula y Γ un conjunto de oraciones, que cumpla con la siguientes propiedades:

C1. $0 \leq Pr(A|\Gamma) \leq 1$

C2. Si $A \in \Gamma$, entonces $Pr(A|\Gamma)=1$.

C3. $Pr(A|\{B\} \cup \Gamma) \cdot Pr(B|\Gamma) = Pr(B|\{A\} \cup \Gamma) \cdot Pr(A|\Gamma)$

C4. $Pr(A \rightarrow B|\Gamma) = Pr(B|\{A\} \cup \Gamma)$.

Cuando Γ es vacío vamos a señalarlo como $\top$ y tendremos lo siguiente:

Tesis Intuicionista de Stalnaker: $Pr(A \rightarrow B|\top) = Pr(B|\{A\} \cup \mathrm{T}) = \dfrac{Pr(A \wedge B)}{Pr(A)}$, si $Pr(A) > 0$.

Que es una forma distinta de reescribir la Tesis original de Stalnaker.

A diferencia de cualquier condicional que extienda la lógica clásica o del condicional material mismo, este condicional satisface la Tesis de Stalnaker sin trivializar. Las razones están explicadas en detalle en el texto de Morgan y Mares.

[19] Ver (Williams, 2016) para más detalles al respecto.

Sin embargo, dos formas sencillas de entender por qué sucede son las siguientes. Primero que nada, es relevante saber que para derivar el resultado de trivialidad de Lewis necesitamos hacer un uso esencial de Tercero Excluido, que en la lógica intuicionista no es un teorema. Otra forma más específica de ver por qué se evita este resultado de trivialidad es pensando cómo se calcula la probabilidad del condicional material en contraposición con la probabilidad del condicional intuicionista. Quien no esté familiarizada con estos temas puede saltarse lo que resta del párrafo, pues es bastante técnico y por razones de espacio no desarrollaré el detalle. Una forma de entender lo que sucede es que si extendemos la lógica clásica con un condicional indicativo que valide MP, Import-Export y que sea más fuerte que el material, entonces este condicional va a colapsar con el condicional material (Fitelson, 2013; Gibbard, 1980). Pero las probabilidades del condicional material siempre son iguales o más altas que la probabilidad condicional — dado que el condicional material en el fondo es una disyunción que nos dice que o bien el antecedente es falso o bien el consecuente es verdadero. Por lo que cuando suponemos que la probabilidad del condicional indicativo (que asumimos que valida MP, Import-Export y que es más fuerte que el material) es la probabilidad condicional (que es una división entre la conjunción de antecedente y consecuente, y el antecedente), llegamos a una contradicción. En cambio en la lógica intuicionista, si bien el condicional indicativo al agregarlo a la lógica colapsa con el condicional intuicionista, este no es interdefinible con la disyunción. De hecho no es interdefinible con ninguna otra conectiva del lenguaje, por lo que tenemos la libertad de definirlo a partir de la famosa *ratio formula*, es decir, la Tesis de Stalnaker.

Naturalmente, va a tener algunos comportamientos un tanto peculiares respecto de lo que estamos acostumbradas a ver en una función de probabilidad condicional, pero que son de esperar dado que aquella relación está representando al condicional intuicionista. En particular va a ser monotónica:

M1. Para toda función Pr que satifaga C1-C4, $Pr(A|\Gamma) \leq Pr(A|\Gamma \cup \{B\})$.

Sin embargo, como hemos visto anteriormente, las paradojas de la monotonía son las más recurrentes en las mejores teorías de los condicionales indicativos.

Volviendo a la pregunta de cuándo una inferencia válida no es aseverable, esta función puede cumplir un rol pragmático/espitémico a la hora de moldear la forma en la que entendemos las reglas del condicional y a la hora de decidir cuándo aseverar un condicional. En tanto que el significado va a estar dado por las reglas del condicional intuicionista, nuestro entendimiento va a completarse a través de la Tesis Intuicionista de Stalnaker. Esto es, siguiendo el espíritu de Jackson (1979; 1987), podemos decir que cuando aseveramos un condicional es porque creemos que el grado de probabilidad del consecuente dado el antecedente es alto. Esto a su vez nos da un buen criterio de relevancia, tal y como propone Douven (2015)[20]. Dado que nuestro condicional respeta la Tesis Intuicionista de Stalnaker, podemos argumentar que un antecedente de un condicional, digamos A, es relevante para el consecuente, en este caso B, cuando la probabilidad condicional de $Pr(B|\Gamma \cup \{A\}) > Pr(B|\Gamma \cup \{\neg A\})$.

Esto es un beneficio fuerte del condicional intuicionista por sobre muchísimos otros condicionales de la literatura. En particular, Jackson (1979) defiende esta tesis pero respecto del condicional material. Nos dice que podemos aseverar un condicional en la medida que la inferencia que nos devuelve aquel condicional sea clásicamente válida y que la probabilidad condicional del consecuente dado el antecedente sea alta. Sin embargo, esta propuesta tiene un problema grave y es que no hay forma de explicar la relación entre la probabilidad condicional y el condicional de la lógica, dado que son dos condicionales distintos. En tanto que para la lógica intuicionista podemos tomar de forma literal la propuesta de Jackson, puesto que solo tenemos un condicional (el intuicionista) cuya probabilidad coincide con la probabilidad condicional.

[20] Douven está hablando de la función de probabilidad condicional clásica. Jackson también.

Por otro lado, una podría preguntarse por los condicionales relevantes. Esta literatura es muy basta e interesante, y si bien suelen invalidar las paradojas de la monotonía, no pueden dar funciones de probabilidad que respeten la Tesis de Stalnaker. A su vez, la literatura se debate mucho en relación a cuáles son los requisitos de relevancia: inclusión de contenido o preservación de tópicos, por nombrar algunos. Aquí tenemos un criterio de relevancia epistémico: un antecedente es relevante cuando la probabilidad del consecuente aumenta si se da el antecedente.

Es importante destacar que este condicional permite tener muchas de las desiderata deseadas en la literatura. Las extensiones de la lógica clásica, como el condicional de Stalnaker y Lewis, no pueden validar MP, Import-Export y ser más fuertes que el condicional material a la vez, así como tampoco pueden satisfacer la Tesis de Stalnaker. Las semánticas dinámicas tampoco. Las propuestas probabilísticas como la de Adams no pueden anidar condicionales (y por ende no pueden satisfacer Import-Export). En tanto que las propuestas trivaluadas satisfacen la Tesis de Stalnaker pero tienen un condicional más débil que el condicional material.[21] En cambio, el condicional intuicionista nos permite validar las tres inferencias que suelen tener consenso en la literatura, su significado está determinado por dos reglas razonables y nos permite validar la Tesis de Stalnaker.

Para cerrar voy a hablar brevemente sobre la interpretación epistémica de la lógica intuicionista.

5.3. Una interpretación epistémica para un condicional epistémico

En esta sección no quiero presentar el detalle de los modelos de Kripke, porque creo que es posible entender la propuesta y la interpretación BHK sin necesidad de introducir todos los tecnicismos —quién esté interesada en conocer la propuesta en detalle, recomiendo ir al texto original de Kripke (1965). La idea principal de los modelos de Kripke es pensar la verdad de las oraciones relativizada a los estados epistémicos de una agente ideal. Para eso se presenta un sistema modal que parte de un mundo actual en el que la agente conoce una serie de cosas (quizás nada) y a medida que avanza el tiempo va obteniendo más y más información certera que va determinando la verdad o la falsedad de las fórmulas del lenguaje. Esta agente ideal va a mantener indeterminado el valor de todas y cada una de las fórmulas hasta no tener una prueba concluyente de ella o de su negación, y a medida que las consiga, dado que son pruebas concluyentes, el valor de verdad de estas oraciones quedará determinado para siempre. A medida que el tiempo avanza, la agente tendrá que elegir sobre qué cosas investigar. Ese conjunto de decisiones lo podemos modelar como una estructura de un árbol que se ramifica en función de las posibles decisiones.

Los modelos de Kripke se presentan en términos de un conjunto de mundos, un mundo designado (el mundo actual desde el que partimos), una relación de accesibilidad reflexiva y transitiva, sobre la que se exige la *condición hereditaria*, —es decir, una restricción que nos dice que si en un mundo una oración es verdadera, en todos los mundos a los que accede, aquella oración seguirá siendo verdadera—, y una función de valuación. Las fórmulas atómicas pueden ser verdaderas o falsas en un mundo, pero dado que no tenemos una negación booleana, vamos a poder tener mundos en los que las fórmulas sean tanto ellas falsas como también su negación. La idea recordemos que es que una fórmula sólo va a ser verdadera en la medida que tengamos una prueba concluyente, por lo que si no tengo una prueba concluyente de A ni de $\neg A$, en el mundo actual ninguna será verdadera. Pero dado que estamos en un contexto bivaluado, esto va a forzar a que ambas sean falsas en el mundo actual. Vamos a escribir $w \vDash^K A$ para decir que un mundo w hace verdadera a A y $w \nvDash^K A$ para decir que w la hace falsa. Suponiendo que tenemos un lenguaje que solo tiene negación y condicionales, la idea es que un condicional $A \to B$ va a ser verdadero en un mundo w si y solo si en todos los mundos a lo que accede w, o bien A es falsa o bien B es verdadera. En caso contrario $A \to B$ será falsa en ese mundo. En tanto que una fórmula $\neg A$ va a ser verdadera en un mundo w si y solo si en todos los mundos a los que accede w, A es falsa y será falsa en caso contrario. Una fórmula va a ser válida si y solo si es verdadera en todo

[21] Ver (Égré *et al.*, 2023).

mundo w y en todo modelo, y lo escribimos como $\vDash^K A$. Un argumento va a ser válido si y solo si en todo mundo w de todo modelo o bien alguna premisa es falsa o bien la conclusión es verdadera. Y lo escribimos como $\Gamma \vDash^K A$. Es decir, en la medida que no tenga contraejemplos, donde un contraejemplo es un mundo w en un modelo en el que todas las premisas son verdaderas y la conclusión falsa.

En términos generales, la idea es que vamos a poder inferir solo aquellas cosas que si sabemos que las premisas son verdaderas, entonces sabemos que la conclusión también lo es. Pero naturalmente, a diferencia de la lógica clásica donde no hay lugar para modelar lo que nosotras sabemos, aquí hay muchas cosas que no vamos a saber. Estos modelos van a invalidar muchas paradojas del condicional material (dado que son correctos y completos respecto de las reglas que presenté anteriormente). Tomemos el ejemplo del asesino en la mansión. Si estamos en un mundo w tal que $w \nvDash^K B$, $w \nvDash^K \neg B$ y $w \nvDash^K A$ y supongamos que w accede a un w' en el que $w' \vDash^K B$ y $w' \nvDash^K A$. En w, $w \nvDash^K B \to A$, dado que w solo accede a w y a w' y en $w' \nvDash^K B \to A$. De lo que se sigue que $w \vDash^K \neg(B \to A)$, sin embargo como bien dijimos, $w \nvDash^K B$.

Por otro lado, aquí se ponen de manifiesto de forma más clara las razones para rechazar instancias de tercero excluido del condicional. Es decir, una agente que desconoce si $A \to B$ o si $\neg(A \to B)$ no está en la obligación de comprometerse con la aseveración de que $(A \to B) \vee \neg(A \to B)$, así como tampoco deberá aseverar instancias del llamado *tercero excluido condicional*, $(A \to B) \vee (A \to \neg B)$, en la medida que su estado epistémico sea tal que no sepa si se dará B o $\neg B$ en caso de que se dé A.

Esta es una lógica de la preservación de evidencia concluyente en la que las oraciones que aseveramos son oraciones que sabemos que son verdaderas. Si bien es verdad que, como argumentan Adams (1975) o Edgington (1995), muchos condicionales indicativos se utilizan para expresar incerteza sobre el mundo, las lógicas que permiten hablar de la incerteza suelen ser extremadamente débiles y tener fuertes restricciones sintácticas. La lógica intuicionista no nos va a permitir aseverar oraciones para las que no tenemos prueba concluyente. Sin embargo, al ser una lógica epistémica nos permite aseverar condicionales tales que si suponemos que tenemos una prueba del antecedente, entonces tenemos certeza de que podemos construir una prueba del consecuente. Nos permite aseverar condicionales que expresan información sobre nuestro estado de conocimiento, mientras que preserva la validez de Modus Ponens, nos permite traducir condicionales anidados vía Import-Export y es un condicional estrictamente más fuerte que el material.

A su vez, dado que valida la Tesis Intuicionista de Stalnaker, podemos dar una pauta de aseveración en tanto probabilidad condicional alta. De esta forma, tenemos un condicional con una interpretación espitémica dada por una agente ideal, al igual que la mayoría de las semánticas de los condicionales, pero que aparte nos permite respetar las intuiciones *naïve* respecto de qué inferencias son válidas. Y también se inserta en el contexto de una lógica fuerte con poder normativo para que rija las prácticas deductivas de las hablantes.

6. Conclusiones

Como argumenté a lo largo de estas páginas, cuando pensamos los modelos del condicional en términos de sus condiciones de verdad, la discusión se empantana, puesto que no hay condiciones de verdad que satisfagan las intuiciones respecto de todos los condicionales del lenguaje. Sin embargo, todas acordamos en que es necesario dar una lógica para este término puesto que su rol es ubicuo en nuestros razonamientos cotidianos y en nuestra práctica científica y filosófica. Por lo que la propuesta es que pensemos su significado en término de sus condiciones de uso. En ese contexto, hay consenso respecto de cuáles serían las inferencias que idealmente querríamos que valide un condicional: en particular queremos que valide MP, Import-Export y que sea estrictamente más fuerte que el condicional material. Argumenté que MP es la inferencia más sustantiva a la hora de pensar el significado de un condicional y que quien no puede instanciar la regla de Eliminación no entiende lo que es un condicional. También argumenté que la regla de Introducción era la mejor forma y la más sencilla de capturar el Test de Ramsey. Y que en la medida en que la negación no se comportase clásicamente, era posible tener un condicional que validase MP e Import-Export y que fuese estrictamente más fuerte que el condicional material. Cabe destacar, que estas tres inferencias implican tanto la regla de Introducción como la de Eliminación, por lo que, si tenemos las tres reglas en una lógica estructural, al menos tendremos un condicional tan fuerte como el intuicionista. También mostré que el condicional intuicionista puede satisfacer una versión de la Tesis de Stalnaker, que en el marco de esta teoría vendrá a cumplir un rol epistémico que complemente el significado otorgado por el par de reglas.

Para cerrar creo que es importante destacar la que creo que es la crítica más fuerte que se le puede hacer a este enfoque. Alguien podría decir que aceptar que el condicional que utilizamos en nuestros razonamientos deductivos diarios es el condicional intuicionista nos obliga a comprometernos con la lógica intuicionista como la lógica de nuestros razonamientos diarios. Creo personalmente que esto es una de las fortalezas de esta propuesta. Que la lógica intuicionista es la lógica detrás del lenguaje natural ha sido defendido con firmeza por Dummett en (1991) y es una tesis que excede los límites de este trabajo. Sin embargo, creo que si bien involucra un cambio de paradigma sustantivo en la forma en la que trabajamos en filosofía y principalmente en matemática, es un trabajo que se ha desarrollado por años por una gran comunidad de filósofas y matemáticas. Por otro lado, es importante destacar que para cualquier propuesta de semántica que se presente, o bien va a involucrar una lógica divergente de la clásica (como en el caso de las semánticas dinámicas o las semánticas trivaluadas) o bien va a ser una extensión de la lógica clásica que pierda MP o Import-Export —como son el caso de la lógica propuesta por McGee (1985) o Stalnaker (1968). Y si bien sostener esta tesis tiene una implicancia muy fuerte, puesto que las prácticas matemáticas y científicas se basan en la lógica clásica, en la cotidianeidad es difícil sostener que usamos algo parecido a la lógica clásica. Por eso creo que la lógica intuicionista es un buen punto de partida para repensar la lógica detrás de nuestras prácticas lingüísticas, y en particular de nuestros condicionales.

Agradecimientos

Quiero agradecer a mis compañeros y compañeras del Buenos Aires Logic Group que me acompañaron en el proceso de producción de este trabajo. Quiero agradecer particularmente a Damián Szmuc que trabajó codo a codo conmigo en el proceso de escritura de la tesis doctoral, de la cual formó parte este artículo. También le agradezco a Federico Pailos, mi director de doctorado, y a Lucas Rosenblatt, mi co-director, a Eleonara Cresto, Natalia Buacar y Diego Tajer por la lectura atenta y el sus comentarios pertinentes. Y a Paula Teijeiro y a Camillo Fiore que me acompañan siempre en el proceso de hacer filosofía y lógica. Quiero agradecerle también a los referis que me leyeron atentamente y ayudaron a mejorar este trabajo. Quiero agradecerle a la UBA y al CONICET por el financiamiento y al proyecto PLEXUS («Philosophical, Logical and Empirical Routes to Substructurality») [grant agreement no. 101086295], una Marie Skłodowska-Curie action financiada por la Unión Europea bajo el Horizon Europe Research and Innovation Programme.

REFERENCES

Adams, E. W. (1975). *The logic of conditionals: An application of probability to deductive logic*. Springer Science & Business Media. https://doi.org/10.1080/00201746508601430

Belnap, N. D. (1962). Tonk, plonk and plink. *Analysis*, *22*(6), 130–134. https://doi.org/10.1093/analys/22.6.130

Belnap, N. D. (1970). Conditional assertion and restricted quantification. *Noûs*, 1-12. https://doi.org/10.2307/2214285

Birman, R. (2015). *What the tortoise said to Kripke: The adoption problem and the epistemology of logic*. Tesis Doctoral, City University of New York. https ://academicworks.cuny.edu/gc_etds/603

Brandom, R. (1994). *Making it explicit: Reasoning, representing, and discursive commitment*. Harvard University Press. https://doi.org/10.2307/2956391

Brandom, R. (2009). *Articulating reasons: An introduction to inferentialism*. Harvard University Press. https://doi.org/10.1093/mind/110.439.721

Brown, D. (1954). What the tortoise taught us. *Mind*, *63*(250), 170-179. https:// doi. org/10.2307/2268398

Buacar, N. (2015). *La justificación de la deducción*. Universidad de Buenos Aires.

Carroll, L. (1895). What the tortoise said to Achilles. *Mind*, *4*(14), 278-280. https ://doi.org/10.1093/mind/iv.14.278

De Finetti, B. (1936). La logique de la probabilité. *Actes du congrès international de philosophie scientifique*, *4*, 1-9. https://doi.org/10.1007/BF00996317

Dicher, B. (2016). Weak disharmony: Some lessons for proof-theoretic semantics. *Review of Symbolic Logic*, (3), 1-20. https://doi.org/10.1017/s1755020316000162

Dicher, B. (2020). Ask not what bilateralist intuitionists can do for cut, but what cut can do for bilateralist intuitionism. *Analysis*, *80*(1), 30-40. https://doi.org/10.1093/analys/anz023

Douven, I. (2015). *The epistemology of indicative conditionals: Formal and empirical approaches*. Cambridge University Press. https://doi.org/10.1017/CBO9781316275962

Douven, I., Elqayam, S., & Krzyżanowska, K. (2023). Inferentialism: A manifesto. *Conditionals: Logic, linguistics and psychology* (pp. 175-221). Springer. https://doi.org/10.1007/978-3-031-05682-6_7

Dummett, M. (1991). *The logical basis of metaphysics*. Harvard University Press.

Edgington, D. (1995). On conditionals. *Mind*, *104*(414), 235–329. https://doi.org/10.1093/mind/104.414.235

Edgington, D. (2020). Indicative Conditionals. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2020). Metaphysics Research Lab, Stanford University.

Égré, P., Rossi, L., & Sprenger, J. (2021). De Finettian logics of indicative conditionals, part I: Trivalent semantics and validity. *Journal of Philosophical Logic*, *50*(2), 187-213. https://doi.org/10.1007/s10992-020-09549-6

Égré, P., Rossi, L., & Sprenger, J. (2023). Gibbardian collapse and trivalent conditionals. *Conditionals: Logic, linguistics and psychology* (pp. 37-71). Springer. https://doi.org/10.1007/978-3-031-05682-6_3

Égré, P., Rossi, L., & Sprenger, J. (2024). Certain and uncertain inference with trivalent conditionals. *Australasian Journal of Philosophy*. https://doi.org/10.1080/00048402.2025.2475882

Finn, S. (2019). The adoption problem and anti-exceptionalism about logic. *Australian Journal of Logic*, 16 (7):231. https://doi.org/10.26686/ajl.v16i7.5916

Fiore, C. G. (2022). What the adoption problem does not show. *Análisis Filosófico*, *42*(1), 79-103. https://doi.org/10.36446/af.2022.402

Fitelson, B. (2013). Gibbard's collapse theorem for the indicative conditional: An axiomatic approach. *Automated reasoning and mathematics*, *7788*. https://doi.org/10.1007/978-3-642-36675-8_10

Fries, C. C. (1952). *The structure of English. an introduction to the construction of English sentences*. Harcourt. https://books.google.com.ar/books?id=XwkFAQAAIAAJ

Gärdenfors, P. (1984). The dynamics of belief as a basis for logic. *The British Journal for the Philosophy of Science*, *35*(1), 1-10. https://doi.org/10.2307/2274226

Gibbard, A. (1980). Two recent theories of conditionals. In W. L. Harper, R. Stalnaker, & G. Pearce (Eds.), *Ifs* (pp. 211-247). Springer. https://doi.org/10.1007/978-94-009-9117-0_10

Gillies, A. S. (2004). Epistemic conditionals and conditional epistemics. *Noûs*, *38*(4), 585-616. https://doi.org/10.1111/j.0029-4624.2004.00485.x

Grice, P. (1989). *Studies in the way of words*. Harvard University Press.

Hallebeek, J. (1986). Las palabras funcionales del español. *Boletın de la AEPE*, (34-35), 205–216.

Harper, W. L., Stalnaker, R., & Pearce, G. (Eds.). (1981). *Ifs*. D. Reidel.

Iacona, A. (2023). Valid arguments as true conditionals. *Mind*, *132*(526), 428-451. https://doi.org/10.1093/mind/fzac026

Jackson, F. (1979). On assertion and indicative conditionals. *The Philosophical Review*, *88*(4), 565-589. https://doi.org/10.2307/2184845

Jackson, F. (1987). *Conditionals*. Blackwell, Oxford. https://doi.org/10.2307/2219834

Kratzer, A. (1979). Conditional necessity and possibility. *Semantics from different points of view* (pp. 117-147). Springer. https://doi.org/10.1007/978-3-642-67458-7_9

Kripke, S. A. (1965). Semantical analysis of intuitionistic logic I. *Studies in logic and the foundations of mathematics* (pp. 92-130). Elsevier. https://doi.org/10.2307/2270547

Kripke, S. A. (2023). The question of logic. *Mind*, *133*(529), 1-36. https://doi.org/10.1093/mind/fzad008

Lewis, D. (1976). Probabilities of conditionals and conditional probabilities. In W. L. Harper, R. Stalnaker, & G. Pearce (Eds.), *Ifs* (pp. 129-147). Springer. https://doi.org/10.2307/2184279

McGee, V. (1985). A counterexample to modus ponens. *The Journal of Philosophy*, *82*(9), 462-471. https://doi.org/10.2307/2026276

Morgan, C. G., & Mares, E. D. (1995). Conditionals, probability, and non-triviality. *Journal of Philosophical Logic*, *24*, 455-467. https://doi.org/10.1007/bf01052599

Negri, S. (2002). Varieties of linear calculi. *Journal of Philosophical Logic*, *31*(6), 569-590. https://doi.org/10.1023/a:1021264102972

Oberauer, K., & Wilhelm, O. (2003). The meaning(s) of conditionals: Conditional probabilities, mental models, and personal utilities. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(4), 680. https://doi.org/10.1037/0278-7393.29.4.680

Prawitz, D. (2006). *Natural deduction: A proof-theoretical study*. Courier Dover Publications. https://doi.org/10.2307/2271676

Prior, A. N. (1960). The runabout inference-ticket. *Analysis*, *21*(2), 38-39. https://doi.org/10.1093/analys/21.2.38

Prior, A. N. (1964). Conjunction and contonktion revisited. *Analysis*, *24*(6), 191-195. https://doi.org/10.1093/analys/24.6.191

Ramsey, F. P. (2016). Truth and probability. *Readings in formal epistemology*. Springer. https://doi.org/10.1007/978-3-319-20451-2_3

Read, S. (2010). General-elimination harmony and the meaning of the logical constants.

Journal of philosophical logic, *39*(5), 557-576. https://doi.org/10.1007/s10992-010-9133-7

Rips, L. J. (1994). *The psychology of proof: Deductive reasoning in human thinking*. MIT Press.

Ryle, G. (1950). The concept of mind. *British Journal for the Philosophy of Science*, *1*(4), 328-332.

Sellars, W. (1953). Inference and meaning. *Mind*, *62*(247), 313–338. https://doi.org/10.1093/mind/lxii.247.313

Stalnaker, R. C. (1968). A theory of conditionals. In W. L. Harper, R. Stalnaker, & G. Pearce (Eds.), *Ifs* (pp. 41-55). Springer. https://doi.org/10.1007/978-94-009-9117-0_2

Stalnaker, R. C. (1976). Indicative conditionals. In W. L. Harper, R. Stalnaker, & G. Pearce (Eds.), *Ifs* (pp. 193-210). Springer. https://doi.org/10.1007/bf02379021

Steinberger, F. (2011). What harmony could and could not be. *Australasian Journal of Philosophy*, *89*(4), 617-639. https://doi.org/10.1080/00048402.2010.528781

Stevenson, J. T. (2018). Roundabout the runabout inference-ticket. *Thinking about logic* (pp. 43-49). Routledge. https://doi.org/10.1093/analys/21.6.124

Weatherson, B. (2003). From classical to intuitionistic probability. *Notre Dame Journal of Formal Logic*, *44*(2), 111-123. https://doi.org/10.1305/ndjfl/1082637807

Williams, R. (2016). Probability and nonclassical logic. In A. Hajek & C. Hitchcock (Eds.), *The Oxford handbook of probability and philosophy*. Oxford University Press.

Williamson, T. (2020). *Suppose and tell: The semantics and heuristics of conditionals*. Oxford University Press. https://doi.org/10.1080/01445340.2021.1958648

Yalcin, S. (2007). Epistemic modals. *Mind*, *116*(464), 983–1026. https://doi.org/10.1093/mind/fzm983

Mariela Rubin es doctora en filosofía por la Universidad de Buenos Aires. Actualmente es becaria pos-doctoral en CONICET. Se dedica principalmente a la lógica y se especializa en condicionales. También es docente de lógica en la Facultad de Filosofía y Letras, UBA y forma parte del comité editorial de *Análisis Filosófico*. Sus principales intereses son los condicionales indicativos, la teoría de la prueba, la pregunta por el significado de las constantes lógicas, el inferencialismo y la normativdad de la lógica, así como también la relación entre creencias y racionalidad.

Dirección Postal: Bulnes 642, CP: 1176, Buenos Aires, Argentina – marubin@gmail.com – ORCID address: 0000-0002-9392-3618

Theoria, 2025, 40(2), 130-153
https://doi.org/10.1387/theoria.26981

THEORIA
ISSN 0495-4548 – eISSN 2171-679X

Agencia, relativismo y conocimiento extendido

(Agency, relativism and extended knowledge)

Dubian Cañas*
Universidad Nacional Autónoma de México
https://orcid.org/0000-0002-2134-1045

Palabras clave

Agencia epistémica
Escepticismo
Externismo activo
Fiabilismo de virtudes
Metaepistemología

RESUMEN: En este artículo examino si la teoría de conocimiento extendido (TCE) puede dar una explicación completamente satisfactoria de la agencia epistémica combinando el externismo activo y el fiabilismo de virtudes. TCE sería una teoría completamente satisfactoria si explica cómo los agentes son responsables epistémicamente porque tienen y manifiestan una disposición cognitiva a responder apropiadamente ante razones derrotantes de su estatus epistémico. Mi tesis negativa es que TCE no logra este cometido porque sus principios explicativos de primer orden conducen al escepticismo. En consecuencia, la teoría debería abandonarse. Mi tesis positiva, sin embargo, es que TCE puede dar una explicación completamente satisfactoria de la agencia si adopta el relativismo epistémico como principio de segundo orden. Aunque esta maniobra implica revisar la metaepistemología detrás de TCE, permite conservar la combinación de externismo activo y fiabilismo de virtudes constitutiva de esta teoría epistemológica.

Keywords

Active externalism
Epistemic agency
Metaepistemology
Scepticism
Virtue reliabilism

ABSTRACT: I examine whether the theory of extended knowledge (TEK) can provide a fully satisfactory account of epistemic agency by combining active externalism and virtue reliabilism. TEK would be a fully satisfactory theory if it could explain how agents are epistemically responsible by both having and manifesting a cognitive disposition to appropriately respond to defeating reasons vis-à-vis their epistemic status. My negative thesis is that TEK fails to achieve this goal because its first-order explanatory principles lead to scepticism. As a result, the theory should be abandoned. My positive thesis, however, is that TEK can provide a fully satisfactory explanation of agency if epistemic relativism is adopted as a second-order principle. Although this manoeuvre involves adjusting the metaepistemology underlying TEK, it makes it possible to retain the combination of active externalism and virtue reliabilism that constitutes this epistemological theory.

1. Introducción

La teoría de conocimiento extendido (TCE) es la combinación de externismo activo y fiabilismo de virtudes. Uno de sus objetivos es explicar la agencia epistémica en la adquisición de conocimiento mediante procesos de formación de creencias, los cuales son realizados por sistemas de cognición extendida y distribuida. TCE sería una teoría completamente satisfactoria de la agencia epistémica si explica cómo los agentes son responsables epistémicamente porque tienen y manifiestan una disposición cognitiva a responder apropiadamente ante razones derrotantes de su estatus epistémico. En este artículo evalúo a TCE como teoría de la agencia epistémica y defiendo las siguientes dos tesis.

* ***Correspondence to:*** Dubian Cañas. Universidad Nacional Autónoma de México, Edificio E - Unidad de Posgrado, Circuito de Posgrados, Ciudad Universitaria, 04510, Ciudad de México, México – dubiandrescanas@gmail.com – https://orcid.org/0000-0002-2134-1045

How to cite: Cañas, Dubian (2025). «Agencia, relativismo y conocimiento extendido»; *Theoria. An International Journal for Theory, History and Foundations of Science,* 40(2), 130-153. (https://doi.org/10.1387/theoria.26981).

Received: 02 October, 2024; Final version: 01 July, 2025.
ISSN 0495-4548 - eISSN 2171-679X / © 2025 UPV/EHU Press

Mi tesis negativa es que TCE no puede dar una explicación completamente satisfactoria de la agencia epistémica porque sus principios explicativos llevan conjuntamente al escepticismo. Este hecho confronta a sus defensores con un dilema: TCE debe abandonarse si renuncian a sus principios de primer orden o TCE puede conservarse si renuncian a sus principios de segundo orden. Mi sugerencia es optar por la segunda opción del dilema porque lo más prudente es ajustar la metaepistemología de una teoría epistemológica antes de rechazar dicha teoría. Así, mi tesis positiva es que, si TCE acepta el relativismo epistémico, entonces puede dar una explicación completamente satisfactoria de la agencia epistémica sin caer en el escepticismo. Esto implica renunciar al absolutismo epistémico como principio de segundo orden, pero permite preservar la combinación de externismo activo y fiabilismo de virtudes.

La sección 2 presenta el externismo activo y el fiabilismo de virtudes como principios de primer orden de TCE. La sección 3 explica cómo TCE daría cuenta de la agencia epistémica a partir de estos principios, especificando qué tiene que cumplir esta teoría para dar una explicación completamente satisfactoria de este fenómeno. La sección 4 propone un argumento escéptico que establece mi tesis negativa: TCE no puede dar una explicación completamente satisfactoria de la agencia epistémica porque la combinación de externismo activo y fiabilismo de virtudes conduce al escepticismo. La sección 5 defiende mi tesis positiva: TCE puede dar una explicación completamente satisfactoria de la agencia epistémica adoptando el relativismo epistémico como principio de segundo orden. Si esta línea de argumento funciona, TCE puede dejar intactos sus principios explicativos de primer orden con hacer un solo ajuste en su metaepistemología.

2. *Conocimiento extendido*

TCE emplea sistemáticamente los recursos del externismo, tanto en filosofía de la mente como en epistemología, para estudiar las dimensiones artefactual y social del conocimiento. TCE consiste en la combinación de dos principios explicativos de primer orden: *externismo activo*, sobre la extensión y distribución de la cognición, y *fiabilismo de virtudes*, sobre el estatus epistémico de los procesos de formación de creencias.

2.1. Externismo activo

Este principio establece que la cognición puede ampliarse más allá del organismo de los individuos a través del uso de artefactos y de la cooperación con otros individuos. En términos técnicos:

> (EXTERNISMO ACTIVO) Los procesos cognitivos están centrados en las facultades orgánicas de agentes individuales, pero también están *extendidos* en artefactos que los agentes emplean y distribuidos entre otros individuos con los que cooperan, interactuando continua y recíprocamente unos con otros.

La idea de que la cognición no es exclusivamente intraorgánica se basa en dos hipótesis de las ciencias cognitivas, las cuales consideran sistemas y procesos cognitivos como unidades de análisis apelando a la teoría de sistemas dinámicos (Palermos, 2014b). Primero, la hipótesis de *cognición extendida* enfatiza que los artefactos son extensiones del sistema cognitivo de un agente, de modo que un sistema cognitivo puede incluir artefactos como partes propias del mismo. Por ejemplo:

> *Caso #1: Observación telescópica*. Galileo cree que Venus tiene fases como la Luna haciendo observaciones a través del telescopio. Inicialmente, sólo logra ver una imagen borrosa y distorsionada de Venus. Minutos después, bajo condiciones atmosféricas más favorables, estabiliza la montura y ajusta los lentes del telescopio, moviéndolo cuidadosamente hasta obtener una imagen nítida

> del planeta en el ocular. Finalmente, dibuja la forma de Venus en una hoja de papel, mirando varias veces la imagen que aparece en el ocular hasta completar un dibujo preciso de la misma. Galileo repite este mismo proceso durante meses, obteniendo reportes observacionales de las distintas fases de Venus según su posición respecto del Sol.[1]

Este caso ilustra un proceso cognitivo extendido realizado por un sistema cognitivo igualmente extendido. El sistema extendido está compuesto por el agente (sus facultades cognitivas orgánicas) y el artefacto externo que emplea (el telescopio). En la medida en que Galileo *interactúa continua y recíprocamente* con el telescopio, ambos conforman un sistema cognitivo *acoplado* que genera un proceso cognitivo extendido. Este proceso está *integrado apropiadamente* al carácter cognitivo del agente, quien emplea dicho proceso para formar una creencia específica.[2]

Por su parte, la hipótesis de *cognición distribuida* subraya que los procesos cognitivos también están distribuidos entre los miembros de un grupo, de modo que un sistema cognitivo puede incluir otros individuos como partes propias. Por ejemplo:

> *Caso #2: Validación de datos.* Belarmino escribe una carta a los matemáticos del Colegio Romano solicitando validar los reportes observacionales de Galileo. Clavio se reúne con los demás expertos y les dice: «Bueno, hemos estado observando a Venus con nuestros telescopios durante algún tiempo», y luego pregunta: «Grienberger, ¿cuál era la forma de Venus vespertino cuando alcanzó su máxima elongación? La mediste con el cuadrante que construiste, ¿no?» Grienberger responde: «Sí, medí su elongación y vi con mi telescopio que había pasado de gibosa a creciente». Lembo, uno de sus estudiantes, agrega: «Correcto, y aproximadamente un mes después, lo observamos como una delgada medialuna menguante, que se hacía cada vez más grande en tamaño aparente». Clavio pregunta: «Y tú, Maelcote, ¿qué recuerdas?» Él responde: «Cerca de cinco meses después de las mediciones de Grienberger, observé a Venus como estrella de la mañana, en fase de creciente, con su parte iluminada aumentando a medida que su diámetro aparente se reducía, y también pude determinar con el cuadrante que estaba próximo al otro punto de máxima elongación». Grienberger interpela a los demás: «¿También notaron que la parte iluminada de Venus siempre estaba de cara al Sol?» Los tres asintieron: «¡Así es!» Entonces Lembo exclama: «¡Todo esto sugiere que su centro de movimiento es el Sol, así como la Tierra es el centro de movimiento de la Luna!». Clavio añade: «Sí... Los matemáticos tendremos que revisar el modelo Ptolemaico a la luz de estos datos... El punto importante, para responder a Belarmino, es que nuestras observaciones coinciden con las de Galileo». Días después, Belarmino recibe en su despacho del Santo Oficio una carta firmada por los cuatro matemáticos donde validan los reportes observacionales de Galileo, formando así la creencia sobre las fases de Venus.

Este caso ilustra un proceso cognitivo distribuido realizado por un sistema de cognición distribuida. Dicho sistema está compuesto por los agentes individuales y los artefactos como sus extensiones cognitivas.[3] En la medida

---

[1] Todos los casos considerados en este artículo derivan del famoso episodio histórico de Galileo usado con frecuencia en los debates sobre relativismo en epistemología (ej. Baghramian & Coliva, 2019; Bland, 2018; Boghossian, 2006; Kusch, 2021). Mis reconstrucciones esquemáticas se basan directamente en fuentes históricas primarias y secundarias autorizadas, incluyendo Biagioli (1993, 2006), Biancani (1620), Carolino (2008), Dear (1987), Feldhay (1995), Galilei (1610/1989, 1890-1909), Gorman (2020), Lattis (1994), Scheiner (1612), van Helden (1996) y Westfall (1985, 1989).

[2] La idea general es que, «en estos casos, el organismo humano está vinculado a una entidad externa en una interacción bidireccional que crea un sistema acoplado que puede verse como un *sistema cognitivo* en sentido propio» (Clark & Chalmers, 1998, p. 8).

[3] La diferencia entre «extensión» y «distribución» es sólo de *énfasis*, pues la hipótesis de cognición distribuida también tiene en cuenta a los artefactos como extensiones cognitivas: esta hipótesis «amplia el alcance de lo que se considera cognitivo más allá del individuo para incluir interacciones entre personas y con recursos y materiales en el ambiente» (Hollan *et al.*, 2000, p. 2). A lo largo del artículo me referiré a sistemas cognitivos «extendidos y distribuidos» porque TCE adopta tanto la hipótesis de cognición extendida como la hipótesis de cognición distribuida.

en que los matemáticos interactúan continua y recíprocamente entre sí (junto con sus telescopios y cuadrantes), constituyen un sistema cognitivo más amplio que consiste en el acoplamiento de las facultades cognitivas orgánicas de estos agentes, generando así un proceso cognitivo distribuido. Aunque un proceso distribuido puede estar integrado apropiadamente al carácter cognitivo de uno solo de sus agentes individuales para producir una creencia particular (por ejemplo, integrarse al carácter cognitivo de Clavio para formar la creencia de que el modelo Ptolemaico es revisable), el acoplamiento de los caracteres cognitivos individuales «permite la existencia de caracteres cognitivos distribuidos [...] que emerge de las interacciones socio-epistémicas de los miembros del grupo» (Palermos & Pritchard, 2016, p. 118). Por ejemplo, las interacciones entre los matemáticos del Colegio Romano dan lugar a un sistema cognitivo que es diferente a la suma de las facultades cognitivas orgánicas de cada matemático.

Hay que destacar algunos aspectos generales detrás de las dos hipótesis anteriores. Un *sistema cognitivo* se concibe como un conjunto de elementos interrelacionados que constituyen el carácter cognitivo de un agente. La interacción dinámica entre esos elementos genera y se basa en procesos cognitivos de distinto tipo, y los sistemas se individúan a partir de los procesos cognitivos que realizan. El *carácter cognitivo* consta principalmente de las facultades cognitivas orgánicas «del cerebro/sistema nervioso central, incluyendo las facultades perceptuales naturales, la memoria y el sistema doxástico en general» (Palermos, 2015, p. 2959). Sin embargo, también se compone de «métodos adquiridos de investigación, incluyendo los que involucran entrenamiento altamente especializado o incluso tecnología avanzada» (Greco, 1999, p. 287). Adicionalmente, los elementos externos al organismo, a través de los cuales se extienden y distribuyen los procesos cognitivos, cuentan como partes propias del sistema cognitivo cuando tiene lugar una *integración cognitiva*. Según Greco (2010), «la integración cognitiva es una función de la cooperación e interacción, o de la interacción cooperativa con otros aspectos del sistema cognitivo» (p. 152).

Más precisamente, los agentes integran apropiadamente otros elementos a su carácter cognitivo a través de la *interacción continua y recíproca* en términos de *bucles de retroalimentación*. En estos bucles, uno de los elementos constituyentes del sistema cognitivo causa efectos en otro elemento constituyente, que a su vez genera efectos en el comportamiento inicial del uno, a partir del cual se causa un nuevo efecto en el otro, y así sucesivamente. Como resultado de esta forma de interacción, emerge un sistema cognitivo extendido o distribuido más amplio, cuya existencia debe postularse por las siguientes dos razones.[4]

La primera es que hay *propiedades emergentes irreductibles*: la interacción continua y recíproca «genera nuevas propiedades sistémicas [...] que pertenecen sólo al sistema en general y no a los subsistemas contribuyentes por separado» (Palermos & Pritchard, 2016, p. 113). En Caso #2, ninguno de los matemáticos produce individualmente la creencia sobre la validación de los reportes observacionales de Galileo, sino sólo la interacción cooperativa entre todos los miembros del grupo. La segunda razón es que los *elementos constituyentes del sistema son interdependientes*: dado que la interacción se basa en bucles de retroalimentación, es imposible explicar las actividades de alguno de los elementos del sistema con independencia del sistema completo, pues el comportamiento de un subsistema siempre determina el comportamiento de otro subsistema (y viceversa) (Palermos, 2014b, pp. 15-18). En Caso #1, el telescopio se iba ajustando mientras Galileo corregía simultáneamente sus dibujos, generando distintas retroalimentaciones sucesivas hasta tener una imagen nítida en el ocular y completar a su vez un dibujo preciso de la figura de Venus en el papel. Claramente, este proceso cognitivo sólo puede adscribirse al sistema cognitivo extendido, y las actividades de Galileo y lo que sucede con el telescopio sólo pueden entenderse en el contexto más amplio del sistema completo y su dinámica.[5]

---

4 Palermos (2014b) desarrolla con mucho detalle esta visión de la *interacción continua y recíproca* sobre la base de bucles de retroalimentación, que cuenta en últimas como criterio de constitución de sistemas cognitivos extendidos y distribuidos.

5 Parafraseando a Clark (2008, p. 24), hay un sistema cognitivo más amplio, compuesto por dos subsistemas, *S* y *O*, cuando *S* afecta continuamente a *O* y *S* es afectado simultáneamente por la actividad de *O*.

https://doi.org/10.1387/theoria.26981

2.2. Fiabilismo de virtudes

Los procesos de cognición extendida y distribuida incluyen *procesos de formación de creencias* a las que puede adjudicarse un estatus epistémico positivo. Para ver cómo las creencias de un agente satisfacen estándares epistémicos, TCE recurre a una variante del externismo en epistemología que permite caracterizar los procesos de formación de creencias como *procesos conducentes al conocimiento*. En términos abstractos:

> (FIABILISMO DE VIRTUDES) Si *S* conoce, entonces la creencia de *S* de que *p* es verdadera en virtud de la manifestación de la habilidad cognitiva del proceso de formación de creencias, de modo que el éxito sea atribuido a *S*. (Palermos & Pritchard, 2016, p. 214)

Este principio concibe el conocimiento en términos de «creencia verdadera acreditable» (Greco, 1999, 2004, 2007, 2010). Según esta forma de fiabilismo, debe existir una relación de *relevancia explicativa* entre el estatus de la creencia verdadera y el estatus del proceso fiable de formación de esa creencia a fin de evitar la suerte epistémica involucrada en los casos Gettier. La idea es que una creencia verdadera cuenta como conocimiento sólo si el sujeto la ha formado en virtud de cierta habilidad que garantiza que la creencia verdadera no fue adquirida por suerte. Así, para tener conocimiento no es suficiente con que un proceso sea fiable (en el sentido de producir un mayor porcentaje de creencias verdaderas que falsas) ni que un sujeto disponga de algunas habilidades cognitivas como partes integrales de su carácter cognitivo. Antes bien, el conocimiento requiere una *relación causal* entre la habilidad y la verdad, donde la habilidad es la que produce de hecho la creencia verdadera. De esta forma, la creencia verdadera es acreditable al agente porque la ha adquirido de manera exitosa más que afortunada. Para este tipo de fiabilismo, las atribuciones de conocimiento tienen como condición necesaria las atribuciones de crédito epistémico.

TCE también plantea que un proceso de formación de creencias es una habilidad cognitiva cuando el agente integra dicho proceso a su carácter cognitivo. Para esto, el proceso puede ser una *disposición normal* del agente y debe estar en *interacción continua y recíproca* con otros aspectos de su cognición. En específico, que el proceso no sea extraño ni esporádico en relación con el carácter cognitivo del agente, y que el proceso haya conducido al éxito epistémico del agente en el pasado, permite que ambos interactúen mutuamente entre sí de manera continua (Palermos, 2015, p. 2962). Esta interacción cooperativa entre el agente y el proceso de formación de creencias da lugar a la integración cognitiva en virtud de la cual el proceso es cognitivamente hábil, respaldando así la atribución de crédito epistémico necesaria para el conocimiento.

EXTERNISMO ACTIVO y FIABILISMO DE VIRTUDES guardan una relación de dependencia que no es trivial. El primero permite caracterizar el segundo para entender el estatus epistémico de los procesos de formación de creencias. La idea abstracta es esta: un proceso *F* es una habilidad cognitiva del agente *S* que produce la creencia verdadera de que *p*, porque *F* está integrado apropiadamente al carácter cognitivo de *S* en la medida en que *S* interactúe de manera continua y recíproca con *F*. Así, la formulación adecuada de la combinación de estos dos principios se expresa mejor bajo el siguiente principio débil de agencia cognitiva:

> (COGA-WEAK) Si *S* sabe que *p*, entonces la creencia verdadera de *S* de que *p* es el producto de un proceso fiable de formación que está integrado apropiadamente al carácter cognitivo de *S*, de modo que el éxito cognitivo es acreditable en un grado significativo a su agencia cognitiva. (Pritchard, 2010, pp. 136-137)

Hay dos presupuestos importantes detrás de COGA-WEAK.[6] En primer lugar, este principio concibe la *agencia cognitiva* como un atributo de seres humanos, que consiste en la capacidad de ejercer control sobre los procesos

[6] Este principio presenta una versión débil de FIABILISMO DE VIRTUDES, pues sólo toma la manifestación de la habilidad cognitiva como *condición necesaria* del conocimiento y reconoce que existen *otros factores relevantes* aparte de la habilidad que contribuyen al éxito epistémico del agente. Esto contrasta con versiones fuertes de FIABILISMO DE VIRTUDES, que conciben la manifestación de la ha-

de formación de creencias en términos de integrar, configurar, mantener, monitorear y modificar dichos procesos. La agencia cognitiva es primariamente humana porque la cognición está centrada en el organismo. EXTERNISMO ACTIVO asume que «el organismo (y dentro del organismo el cerebro/sistema nervioso central) permanece siendo el centro y el elemento más activo», si bien «el procesamiento cognitivo humano (algunas veces) se extiende en el ambiente que rodea al organismo» (Clark, 2007, p. 192). Además, los humanos son quienes reciben adscripciones de crédito epistémico porque tienen intencionalidad. FIABILISMO DE VIRTUDES asume que las «atribuciones de responsabilidad, elogio o meramente acción neutral (es decir, atribuciones de crédito positivo, negativo o meramente neutral) se asocian tradicionalmente con agentes intencionales» (Palermos, 2015, p. 2972).

Sobre esta base, el *agente* que recibe atribuciones de crédito epistémico puede entenderse en términos tanto individuales como grupales. Algunas veces, el crédito epistémico se atribuye significativamente a la agencia cognitiva de las facultades orgánicas de un sujeto individual. En otras ocasiones, se atribuye más bien al acoplamiento coordinado de las facultades cognitivas de los individuos que constituyen un grupo. En Caso #1, Galileo es quien recibe el crédito por formar la creencia verdadera sobre las fases de Venus con el telescopio. Aquí, «el éxito cognitivo del agente aún es acreditable significativamente a sus facultades cognitivas orgánicas, aunque la creencia verdadera sea el producto de un proceso extendido» (Palermos, 2015, p. 2963). En Caso #2, Clavio recibe crédito epistémico por formar la creencia sobre los reportes observacionales de Galileo como producto del proceso distribuido del conjunto de matemáticos del Colegio Romano. El crédito restante lo tiene la agencia cognitiva de su comunidad epistémica que «contribuye al desarrollo de dicho proceso de formación de creencias, proporcionando incluso más subprocesos de formación de creencias o datos producidos por medio de ellos» (Palermos, 2015, p. 2972).

Con respecto a las atribuciones de crédito epistémico de *agentes grupales*, la unidad de análisis sería «caracteres cognitivos distribuidos que consisten en la habilidad cognitiva colectiva que emerge de las interacciones socio-epistémicas de los miembros del grupo» (Palermos & Pritchard, 2016, p. 118). En Caso #2, la creencia verdadera sobre los reportes observacionales de Galileo fue producida sólo por el grupo de matemáticos. Por lo tanto, el éxito epistémico sólo es *acreditable* a la agencia colectiva del grupo como un todo. La idea es que «el éxito cognitivo colectivo de creer la verdad de alguna proposición (científica) será todavía acreditable significativamente a la agencia cognitiva del grupo, esto es, al acoplamiento de las facultades cognitivas de los miembros individuales» (Palermos & Pritchard, 2016, p. 119). Así, dependiendo de cuál sea el foco de la evaluación epistémica en cada caso, el agente epistémico puede especificarse como uno de los individuos o como el grupo en general cuando se trata de sistemas cognitivos extendidos y distribuidos.

El otro presupuesto detrás de COGA-WEAK es que la noción de interacción continua y recíproca, que da lugar a la integración cognitiva, es un criterio tanto de lo cognitivo como de lo epistémico. Con respecto a EXTERNISMO ACTIVO, vimos que este criterio facilita la individuación de *sistemas cognitivos extendidos y distribuidos*:

> Para tener un sistema cognitivo extendido o incluso distribuido [...], todo lo que se requiere es que los miembros contribuyentes (a saber, los agentes cognitivos relevantes y sus artefactos) *interactúen continua y recíprocamente* (sobre la base de bucles de retroalimentación) unos con otros. (Palermos & Pritchard, 2016, pp. 112-113)

---

bilidad como *condición necesaria y suficiente* del conocimiento, y proponen que el éxito epistémico del agente es atribuible *primariamente* a su agencia cognitiva. Bajo estas versiones fuertes, el principio correspondiente sería COGA-STRONG, según el cual hay conocimiento si y sólo si el éxito epistémico de *S* es primariamente acreditable a su habilidad cognitiva (Pritchard, 2010, p. 137), o su iteración COGA-STRONG*, donde el éxito epistémico de *S* como acreditable primariamente a su habilidad cognitiva es sólo condición necesaria del conocimiento (Pritchard, 2010, p. 140). En todo caso, Pritchard (2010, pp. 148-149) sostiene que COGA-WEAK es más adecuado para explicar casos de adquisición de conocimiento con base en procesos de cognición extendida y distribuida, donde típicamente la contribución de la habilidad cognitiva al éxito epistémico del agente es significativa más que primaria de acuerdo con nuestras intuiciones.

https://doi.org/10.1387/theoria.26981

Con respecto a FIABILISMO DE VIRTUDES, sin embargo, esta interacción continua y recíproca no es más que el criterio de individuación de *procesos conducentes al conocimiento*:

> De esta manera, podemos decir que la única condición necesaria y suficiente para que un proceso cuente como conducente al conocimiento es que el proceso interactúe con el resto del carácter cognitivo del agente. (Palermos, 2014a, p. 1941)

Por lo tanto, hay una caracterización de FIABILISMO DE VIRTUDES en términos de EXTERNISMO ACTIVO de acuerdo con COGA-WEAK. TCE da cuenta de la agencia epistémica con base en esta combinación de principios explicativos.

3. Agencia epistémica

Dado que COGA-WEAK está en las coordenadas del externismo epistemológico, TCE concibe la agencia epistémica en términos condicionales:

> Si el proceso de formación de creencias interactúa cooperativamente con otros aspectos del sistema cognitivo, aquel puede monitorearse continuamente en un segundo plano, de modo que, si hay algún error con el mismo, entonces el agente será capaz de detectarlo y responder apropiadamente. De otro modo, si el agente no tiene razones negativas sobre su proceso de formación de creencias, puede estar justificado subjetivamente en emplear el proceso relevante *por defecto*, incluso si carece en absoluto de razones positivas de si el proceso es fiable o de por qué lo sería. (Palermos, 2015, p. 2961)

COGA-WEAK rescata la responsabilidad epistémica del agente. El agente es responsable epistémicamente frente a sus compromisos doxásticos aun cuando no posea evidencia a favor de la habilidad cognitiva de proceso de formación de creencias. El fiabilismo de procesos ha enseñado que un sujeto es inocente epistémicamente al adquirir creencias con base en un proceso fiable de formación, a menos que se demuestre que el proceso es culpable epistémicamente y aun cuando ese sujeto no tenga razones positivas a favor de su inocencia (Alston, 1982, p. 7; ver también Goldman, 1986). Sin embargo, el fiabilismo de procesos más conservador sostiene que las atribuciones de conocimiento no tienen como condición necesaria las atribuciones de crédito epistémico (Kornblith, 2016). Un sujeto tiene conocimiento si su creencia está *justificada objetivamente* porque ha sido el resultado de un proceso fiable. Sin embargo, el sujeto no está *justificado subjetivamente* porque no puede *dar cuenta de* la fiabilidad de su proceso de formación de creencias. Por ejemplo, Belarmino tuvo conocimiento de los reportes observacionales de Galileo. Sin embargo, no recibe ningún crédito epistémico porque al parecer no era competente para determinar la fiabilidad del conjunto de observaciones astronómicas del Colegio Romano que validaron dichos reportes.

En contraste, COGA-WEAK lleva a tomar la responsabilidad epistémica como un elemento constitutivo del conocimiento, según el cual «un sujeto debe poseer algún grado de entendimiento de su fiabilidad, en particular de sus limitaciones: de cómo es posible estar equivocado» (Williams, 2015, p. 258). El entendimiento que el agente tiene sobre el estatus epistémico de sus procesos de formación de creencias adquiere dos aspectos. De un lado, el agente es responsable epistémicamente sin tener evidencia a favor de la habilidad cognitiva de su proceso de formación, cuando integra dicho proceso a su carácter cognitivo para formar creencias verdaderas bajo ciertas circunstancias en las que el uso del proceso ha garantizado el éxito epistémico. En otras palabras, el agente está *justificado subjetivamente* al obtener respaldo epistémico *prima facie* que lo acredita a seguir usando el proceso, porque ha sido epistémicamente exitoso en situaciones semejantes y no hay razones negativas de mayor peso para dudar. Por ejemplo, Galileo sí recibe crédito epistémico por su conocimiento de las fases de Venus porque, a diferencia de Belarmino, había usado previamente la observación telescópica con los cuatro satélites de Júpiter y otros descubrimientos en su *Side-*

reus nuncius (Galilei, 1610/1989). Además, ya no tenía razones en contra de la fiabilidad de la observación telescópica ni de sus observaciones particulares de dichas fases.

De otro lado, el agente es responsable epistémicamente al tener evidencia en contra del proceso de formación de creencias cuando tiene una disposición a *detectar* el error con su proceso de formación y a *responder* apropiadamente haciendo la revisión correspondiente de su compromiso doxástico. En este segundo caso, el agente está *justificado subjetivamente* al ser sensible a razones que ponen en cuestión su respaldo epistémico *prima facie*. Por ejemplo, los matemáticos del Colegio Romano concluyeron que los reportes observacionales de Galileo eran inconsistentes con el modelo Ptolemaico, usado tradicionalmente para formar creencias sobre el movimiento orbital de Venus. Clavio reconoció la necesidad de construir un modelo alternativo, y su comunidad epistémica refinó el modelo de Tycho Brahe durante años hasta aceptarlo oficialmente en *Sphaera mundi* una década más tarde (Biancani, 1620). Esta disposición del agente para ejercer control sobre el proceso de formación de creencias está garantizada bajo un presupuesto doble. Primero, el proceso pertenece al carácter cognitivo: «el proceso de formación de creencias está integrado al carácter cognitivo de uno, de modo que uno estaría en posición de responder si algo estuviera mal (con el proceso)» (Palermos, 2015, p. 2961).[7] Segundo, las razones derrotantes son creencias que también pertenecen al carácter cognitivo: «lo que cuenta como evidencia derrotante no es independiente de las disposiciones cognitivas de facto del sujeto» (Greco, 2010, p. 167).

Teniendo esto en cuenta, la forma técnica de formular los dos aspectos de la agencia epistémica según TCE sería la siguiente. Primero tenemos la *agencia epistémica sin derrotantes*:

> (AGENCIA$_1$) Si *S* no tiene razones negativas sobre su proceso de formación de creencias, *F*, entonces *S* es responsable epistémicamente *por defecto* en creer que *p* con base en *F*.

También podemos caracterizar la *agencia epistémica ante derrotantes* de la siguiente forma:

> (AGENCIA$_2$) Si *S* tiene razones negativas sobre su proceso de formación de creencias, *F*, entonces *S* es responsable epistémicamente al estar dispuesto a detectar el error en *F* y a revisar su creencia de que *p* basada en *F*.

Esta caracterización, sin embargo, sólo ofrece una explicación incompleta de la agencia epistémica. AGENCIA$_2$ tendría que reformularse de tal modo que pueda dar cuenta de cómo el agente *de hecho* detecta el error y revisa su creencia en virtud de su agencia cognitiva. Aparte de tener la disposición para hacer ambas cosas, se espera que el agente sea responsable epistémicamente descubriendo y revisando lo que está mal. En este caso, el control doxástico responsable epistémicamente requiere que la creencia del sujeto esté sujeta a escrutinio crítico cuando existen razones para poner en cuestión el estatus epistémico del proceso de formación. Para este fin, el agente debe *manifestar su disposición* a revisar su creencia ante razones derrotantes. En términos técnicos:

> (AGENCIA*) Si *S* tiene razones negativas sobre su proceso de formación de creencias, *F*, entonces *S* es responsable epistémicamente al manifestar su disposición a detectar el error en *F* y a revisar su creencia de que *p* basada en *F*.

Esta reformulación es importante porque AGENCIA$_2$ implicaría atribuir responsabilidad epistémica a un agente por el simple hecho de tener la disposición a responder apropiadamente ante razones derrotantes, incluso cuando no lleve a cabo dicha respuesta y continue sosteniendo dogmáticamente su creencia de que *p* con base en *F*. Por ejemplo, si el Colegio Romano no hubiese abandonado la creencia de que Venus gira alrededor de la Tierra según el modelo Ptole-

---

7 Palermos (2014a) también sugiere que la agencia cognitiva se basa en mecanismos de monitoreo y control de procesos cognitivos, en particular la fluidez: «los sujetos recurren al razonamiento analítico sólo cuando experimentan el efecto metacognitivo de la carencia de fluidez» (p. 1942, n. 17).

maico, después de detectar su inconsistencia con los reportes observacionales de Galileo, entonces este agente grupal no hubiese sido responsable epistémicamente ante esos reportes. En cambio, sabemos que los matemáticos Jesuitas no sólo abandonaron dicha creencia, sino que respondieron ante la evidencia observacional desarrollando conjuntamente un modelo alternativo donde Venus se mueve alrededor del Sol. El Colegio Romano sí fue responsable epistémicamente como agente grupal manifestando su disposición a revisar sus compromisos doxásticos originales.

Así pues, la agencia epistémica ante derrotantes conjuga estos dos aspectos: *tener cierta disposición cognitiva y manifestar dicha disposición cognitiva*. Greco (2010) define la responsabilidad epistémica así: «La creencia de *S* de que *p* es responsable epistémicamente […] si y sólo si la creencia de *S* de que *p* resulta de las disposiciones intelectuales que *S* manifiesta cuando *S* está motivado a creer la verdad» (p. 43). De la misma manera que las atribuciones de crédito epistémico involucran una relación causal entre la manifestación de la habilidad cognitiva y el éxito epistémico al formar la creencia verdadera, la responsabilidad epistémica (y el estatus de justificación subjetiva) implica una relación del mismo tipo entre la manifestación de la habilidad y la respuesta apropiada frente a razones derrotantes. AGENCIA* captura este otro aspecto que la concepción original de TCE sobre la agencia no toma en consideración.

A primera vista, TCE podría dar cuenta de AGENCIA* con base en sus principios explicativos. Recordemos que su elemento clave es la idea de que «un proceso de formación de creencias cuenta como una habilidad cognitiva y, por lo tanto, como conducente al conocimiento, si y sólo si es un proceso fiable de formación de creencias que está integrado al carácter cognitivo del agente con base en un proceso de interacción cooperativa con el mismo» (Palermos, 2014a, p. 1944). Por desgracia, mostraré a continuación que esta combinación de EXTERNISMO ACTIVO y FIABILISMO DE VIRTUDES en términos de COGA-WEAK hace que TCE no pueda explicar cómo el agente exhibe AGENCIA*. Por lo tanto, TCE no sería una teoría completamente satisfactoria de la agencia epistémica.

4. *Tesis negativa: TCE conduce al escepticismo*

La objeción escéptica plantea esto: si TCE caracteriza FIABILISMO DE VIRTUDES a partir de EXTERNISMO ACTIVO de acuerdo con COGA-WEAK, entonces TCE no puede explicar AGENCIA* sin caer en el escepticismo. Para justificar esta afirmación, el escéptico tiene disponible el siguiente argumento:

(ARGUMENTO GENERAL)

[Premisa 1] *S* sabe que *p* sólo si el éxito epistémico de *S* al formar la creencia verdadera de que *p*, con base en un proceso *F* integrado apropiadamente a su carácter cognitivo, es acreditable significativamente a su agencia cognitiva.

[Premisa 2] El éxito epistémico de *S* al formar la creencia verdadera de que *p* con base en *F* es acreditable significativamente a su agencia cognitiva sólo si se satisface el principio de DISCRIMINACIÓN respecto de AGENCIA* y sólo si se satisface el principio de NO-DUDA respecto de AGENCIA_1.

[Premisa 3] No se satisface el principio de DISCRIMINACIÓN respecto de AGENCIA*.

[Conclusión] Por lo tanto, es falso que el éxito epistémico de *S* al formar la creencia verdadera de que *p* con base en *F* sea acreditable significativamente a su agencia cognitiva (y, por lo tanto, es falso que *S* sabe que *p*).

No se aprecia ninguna relación obvia entre las premisas de este argumento y lo que la objeción escéptica plantea. Hace falta hacer explícitas las razones que justifican algunas de las premisas para ver la conexión. Si [Premisa 1] es simplemente otra manera de formular FIABILISMO DE VIRTUDES, la solidez de ARGUMENTO GENERAL reside en los elementos adicionales que introduce en [Premisa 2] y [Premisa 3].

4.1. Segunda premisa escéptica

> *[Premisa 2] El éxito epistémico de S al formar la creencia verdadera de que p, con base en un proceso F integrado apropiadamente a su carácter cognitivo, es acreditable significativamente a su agencia cognitiva sólo si se satisface el principio de DISCRIMINACIÓN respecto de AGENCIA* […].*

Esta afirmación tiene que ver con las condiciones mínimas (aunque suficientes) bajo las cuales la habilidad cognitiva explica de manera relevante el éxito epistémico del agente. Según FIABILISMO DE VIRTUDES, la relación entre el estatus epistémico de las creencias y el estatus epistémico de los procesos de formación de creencias es una *relación de explicación*: una creencia verdadera adquiere el estatus de conocimiento sólo si *fue producida gracias a la manifestación de la habilidad cognitiva* (Palermos, 2015, pp. 2960-2961; Palermos & Pritchard, 2016, pp. 113-114). De igual manera, un agente merece adjudicación de crédito epistémico si emplea procesos que son conducentes al conocimiento en tanto que dichos procesos cuentan como sus habilidades cognitivas (Pritchard, 2010).

Sobre esta base, ARGUMENTO GENERAL toma la explicación del éxito epistémico como una *explicación contrastiva* (Garfinkel, 1981; Lipton, 1990). Para entender cómo un proceso que produce una creencia verdadera cuenta como una habilidad cognitiva (de modo que el agente merezca adjudicación de crédito epistémico por formar dicha creencia), no es suficiente con que la habilidad cognitiva produzca la creencia verdadera, sino que además dicha habilidad sea el *explanans relevante* de la creencia verdadera. En otras palabras, la explicación del éxito epistémico no consiste en saber únicamente por qué el proceso es cognitivamente hábil, sino más bien en saber por qué el proceso es conducente al conocimiento *en contraste con* al menos otro proceso que, si bien es potencialmente hábil para generar la creencia verdadera, no es conducente al conocimiento.

Adicionalmente, esta explicación contrastiva ha de satisfacer dos condiciones mínimas que pueden formularse en los siguientes términos: la habilidad cognitiva explica el éxito epistémico del agente *S* al formar la creencia verdadera de que *p*, con base en un proceso *F* integrado apropiadamente a su carácter cognitivo, si y sólo si: (a) *S* cree que *p* con base en *F* y (b) *S* excluye otros procesos distintos a *F*, ya sea porque se cumple alguno de los siguientes dos principios en conformidad con AGENCIA_1 y AGENCIA*. Este es el principio correspondiente a AGENCIA_1:

> (NO-DUDA) No hay otros procesos potencialmente hábiles que llevan a *S* a dudar que el proceso *F*, con base en el cual formó la creencia de que *p*, es conducente al conocimiento.

Y este es el principio correspondiente a AGENCIA*:

> (DISCRIMINACIÓN) *S* discrimina el proceso que es cognitivamente hábil cuando hay otros procesos potencialmente hábiles que llevan a *S* a dudar que el proceso *F*, con base en el cual formó la creencia de que *p*, es conducente al conocimiento.

El agente debe satisfacer NO-DUDA a fin de ser responsable epistémicamente de acuerdo con AGENCIA_1. La agencia epistémica sin razones derrotantes sólo requiere que el agente esté *justificado subjetivamente* en continuar empleando el proceso integrado a su carácter cognitivo a menos que tenga buenas razones para dudar del mismo. Sin embargo, NO-DUDA no tiene por qué satisfacerse en aquellas situaciones donde el agente encuentra buenas razones para dudar de su proceso de formación de creencias. Ante razones derrotantes, el agente ha de satisfacer DISCRIMINACIÓN para ser responsable epistémicamente en términos de AGENCIA*.

En este segundo caso, el agente es responsable epistémicamente porque no sólo tiene la disposición a descubrir que hay un error con su proceso de formación de creencias y a revisar críticamente su compromiso doxástico, sino que también manifiesta las disposiciones relevantes para ejecutar dicha tarea. En particular, el agente se entera de que hay otros procesos potencialmente hábiles y entonces discrimina el proceso que es conducente al conocimiento. Es importante señalar que la creencia en la existencia de otros procesos potencialmente hábiles es una razón que derrota la justificación subjetiva que el agente tenía al emplear por defecto su proceso de formación de creencias. Con la satisfacción de DISCRIMINACIÓN se restaura su respaldo epistémico *prima facie*. De ahí que, en situaciones donde hay procesos alternativos potencialmente hábiles, no es suficiente con que el agente esté en posición de responder apropiadamente. Antes bien, también se requiere que, con base en manifestar sus disposiciones cognitivas, el agente identifique el proceso relevante que explica el éxito epistémico luego de haber adquirido razones derrotantes en su contra.[8]

Para entender mejor la conexión entre DISCRIMINACIÓN y AGENCIA*, hay que precisar las razones que llevarían al agente a poner en cuestión su respaldo epistémico *prima facie*. Existen al menos dos tipos:

(R1) *Evidencia derrotante de primer orden*: *S* duda de su proceso de formación de creencias cuando tiene evidencia de una inconsistencia (A) entre la creencia de que *p* formada a partir de ese proceso y otro elemento de su propio conjunto de creencias, o (B) entre la creencia de que *p* formada a partir de ese proceso y la creencia de que ~*p* adquirida por otro agente con base en un proceso de formación de creencias diferente.

(R2) *Evidencia derrotante de segundo orden*: *S* duda de su proceso de formación de creencias cuando tiene evidencia de la existencia de procesos alternativos, los cuales (C) fueron tomados como exitosos en el pasado y son rechazados en el presente, o (D) se emplean en el presente y están en conflicto con su proceso de formación de creencias.[9]

(R1) es una formulación general de la subdeterminación, (A) cuando es holista y (B) cuando es contrastiva en casos de desacuerdo epistémico. Un ejemplo de (A) sería el de un científico que encuentra una inconsistencia entre las predicciones de su teoría en contrastación y los datos experimentales que obtuvo a partir de la ejecución coordinada del experimento por parte de su grupo de investigación. Un ejemplo de (B) sería cuando otro científico, que ha corroborado una hipótesis mediante un experimento con su equipo de trabajo, encuentra un artículo recién publicado de otro grupo de investigación que ha replicado el experimento original con resultados inconsistentes con su propia hipótesis. La *evidencia derrotante de primer orden* es un lugar común en nuestras prácticas (científicas) ordinarias.

(R2) tiene que ver con una situación un poco más compleja. La *evidencia derrotante de segundo orden* es aquella que el *analista* de las prácticas (científicas) ordinarias pone a disposición del *actor* de dichas prácticas. Esta evidencia amenaza la legitimidad de la inferencia eliminativa involucrada en las explicaciones contrastivas en casos importantes. El

---

8 Así, sentimientos metacognitivos fiables de monitoreo y control, en particular el «sentimiento de error» determinado por la experiencia de fluidez (Fernandez Cruz *et al.*, 2016; Schwartz & Jemstedt, 2021), jugarían un papel *sintomático* en la respuesta apropiada del agente ante la razón derrotante: el agente *detecta* el error a partir de experimentar dicho sentimiento metacognitivo y con base en este indicador *revisa* su compromiso doxástico valiéndose de sus procesos cognitivos. Más precisamente, la duda que el agente tiene respecto de su proceso de formación de creencias después de creer que hay otros procesos alternativos potencialmente hábiles vendrá acompañada del sentimiento de error determinado por la falta de fluidez. Con base en este sentimiento metacognitivo, la estrategia cognitiva que el agente decidirá ejecutar para responder a la razón derrotante consiste en *discriminar* el proceso que es conducente al conocimiento. Como con todo sentimiento metacognitivo, la falta de fluidez es una mera «señal» de error que motiva al agente a emplear otros procesos cognitivos para resolver el problema. Sin embargo, la metacognición del agente no se encarga de ejecutar la estrategia cognitiva para remediarlo. La discriminación es la estrategia cognitiva en este caso. De hecho, la discriminación podría entenderse como una disposición cognitiva del agente cuya manifestación emplearía herramientas cognitivas adicionales, entre ellas el razonamiento analítico. Agradezco a un revisor anónimo de THEORIA por pedirme aclarar este punto.

9 Esta taxonomía de razones derrotantes se concentra en derrotantes *psicológicos socavadores*: creencias que el agente posee actualmente y que derrotan el estatus epistémico del proceso de formación de creencias. Desde esta visión, los derrotantes psicológicos tienen fuerza normativa porque forman parte del carácter cognitivo del agente. Recordemos que «lo que cuenta como evidencia derrotante no es independiente de las disposiciones cognitivas de facto del sujeto», sino que es «parcialmente una función de los hechos contingentes sobre el sujeto y su carácter cognitivo» (Greco, 2010, p. 167). Sobre tipos de derrotantes, ver Pollock & Cruz (1999, pp. 196-197) y Sudduth (2008).

ejemplo paradigmático de (C) es el argumento de las alternativas inconcebidas basado en evidencia histórica (Stanford, 2006). La metaevidencia derrotante aquí es que los agentes de hoy no conciben *procesos cognitivamente hábiles, alternativos y lo suficientemente distintos* a los procesos con los que forman sus creencias actuales, tal y como los agentes del pasado no concibieron los procesos que los agentes de hoy emplean para formar esas creencias. Los ejemplos de (D) son los mismos casos de desacuerdo epistémico que (B) tiene en cuenta. Sin embargo, aquí la metaevidencia derrotante es el reconocimiento de *la existencia de desacuerdo* más que el reconocimiento de evidencia que derrotaría el estatus epistémico del proceso de formación de una de las partes rivales.

La función de (R1)-(R2) es mostrar al agente que algo anda mal con su proceso de formación de creencias. Así, se espera que el agente responda apropiadamente frente a su compromiso doxástico según el tipo de razones (A)-(D). Sin embargo, es importante señalar que no todas las razones derrotantes que cabrían dentro de estas cuatro opciones serían relevantes de acuerdo con [Premisa 2]. El escéptico necesita considerar únicamente *alternativas existentes* de procesos de formación de creencias para definir el espacio de posibilidades relevantes. El escéptico puede conceder que las «fantasías cartesianas» son alternativas escépticas implausibles, de modo que constituyen posibilidades remotas que no ponen en cuestión el respaldo epistémico *prima facie*. Esto permite diferenciarlas de los casos de prácticas (científicas) ordinarias que serían relevantes para tomar ARGUMENTO GENERAL seriamente.

Es fácil apreciar esta diferencia. De un lado, los escenarios escépticos no clasificarían dentro de (R1). Es implausible que algún agente incluya una hipótesis escéptica en su conjunto de creencias, hipótesis que estaría en contradicción con otras de sus creencias ordinarias. Lo mismo sería aplicable al caso implausible de que un agente encuentre a otro agente que está en desacuerdo con él porque este último suscribe una hipótesis escéptica. De otro lado, los escenarios escépticos tampoco clasificarían dentro de (R2). La razón es simple: no tenemos evidencia de que haya, haya habido o podría haber un proceso de formación que respalde la creencia en un escenario escéptico, que además compita con procesos que los agentes emplean en el presente. En contraste, la evidencia que apoya el problema de las alternativas inconcebidas o el problema de la existencia de desacuerdo muestra procesos alternativos que realmente han sido usados o se emplean en la actualidad. La historia de la ciencia proporciona este tipo de evidencia.

Para resumir: el agente debe satisfacer DISCRIMINACIÓN para ser responsable epistémicamente de acuerdo con AGENCIA*. Ante razones que ponen en duda su proceso de formación de creencias, el agente responde apropiadamente revisando su compromiso doxástico cuando identifica el proceso cognitivamente hábil que es relevante para explicar su éxito epistémico al formar la creencia verdadera. Esta discriminación que el agente hace del proceso conducente al conocimiento se basa en la explicación contrastiva de la habilidad cognitiva.

4.2. Tercera premisa escéptica

> *[Premisa 3] No se satisface el principio de DISCRIMINACIÓN respecto de AGENCIA*.*

¿Cómo discriminaríamos, en el espacio de posibilidades relevantes, aquel proceso de formación de creencias que explica el éxito epistémico al formar la creencia verdadera? Recordemos que «todo lo que necesitamos para aceptar que un proceso es conducente al conocimiento es que esté integrado al carácter cognitivo del sujeto al interactuar cooperativamente con él» (Palermos, 2014a, pp. 1942-1943). Si esto es así, identificar la habilidad cognitiva para hacer la discriminación correspondiente consistiría en individuar la *integración cognitiva apropiada* de la cual la habilidad es una función. Para este fin, la condición de integración cognitiva en COGA-WEAK puede interpretarse considerando dos niveles de individuación. Este es el primero:

> (INTEGRACIÓN DE GRANO GRUESO) *S* integra apropiadamente el proceso de formación de creencias, *F*, a su carácter cognitivo, si y sólo si: *S* y *F interactúan continua y recíprocamente con base en bucles de retroalimentación*, bajo la precondición de que *F* es una *disposición normal* de *S*.

Para identificar la habilidad cognitiva, la interacción continua y recíproca sería el único criterio relevante de individuación de la integración cognitiva entre el agente y su proceso de formación de creencias. Mientras que la normalidad y la disposicionalidad del proceso son únicamente «precondiciones prácticas» que *facilitan* la integración cognitiva, «el efecto decisivo de la integración cognitiva [...] sólo estará garantizado si el carácter cognitivo del agente interactúa mutuamente con el proceso relevante» (Palermos, 2014a, p. 1941). Bajo esta interpretación, el agente satisfaría DISCRIMINACIÓN mediante la especificación del proceso que interactúa continua y recíprocamente con su carácter cognitivo, interacción facilitada por el hecho de que el proceso es una disposición normal del agente para formar creencias.

Sin embargo, INTEGRACIÓN DE GRANO GRUESO es demasiado fácil de satisfacer. Esta interpretación de la integración cognitiva apropiada conlleva el siguiente problema: el agente sería incapaz de discriminar el proceso de formación de creencias que es conducente al conocimiento en la medida en que existan procesos alternativos, relevantes para un mismo conjunto de creencias verdaderas, que son disposiciones normales del agente y que interactúan continua y recíprocamente con su carácter cognitivo. Consideremos otro caso:

> *Caso #3: Confirmación de hipótesis*. El matemático Castelli formó la creencia sobre las fases de Venus después de predecirlas usando el modelo Copernicano y preguntó a su maestro Galileo si había observado telescópicamente este fenómeno. Galileo no sólo respondió públicamente que sus observaciones telescópicas individuales confirmaban el modelo Copernicano, sino que se atribuyó exclusivamente haber descubierto con dichas observaciones que Venus gira alrededor del Sol y que las manchas solares son objetos reales. Mientras tanto Scheiner (1612), un Jesuita que colaboraba con las investigaciones del Colegio Romano desde Alemania, afirmó públicamente haber descubierto que las manchas solares eran objetos reales. Además, enfatizó que la verdad de este descubrimiento y de las fases de Venus estaba respaldada en sus observaciones personales y las de sus colaboradores, así como en la autoridad epistémica de otros matemáticos y filósofos naturales.

Este caso ilustra tres procesos de formación de creencias que satisfacen INTEGRACIÓN DE GRANO GRUESO, pero donde no puede especificarse con claridad cuál de ellos es la habilidad cognitiva que explica el éxito epistémico. Los tres procesos eran disposiciones normales de los agentes, pues todos habían aceptado el uso de modelos matemáticos para hacer predicciones verdaderas sobre el mundo físico y de telescopios para obtener datos observacionales. Aún más importante, los agentes interactuaron cooperativamente con los modelos,[10] con sus telescopios y con otros agentes para formar creencias verdaderas. Sin embargo, ¿cuál de los tres procesos merece adjudicación de crédito epistémico? Sabemos que Galileo no reconoció ningún papel al proceso inferencial de Castelli, atribuyendo el crédito epistémico únicamente a sus observaciones telescópicas individuales en cuanto proceso extendido de formación de creencias. También sabemos que Galileo acusó de plagio a Scheiner sobre las manchas solares, que tampoco reconoció la fiabilidad de las observaciones de los Jesuitas en cuanto proceso distribuido de formación de creencias, y que rechazó la autoridad epistémica de las comunidades matemática y filosófica. En cambio, Scheiner sí dio crédito epistémico a estas comunidades, no sólo a sí mismo o a Galileo.

En términos generales, hay un gran número de procesos asociados a la producción de un mismo conjunto de creencias verdaderas particulares que cumplen con el criterio de interacción cooperativa más las precondiciones prácticas que facilitan esa interacción. La manera como INTEGRACIÓN DE GRANO GRUESO tipifica los procesos de formación de creencias es tan general que el agente no estará en posición de saber cuál es el proceso cognitivamente hábil que explica el éxito epistémico. En consecuencia, DISCRIMINACIÓN no se cumple en relación con AGENCIA*. Este diagnóstico no es más que el problema de la generalidad endilgado a las teorías fiabilistas (Conee & Feld-

[10] Palermos (2015) propone que las teorías científicas «también cuentan como procesos de formación de creencias que pueden extender el carácter cognitivo del agente más allá de sus facultades cognitivas orgánicas» (p. 2966).

man, 1998). TCE enfrenta una versión especial de esta objeción si concibe la integración cognitiva apropiada en términos de INTEGRACIÓN DE GRANO GRUESO. La razón es que *la interacción continua y recíproca basada en bucles de retroalimentación bajo precondiciones prácticas* no ofrece un criterio principioso que permita especificar el nivel relevante de integración cognitiva para identificar la habilidad cognitiva. El escéptico insiste en que la interacción cooperativa, aunque permite individuar procesos *cognitivos* del agente, no es suficiente para individuar sus procesos cognitivos *relevantes epistémicamente*.[11]

La salida a esta dificultad sería considerar la integración cognitiva apropiada con base en un segundo nivel de individuación:

> (INTEGRACIÓN DE GRANO FINO) *S* integra apropiadamente el proceso *F* a su carácter cognitivo si y sólo si: *S* y *F* interactúan continua y recíprocamente con base en bucles de retroalimentación, bajo la precondición de que *F* es una disposición normal de *S*, *en función de los intereses y objetivos pragmáticos que operan en el contexto de la práctica epistémica donde S forma creencias con base en F*.[12]

Según esta interpretación, los objetivos e intereses del agente en el marco de su práctica epistémica son relevantes para individuar la integración cognitiva apropiada requerida para que un proceso sea cognitivamente hábil. Como Greco (2007) sostiene, los parámetros para especificar el nivel relevante de integración cognitiva «se definen por los intereses y propósitos que operan en el contexto relevante del razonamiento práctico» (p. 62). En este caso, dentro de un conjunto de procesos cognitivos, el agente discrimina aquel proceso de formación de creencias que es *relevante epistémicamente* dependiendo del contexto de razonamiento práctico. En determinados contextos y para ciertos propósitos, un proceso será cognitivamente hábil, pero no lo será en otros contextos y para propósitos diferentes.

En relación con Caso #3, sabemos que la ciencia de la época estaba configurada por las dinámicas y valores del sistema de patrocinio de la cultura cortesana en el marco del absolutismo político. Esta cultura patrocinaba científicos para trabajar en la adquisición de verdades novedosas y promovía disputas entre ellos con el fin de legitimar su estatus social. La práctica epistémica de Galileo, en la cual se enmarca su proceso extendido de formación de creencias, incluía el objetivo de establecerse a sí mismo a los ojos de la nobleza y la Iglesia como la única autoridad epistémica en matemáticas y filosofía natural en competencia con los Jesuitas. Por su parte, el proceso distribuido de formación de creencias de Scheiner fue cognitivamente hábil relativo a otra práctica epistémica, cuyo propósito general era convertir al Colegio Romano en la máxima autoridad científica ante las cortes y el papado, desplazando así la figura de Galileo de la escena.

Este caso involucra dos razonamientos prácticos diferentes, donde la motivación de ambos agentes por adquirir verdades novedosas como las fases de Venus y las manchas solares fue congruente con los propósitos e intereses de sus

---

11 El problema de la generalidad es principalmente una dificultad para la evaluación epistémica como tarea de los epistemólogos. Sin embargo, también es un problema para los agentes epistémicos cuando ellos mismos son quienes evalúan sus propias creencias, como sucede cuando son responsables epistémicamente de acuerdo con AGENCIA*. Así, el problema de la generalidad se convierte en un obstáculo tanto para los defensores de TCE como para los agentes mismos. Contrariamente, Goldman (2021) sostiene que la generalidad no es un problema ni para las teorías fiabilistas ni para los agentes epistémicos. Sin embargo, su argumento no funciona en relación con casos de agencia ante derrotantes.

12 INTEGRACIÓN DE GRANO GRUESO e INTEGRACIÓN DE GRANO FINO no deben confundirse con COGA-WEAK y las dos variantes de COGA-STRONG. Las primeras son formulaciones de *integración cognitiva* que relacionan la *agencia cognitiva* con la *interacción continua y recíproca* del agente con su proceso de formación de creencias, mientras que las últimas son versiones de FIABILISMO DE VIRTUDES que establecen una relación entre *conocimiento* y *agencia cognitiva*. Por supuesto, dado que TCE caracteriza FIABILISMO DE VIRTUDES en términos de EXTERNISMO ACTIVO de acuerdo con COGA-WEAK, habría una relación transitiva entre *conocimiento* e *interacción continua y recíproca*: si *S* sabe que *p*, entonces *S* interactúa continua y recíprocamente con el proceso *F* al formar la creencia verdadera de que *p* con base en *F*.

respectivas prácticas epistémicas. En general, INTEGRACIÓN DE GRANO FINO implica que los agentes deciden el nivel de integración cognitiva que es relevante epistémicamente en función de su razonamiento práctico. Además, los propósitos e intereses de las prácticas epistémicas no son independientes del carácter cognitivo de los agentes mismos; simplemente forman parte de los hechos contingentes de la agencia cognitiva de seres humanos (Code, 1987). Este punto es mucho más evidente con procesos extendidos y distribuidos que son altamente sofisticados en términos epistémicos, como los que producen conocimiento científico.

Con INTEGRACIÓN DE GRANO FINO se resuelve el problema de la generalidad si se admite que «la fiabilidad es fiabilidad para propósitos particulares» (Williams, 2015, p. 267; ver también Brandom, 1998; Cartwright, 2020; Kusch, 2002). Además, esta solución pragmatista permitiría la satisfacción de DISCRIMINACIÓN respecto de AGENCIA* si el escéptico recurre al problema de la generalidad para motivar [Premisa 3] de ARGUMENTO GENERAL. Sin embargo, incluso si se concede INTEGRACIÓN DE GRANO FINO, el escéptico podría establecer [Premisa 3] apelando a una de las consecuencias que derivarían de esta respuesta pragmatista a dicho problema. Para el escéptico, esta maniobra parece abrir la puerta a una forma de relativismo epistémico.

5. *Tesis positiva: TCE abraza el relativismo*

El escéptico argumenta que, si los criterios de habilidad cognitiva dependen de los objetivos e intereses situados de los agentes, y si estos objetivos e intereses varían en las prácticas epistémicas y a través de ellas, entonces la discriminación del proceso que es conducente al conocimiento no sería una *elección neutral*. Antes bien, la discriminación se haría con base en los criterios de habilidad cognitiva relativos a las diferentes prácticas epistémicas. Por supuesto, la discriminación sería neutral sólo si hubiese un criterio independiente de habilidad cognitiva para identificar el proceso relevante que explica el éxito epistémico. El siguiente estándar captura esta condición:

> (METAJUSTIFICACIÓN) Cuando hay creencias y/o procesos de formación en conflicto, existen formas de arbitrar procesos como cognitivamente hábiles que son neutrales e independientes respecto de los caracteres cognitivos. (Seidel, 2014; Kusch, 2017)

Bajo este estándar, DISCRIMINACIÓN se satisface en principio. Sin embargo, el escéptico puede presentar casos ejemplares en los que METAJUSTIFICACIÓN no se cumple. Me refiero a casos donde existen criterios de habilidad cognitiva en conflicto porque dependen de intereses y objetivos diferentes que son relativos a prácticas epistémicas rivales. Aunque el escéptico podría motivar esta postura con tan sólo considerar un caso hipotético de desacuerdo fundamental como una mera posibilidad teórica (Bland, 2018; Hales, 2014; Kusch, 2009), ya he señalado que ARGUMENTO GENERAL funciona si se descartan estas posibilidades remotas. Por esto, el escéptico podría motivar su posición más convincentemente apelando a múltiples controversias científicas ampliamente documentadas y analizadas en los estudios de la ciencia (ej. Bloor, 2011; Kinzel & Kusch, 2018; Kusch, 2021).

Caso #3 admite claramente un análisis relativista. Galileo y Scheiner emplearon dos procesos de formación de creencias diferentes, el uno extendido y el otro distribuido. Con base en estos procesos, ambos agentes formaron las mismas creencias verdaderas, pero también produjeron algunas creencias rivales (por ejemplo, Galileo creía en el modelo de Copérnico, mientras que los Jesuitas creían en el modelo de Brahe). A su vez, estos dos procesos de formación de creencias eran relativos a prácticas epistémicas con propósitos e intereses en conflicto. Estos propósitos e intereses fueron parte constitutiva de los respectivos caracteres cognitivos de los agentes. Galileo descubría *verdades científicas* novedosas para impresionar a sus patrones, mientras que los Jesuitas lo hacían para impresionar a sus superiores. Galileo mantenía que la *autoridad epistémica* residía en él mismo como agente individual, mientras que el Colegio Romano insistía en que la autoridad recaía en ellos como agente grupal. Galileo cultivaba la *humildad epistémica* en términos de oponerse a la pedantería escolástica de sus rivales, mientras que Scheiner y Grienberger practicaban la humildad como forma de abnegación cristiana. Finalmente, Galileo reclamaba su *crédito epistémico* como «matemático y filósofo» de la corte de los Médici, mientras que el Colegio Romano reclamó ser la comunidad científica global de la Iglesia que haría frente a la expansión del protestantismo en Europa ante la posición retrógrada e irrelevante de los Dominicos. Sin sorpresa, este desacuerdo culminaría con el famoso juicio contra Galileo y su arresto domiciliario hasta el día de su muerte.

Sobre esta base, el escéptico diagnostica un problema serio para TCE. Si la condición de integración cognitiva en COGA-WEAK se interpreta bajo INTEGRACIÓN DE GRANO FINO, la satisfacción de DISCRIMINACIÓN implica la neutralidad del criterio de habilidad cognitiva o no. Si la implica, entonces TCE no salvaría AGENCIA* en casos donde METAJUSTIFICACIÓN no se satisface. Esto sería suficiente para establecer [Premisa 3]. Si no implica la neutralidad del criterio, entonces los defensores de TCE deben argumentar por qué esta consecuencia relativista no es devastadora de su teoría. En otras palabras, los defensores de TCE deben elegir entre el escepticismo y el relativismo. Además, si optan por el segundo, deben justificar por qué relativizar TCE sería preferible a aceptar la derrota ante el escéptico. Después de todo, Pritchard (2009, 2015) piensa que la teorización epistemológica no debe conducir al escepticismo ni al relativismo.

Este diagnóstico pone sobre la mesa un dilema frente al cual los defensores de TCE deben tomar una decisión. De un lado, TCE debe abandonarse si renuncian a sus principios de primer orden. Tanto EXTERNISMO ACTIVO como FIABILISMO DE VIRTUDES son principios *constitutivos* de esta teoría. De otro lado, TCE puede conservarse si renuncian a sus principios de segundo orden. El escéptico ha mostrado que, bajo INTEGRACIÓN DE GRANO FINO, la interpretación de la condición de integración cognitiva en COGA-WEAK tiene implicaciones relativistas. Esto sugiere otro lugar en el cual la teoría podría ajustarse. ¿Cuál de los dos caminos conviene seguir?

Mi propuesta es revisar la metaepistemología de TCE por dos razones. La primera es prudencial: al igual que sucede con las teorías científicas, nuestras epistemologías (naturalistas) con cierto grado de fertilidad teórica deben tener la oportunidad de seguir con vida antes de desecharlas definitivamente. TCE ha demostrado ser una teoría fecunda en términos explicativos.[13] La segunda es tentadora: el escepticismo, pero no el relativismo, sí es inaceptable en epistemología. Si se opta por esta vía, los defensores de TCE deben mostrar un par de puntos. En primer lugar, que TCE puede conservar sus principios de primer orden si se acepta el relativismo epistémico. En segundo lugar, que esta relativización de TCE no implica el escepticismo, de modo que sí hay buenas razones para preferir al relativismo sobre el escepticismo.

Mi tesis positiva es que el relativismo epistémico pavimenta el camino para que TCE pueda dar una explicación completamente satisfactoria de la agencia epistémica sin caer en el escepticismo. Si los defensores de TCE conceden que INTEGRACIÓN DE GRANO FINO es la interpretación adecuada de la condición de integración cognitiva en COGA-WEAK (porque, de otra manera, TCE enfrenta el problema de la generalidad al interpretar dicha condición bajo INTEGRACIÓN DE GRANO GRUESO), el escéptico presenta el siguiente argumento para establecer [Premisa 3] de ARGUMENTO GENERAL:

(ARGUMENTO ESPECÍFICO)

(P1) Si hay DISCRIMINACIÓN, entonces hay METAJUSTIFICACIÓN;

(P2) no hay METAJUSTIFICACIÓN;

(C) por lo tanto, no hay DISCRIMINACIÓN.

---

[13] Por ejemplo, TCE permitiría conciliar la epistemología social «analítica» de Goldman con la epistemología social «naturalista» de Fuller (Palermos & Pritchard, 2013), cerrar la brecha entre la epistemología tradicional y la filosofía de la ciencia (Palermos, 2015), mostrar la naturaleza tanto individual como colectiva del conocimiento (Palermos, 2011), y dar cuenta del conocimiento producido por grupos de personas a partir de una noción de agente grupal epistémico (Palermos & Pritchard, 2016).

https://doi.org/10.1387/theoria.26981

(P1) sería un compromiso de segundo orden de TCE, el cual establece una relación de condición necesaria entre los conceptos epistémicos fundamentales de «discriminación» y «metajustificación.» Sobre la base de esta premisa, el agente exhibe AGENCIA* en la medida en que cuenta con un criterio independiente para identificar el proceso de formación de creencias que es conducente al conocimiento. Sin este criterio independiente de habilidad cognitiva, el agente no podría discriminar el proceso relevante.

(P2) asegura que METAJUSTIFICACIÓN no se cumple. Esta premisa se apoya en al menos tres condiciones a partir de las cuales se ha caracterizado el relativismo epistémico recientemente (Ashton, 2020; Kusch, 2016, 2019, 2021). Empleando la terminología de TCE, la formulación de estas condiciones tendría esta forma:

> (RELATIVISMO EPISTÉMICO MÍNIMO)
>
> (DEPENDENCIA) Las creencias tienen un estatus epistémico positivo sólo relativo a procesos de formación que son cognitivamente hábiles. Conjuntos de estos procesos componen caracteres cognitivos de los agentes.
>
> (PLURALIDAD) Hay (ha habido o podría haber) más de un conjunto de procesos que son cognitivamente hábiles para formar creencias.
>
> (NO-METAJUSTIFICACIÓN) Cuando las creencias y/o los procesos entran en conflicto, no existen (al menos en casos importantes) formas de arbitraje de procesos como cognitivamente hábiles que sean neutrales e independientes respecto de dichos procesos.

Relativismo y absolutismo son posiciones rivales en metaepistemología (Bland, 2018; Bloor, 2011; Boghossian, 2006; Siedel, 2014). Además, el absolutismo también es consistente con DEPENDENCIA y PLURALIDAD. No parece problemático aceptar que el estatus de las creencias responsables epistémicamente es relativo a la habilidad cognitiva (según FIABILISMO DE VIRTUDES) ni que lo que cuenta como habilidad cognitiva varía según los caracteres cognitivos de los agentes a través de los contextos de sus prácticas epistémicas (según INTEGRACIÓN DE GRANO FINO). El desacuerdo reside más bien en aceptar o rechazar NO-METAJUSTIFICACIÓN. Mientras que el absolutista sostiene que sí existen criterios independientes para discriminar procesos como cognitivamente hábiles, el relativista niega que dichos criterios sean independientes. El absolutista rechaza NO-METAJUSTIFICACIÓN porque interpreta esta condición con la implicación desafortunada de la «igual validez» endilgada comúnmente al relativismo (Bland, 2018; Boghossian, 2006):

> (IGUALDAD) Todos los procesos son cognitivamente hábiles en igual medida.

Esta tesis resulta inadmisible porque conduce directamente al escepticismo. Si todos los procesos son igualmente hábiles en términos epistémicos, entonces ningún proceso sería hábil y ninguna creencia tendría un estatus epistémico positivo (Boghossian, 2006; Pritchard, 2009). De esto se sigue que ningún agente recibiría crédito epistémico (ni sabría que *p*) porque ningún proceso conduciría al agente a saber que *p*. Como esta es la conclusión de ARGUMENTO GENERAL, el absolutista encuentra una razón para afirmar que DISCRIMINACIÓN requiere que el estándar de METAJUSTIFICACIÓN se satisfaga. Este es su argumento:

1. NO-METAJUSTIFICACIÓN implica IGUALDAD.
2. IGUALDAD implica que los procesos no son cognitivamente hábiles y, por lo tanto, que *S* no merece crédito epistémico.
3. En consecuencia, NO-METAJUSTIFICACIÓN implica que los procesos no son cognitivamente hábiles y, por lo tanto, que *S* no merece crédito epistémico.

A pesar de esto, el relativista tiene una interpretación diferente de NO-METAJUSTIFICACIÓN. Su posición es que esta condición sólo establece que no hay criterios neutrales para arbitrar procesos de formación de creencias, lo cual no significa ni implica (al menos no de forma obvia) IGUALDAD (Ashton, 2019, 2020; Barnes & Bloor, 1982; Field, 2009; Kusch 2019, 2021):

> (NO-NEUTRALIDAD) La discriminación de procesos como cognitivamente hábiles nunca es neutral. La discriminación siempre se hace a partir de un proceso o conjunto de procesos (esto es, a partir de un carácter cognitivo) específico.

El relativista evita la consecuencia escéptica que el absolutista trae a cuento de dos formas. De un lado, abriendo la puerta al respaldo epistémico *relativo*. Según esta idea, el estatus epistémico de las creencias es relativo a la habilidad cognitiva y no hay una manera de determinar la habilidad cognitiva que no sea apelando a los caracteres cognitivos mismos. Los procesos de formación de creencias sólo pueden ser *relativamente hábiles*. Así, un agente merecería adjudicación de crédito epistémico (y de conocimiento) al conceder estándares epistémicos menos demandantes. En este caso, «demandamos menos de nuestros estándares epistémicos si permitimos que varíen con comunidades y si aceptamos que no pueden arbitrarse de manera neutral» (Kusch, 2017, p. 4693). De otro lado, el relativista señala que IGUALDAD es correcta sólo si se presupone una perspectiva más allá de los caracteres cognitivos a partir de la cual discriminar los procesos de formación de creencias. El problema con esto es que el relativista niega la existencia de dicha perspectiva o la posibilidad misma de asumirla. Que NO-METAJUSTIFICACIÓN implique IGUALDAD y que de esta se siga el escepticismo es únicamente «lo que el relativismo parece ser *desde el punto de vista absolutista*» (Bloor, 2011, p. 440; ver también Field, 2009, p. 256). En este caso, el absolutista comete petición de principio al vincular NO-METAJUSTIFICACIÓN con IGUALDAD. El relativista no está obligado a conceder que el respaldo epistémico «real» sea el respaldo epistémico «absoluto» ni que haya una perspectiva enteramente independiente de los caracteres cognitivos para discriminar procesos en términos de su habilidad cognitiva. El relativismo niega justamente lo que el absolutismo está requiriendo. Así las cosas, este es el argumento del relativista de por qué es innecesario rechazar NO-METAJUSTIFICACIÓN:

1. NO-METAJUSTIFICACIÓN implica NO-NEUTRALIDAD.
2. NO-NEUTRALIDAD implica que los procesos sólo son relativamente hábiles y, por lo tanto, que *S* sólo merece relativo crédito epistémico.
3. Por lo tanto, NO-METAJUSTIFICACIÓN implica que los procesos sólo son relativamente hábiles y, por lo tanto, que *S* sólo merece relativo crédito epistémico.

Hasta que el absolutista ofrezca una razón independiente para negar NO-METAJUSTIFICACIÓN, el relativista recomendará atacar ARGUMENTO ESPECÍFICO en otra dirección. Recordemos el argumento: (P1) Si hay DISCRIMINACIÓN, entonces hay METAJUSTIFICACIÓN; (P2) no hay METAJUSTIFICACIÓN; (C) por lo tanto, no hay DISCRIMINACIÓN. El problema con ARGUMENTO ESPECIFICO ya no tendría que ver con (P2), sino con (P1). El relativista toma (P1) como una *presuposición absolutista* que el escéptico atribuye a TCE como uno de sus compromisos implícitos de segundo orden. Según esta presuposición, si no hay un criterio independiente de habilidad cognitiva para arbitrar procesos de formación de creencias, el agente no puede discriminar el proceso que es conducente al conocimiento. Por lo tanto, el escepticismo aparece como una consecuencia inevitable. Los compromisos de segundo orden de TCE se convierten en la puerta de entrada al escepticismo.

Para bloquear ARGUMENTO ESPECÍFICO, el relativista propone entonces una solución bastante simple: sustituir (P1) por su negación y adoptarla como una *presuposición relativista*. Bajo esta nueva presuposición, ~(P1): *hay discriminación y no hay metajustificación*. Como se puede apreciar, esta maniobra es suficiente para evitar el paso de las premisas a la conclusión a través de la regla del *modus tollens*. Si la conclusión de ARGUMENTO ESPECÍFICO no se obtiene, entonces [Premisa 3] de ARGUMENTO GENERAL tampoco se obtiene, con lo cual la objeción del escéptico queda sin respaldo en última instancia.

¿Qué se logra a partir de esto? Un compromiso de segundo orden con el relativismo epistémico permite a TCE tener una visión más laxa acerca de las condiciones bajo las cuales DISCRIMINACIÓN se satisface a fin de que el agente exhiba AGENCIA*. El relativismo hace posible que el agente discrimine el proceso que cuenta como conducente al conocimiento, aunque no lo haga como el criterio de METAJUSTIFICACIÓN demanda; es decir, aunque no lo haga de manera neutral. Los agentes pueden apelar a su propio carácter cognitivo para evaluar procesos y creencias que dependen del mismo, así como para evaluar procesos y creencias de otros agentes. También pueden integrar otros procesos de formación de creencias a su carácter cognitivo para revisar sus compromisos doxásticos y los de otros agentes. Esto deja a los agentes en posición de responder apropiadamente a sus compromisos doxásticos en situaciones donde encuentran razones que ponen en cuestión su respaldo epistémico *prima facie*. Desde una perspectiva relativista, Galileo y los matemáticos del Colegio Romano fueron responsables epistémicamente y merecedores de relativo crédito epistémico. En términos generales, el relativismo es una opción en metaepistemología que evita las consecuencias escépticas a las que el absolutismo conduce. El absolutista pone una vara demasiado alta para que un agente sea responsable epistémicamente ante razones derrotantes.

Para resumir: si TCE abraza un compromiso de segundo orden con el absolutismo epistémico, ARGUMENTO GENERAL sí permite concluir que TCE no explica AGENCIA*. Sin embargo, si TCE abraza un compromiso de segundo orden con el relativismo epistémico, ARGUMENTO GENERAL ya no funciona para establecer esa conclusión. Por lo tanto, TCE puede explicar AGENCIA* si modifica sus compromisos metaepistemológicos implícitos. Por supuesto, adoptar el relativismo no es necesariamente la única alternativa que TCE podría tomar para tratar con el desafío escéptico que enfrenta. Sin embargo, dado el conjunto de razones escépticas que he considerado en esta ocasión, aceptar el relativismo epistémico es una buena manera en la cual TCE puede lidiar con el problema. A diferencia de su rival, el relativismo no lleva al escepticismo y puede incluso cerrarle la puerta.

6. Conclusiones

En este texto evalué a TCE como explicación de la agencia epistémica en relación con agentes que adquieren conocimiento a partir de procesos de formación de creencias realizados por sistemas cognitivos extendidos y distribuidos. Me concentré en discutir la manera como los principios de primer orden de TCE darían cuenta del agente que es sensible a razones derrotantes de su estatus epistémico y que responde a ellas apropiadamente.

Primero caractericé y defendí un resultado pesimista de esta evaluación, a saber, que TCE no puede explicar AGENCIA* sin caer en el escepticismo. Sin embargo, para salvar a TCE en lugar de abandonarla, seguí luego un camino optimista que consiste en modificar la metaepistemología de la teoría. Así, defendí que, si TCE acepta el relativismo epistémico como principio de segundo orden, entonces la teoría puede explicar AGENCIA* sin caer en el escepticismo.

Para justificar mi tesis negativa, elaboré dos argumentos escépticos, uno general y otro específico. ARGUMENTO GENERAL apoya la afirmación de que el agente epistémico es incapaz de exhibir AGENCIA* porque no satisface DISCRIMINACIÓN. En consecuencia, el agente no merece crédito epistémico ni atribución de conocimiento. Además, DISCRIMINACIÓN no se cumple cuando la condición de integración cognitiva en COGA-WEAK se interpreta bajo INTEGRACIÓN DE GRANO GRUESO, pues TCE enfrenta una versión concreta del problema de la generalidad bajo esta interpretación. Pero, cuando la condición de integración cognitiva en COGA-WEAK se interpreta bajo INTEGRACIÓN DE GRANO FINO, TCE tiene consecuencias relativistas. Sobre esta base, ARGUMENTO ESPECÍFICO muestra de forma interesante que la fuente del problema con TCE reside en su compromiso implícito con el absolutismo epistémico como principio de segundo orden. Según este argumento, DISCRIMINACIÓN no se cumple porque el estándar de METAJUSTIFICACIÓN no se satisface y porque TCE estaría comprometida con la presuposición del absolutismo epistémico: *si hay discriminación, entonces hay metajustificación*.

Esto no significa que el escepticismo deba aceptarse y, por lo tanto, que TCE deba rechazarse. Para justificar mi tesis positiva, argumenté que TCE puede evitar el escepticismo si acepta la presuposición del relativismo epistémico: *hay discriminación y no hay metajustificación*. Esta maniobra implica relajar los estándares bajo los cuales DISCRIMINACIÓN se cumple a fin de que el agente sea responsable epistémicamente de acuerdo con AGENCIA*, pero esto no tiene por qué tener las mismas consecuencias escépticas. Por supuesto, el relativismo sería tan inaceptable como el escepticismo sólo si NO-METAJUSTIFICACIÓN implica IGUALDAD, pues aceptar la segunda termina llevando al escepticismo. Sin embargo, aunque el relativista admite que IGUALDAD implica el escepticismo, no está convencido de que NO-METAJUSTIFICACIÓN implique IGUALDAD. Para el relativista, NO-METAJUSTIFICACIÓN implica NO-NEUTRALIDAD y de esta no se sigue el escepticismo. Esto es una buena razón para pensar que el relativismo es preferible al escepticismo, por lo menos al evaluar el poder explicativo de TCE.

Mi propuesta de relativizar TCE, para tener una explicación completamente satisfactoria de la agencia epistémica, puede que no resulte atractiva para los defensores principales de esta teoría. Después de todo, el relativismo sería absolutamente inadmisible para un epistemólogo, o la maniobra podría tomarse como la introducción de una hipótesis *ad hoc*, o intentar ajustar otras partes de la teoría sería más promisorio, o simplemente no habría por qué aceptar mis argumentos escépticos ni los relativistas. Todo esto podría llegar a ser correcto presentando contraargumentos adicionales. Sin embargo, las razones aquí expuestas justifican la afirmación de que TCE corre el riesgo de abrir la puerta al escepticismo y al relativismo debido a la combinación sofisticada de sus principios explicativos.[14] Por mi parte, considero que abrazar el relativismo epistémico para ajustar TCE y no sepultarla es una buena decisión después de evaluar la teoría.

Agradecimientos

Quiero agradecer a Santiago Echeverri y a dos revisores anónimos de THEORIA por sus valiosos comentarios a versiones anteriores de este artículo. También agradezco a Ángeles Eraña, Maite Ezcurdia (†), Martin Kusch y Santiago Arango por las estimulantes conversaciones que tuvimos sobre estos temas. Este trabajo recibió apoyos del proyecto de investigación PAPIIT (#IA400124) "Escepticismo y Racionalidad" adscrito al Instituto de Investigaciones Filosóficas de la UNAM.

---

[14] También pueden objetarme que estos argumentos no atacan exclusivamente a TCE, pues son en principio generalizables a cualquier forma de fiabilismo de virtudes que defienda una visión condicional de agencia epistémica. En un grado menos general, mis críticas serían especialmente difíciles para el conjunto de epistemologías de virtudes que toman la condición de integración cognitiva en COGA-weak en su sentido más amplio: integración cognitiva por interacción cooperativa entre el proceso y otros aspectos del sistema cognitivo del agente (Greco, 2010, p. 152). La razón es que esta condición permite considerar como habilidades cognitivas a procesos internos realizados exclusivamente por facultades orgánicas y no sólo a procesos extendidos y distribuidos. Además, incluso procesos dentro del organismo (al igual que procesos extendidos y distribuidos) pueden modelarse con la teoría de sistemas dinámicos en términos de interacciones con base en bucles de retroalimentación. Por ejemplo, Palermos (2014b) sugiere que procesos cognitivos cerebrales de «orden superior», como lo son procesos de formación de creencias, «sólo pueden surgir a partir de interacciones mutuas entre diferentes componentes procesuales» del sistema cerebral (p. 32).

Este aspecto, sin embargo, no resta valor a mis argumentos contra TCE. Mi evaluación *concreta* de esta teoría puede tomarse como un *caso de estudio* que articula un argumento epistemológico *abstracto* que puede *generalizarse* a cualquier teoría de la agencia epistémica que adopte el fiabilismo de virtudes. Mostrar la aplicación general de mis críticas está obviamente por fuera de los objetivos de este artículo. Esta estrategia argumentativa, que desarrolla argumentos abstractos de aplicabilidad general a partir del análisis iterativo de posiciones concretas, es común en historia y filosofía de la ciencia (Chang, 2011) y tampoco es nueva en epistemología analítica (ej. Carter & McKenna, 2021; Williamson, 2022). Agradezco a un revisor anónimo de THEORIA por hacerme notar y justificar este punto tan importante.

REFERENCES

Alston, W. (1982). Religious experience and religious belief. *Noûs*, 16(1), 3-12. https://doi.org/10.2307/2215404

Ashton, N. (2019). Rethinking epistemic relativism. *Metaphilosophy*, 50(5), 587-607. https://doi.org/10.1111/meta.12389

Ashton, N. (2020). Relativising epistemic advantage. In M. Kusch (Ed.). *The Routledge handbook of philosophy of relativism* (329-338). Routledge. https://doi.org/10.4324/9781351052306

Baghramian, M., & Coliva, A. (2019). *Relativism*. Routledge. https://doi.org/10.4324/9781315145190

Barnes, B., & Bloor, D. (1982). Relativism, rationalism and the sociology of knowledge. In M. Hollis & S. Luke (Eds.). *Rationality and relativism* (21-47). Blackwell.

Biagioli, M. (1993). *Galileo, courtier: The practice of science in the culture of absolutism*. University of Chicago Press. https://doi.org/10.7208/chicago/9780226218977.001.0001

Biagioli, M. (2006). *Galileo's instruments of credit: Telescopes, images, secrecy*. University of Chicago Press. https://doi.org/10.7208/chicago/9780226218977.001.0001

Biancani, G. (1620). *Sphaera mundi, seu cosmographia demonstrativa, ac facili methodo tradita*. Apud haeredes G. Bellagamba.

Bland, S. (2018). *Epistemic relativism and scepticism*. Palgrave Macmillan.

Bloor, D. (2011). Relativism and the sociology of knowledge. In S. Hales (Ed.). *A companion to relativism* (433-455). Blackwell.

Boghossian, P. (2006). *Fear of knowledge*. Clarendon Press.

Brandom, R. (1998). Insights and blindspots of reliabilism. *The Monist*, 81(3), 371-392. https://doi.org/10.5840/monist199881317

Carolino, L. M. (2008). The making of a Tychonic cosmology: Cristoforo Borri and the development of Tycho Brahe's astronomical system. *Journal for the History of Astronomy*, 39(3), 313–344. https://doi.org/10.1177/002182860803900303

Carter, J., & McKenna, R. (2021). Absolutism, relativism and metaepistemology. *Erkenntnis*, 86, 1139–1159. https://doi.org/10.1007/s10670-019-00147-w

Cartwright, N. (2020). X-Why trust science? Reliability, particularity and the tangle of science. *Proceedings of the Aristotelian Society*, 0(0), 1-16. https://doi.org/10.1093/arisoc/aoaa015

Chang, H. (2011). Beyond case-studies: History as philosophy. In S. Mauskopf & T. Schmaltz (Eds.). *Integrating history and philosophy of science: Problems and prospects* (109–124). Springer. https://doi.org/10.1007/978-94-007-1745-9

Clark, A. (2007). Curing cognitive hiccups: A defence of the extended mind. *The Journal of Philosophy*, 104, 163–192. https://www.jstor.org/stable/20620012

Clark, A. (2008). *Supersizing the mind*. Oxford University Press.

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, 58(1), 7–19. https://www.jstor.org/stable/3328150

Code, L. (1987). *Epistemic responsibility*. University Press of New England.

Conee, E., & Feldman, R. (1998). The generality problem for reliabilism. *Philosophical Studies*, 89(1), 1-29. https://doi.org/10.1023/A:1004243308503

Dear, P. (1987). Jesuit mathematical science and the reconstitution of experience in the early seventeenth century. *Studies in History and Philosophy of Science Part A, 18*(2), 133–175. https://doi.org/10.1016/0039-3681(87)90016-1

Feldhay, R. (1995). *Galileo and the church: Political inquisition or critical dialogue?* Cambridge University Press.

Fernandez Cruz, A. L., Arango-Muñoz, S., & Volz, K. G. (2016). Oops, scratch that! Monitoring one's own errors during mental calculation. *Cognition*, 146(C):110-120.

Field, H., (2009). Epistemology without metaphysics. *Philosophical Studies*, 143, 249–90. https://doi.org/10.1007/s11098-009-9338-1

Galilei, G. (1610/1989). *Sidereus nuncius, or the sidereal messenger*. (A. Van Helden, Trans.). University of Chicago Press.

Galilei, G. (1890–1909). *Le opere di Galileo Galilei*. (A. Favaro, Ed., 20 vols.). Barbera.

Garfinkel, A. (1981). *Forms of explanation*. Yale University Press.

Goldman, A. (1986). *Epistemology and cognition*. Harvard University Press.

Goldman, A. (2021). A different solution to the generality problem for process reliabilism. *Philosophical Topics*, 49(2), 105-111. https://doi.org/10.5840/philtopics202149217

Gorman, M. J. (2020). *The scientific counter-revolution: The Jesuits and the invention of modern science*. Bloomsbury Academic.

Greco, J. (1999). Agent reliabilism. In J. Tomberlin (Ed.). *Philosophical perspectives 13: Epistemology* (273–296). Ridgeview Press.

Greco, J. (2004). Knowledge as credit for true belief. In M. DePaul & L. Zagzebski (Eds.). *Intellectual virtue: Perspectives from ethics and epistemology* (111-134). Oxford University Press.

Greco, J. (2007). The nature of ability and the purpose of knowledge. *Philosophical Issues*, 17, 57–69. https://doi.org/10.1111/j.1533-6077.2007.00122.x

Greco, J. (2010). *Achieving knowledge: A virtue-theoretic account of epistemic normativity*. Cambridge University Press.

Hales, S. (2014). Motivations for relativism as a solution to disagreements. *Philosophy*, 89(1), 63–82. https://doi.org/10.1017/S003181911300051X

Hollan, J., Hutchins, E., & Kirsh, D. (2000). Distributed cognition: toward a new foundation for human-computer interaction research. *ACM Trans. Comput. Hum. Interact*, 7(2), 174–196. https://doi.org/10.1145/353485.353487

Kinzel, K., & Kusch, M. (2018). De-idealizing disagreement, rethinking relativism. *International Journal of Philosophical Studies*, 26(1), 40-71. https://doi.org/10.1080/09672559.2017.1411011

Kornblith, H. (2016). Epistemic agency. In M. Fernández (Ed.). *Performance epistemology: Foundations and applications* (1-22). Oxford University Press.

Kusch, M. (2002). *Knowledge by agreement*. Oxford University Press.

Kusch, M. (2009). Boghossian on epistemological and moral relativism: A critique. Unpublished Manuscript.

Kusch, M. (2016). Wittgenstein's on certainty and relativism. In H. A. Wiltsche & S. Rinofner-Kreidl (Eds.). *Analytic and continental philosophy: Methods and perspectives: Proceedings of the 37th international Wittgenstein symposium* (29-46). De Gruyter.

Kusch, M. (2017). Epistemic relativism, scepticism, pluralism. *Synthese*, 194(12), 4687-4703. https://doi.org/10.1007/s11229-016-1041-0

Kusch, M. (2019). Relativist stances, virtues and vices. *Aristotelian Society Supplementary Volume*, 93(1), 271-291. https://doi.org/10.1093/arisup/akz007

Kusch, M. (2021). *Relativism in the philosophy of science*. Cambridge University Press.

Lattis, J. M. (1994). *Between Copernicus and Galileo: Christoph Clavius and the collapse of Ptolemaic cosmology*. University of Chicago Press.

Lipton, P. (1990). Contrastive explanations. In D. Knowles (Ed.). *Explanation and its limits* (247–266). Cambridge University Press.

Palermos, S. O. (2011). *Extending cognition in epistemology: Towards an individualistic social epistemology.* Unpublished Doctoral Thesis. The University of Edinburgh.

Palermos, S. O. (2014a). Knowledge and cognitive integration. *Synthese*, 191(8), 1931–1951. https://doi.org/10.1007/s11229-013-0383-0

Palermos, S. O. (2014b). Loops, constitution and cognitive extension. *Cognitive Systems Research*, 27, 25-41. https://doi.org/10.1016/j.cogsys.2013.04.002

Palermos, S. O. (2015). Active externalism, virtue reliabilism and scientific knowledge. *Synthese*, 192(9), 2955-2986. https://doi.org/10.1007/s11229-015-0695-3

Palermos, O. & Pritchard, D. (2013). Extended knowledge and social epistemology. *Social Epistemology Review and Reply Collective*, 2(8), 105-120.

Palermos, O. & Pritchard, D. (2016). The distribution of epistemic agency. In P. Reider (Ed.). *Social epistemology and epistemic agency: Decentralizing epistemic agency* (109-126). Rowman & Littlefield.

Pollock J., & Cruz, J. (1999). *Contemporary theories of knowledge*. Rowman and Littlefield.

Pritchard, D. (2009). Defusing epistemic relativism. *Synthese*, 166(2), 397–412. https://doi.org/10.1007/s11229-007-9278-2

Pritchard, D. (2010). Cognitive ability and the extended cognition thesis. *Synthese*, 175, 133-151. https://doi.org/10.1007/s11229-010-9738-y

Pritchard, D. (2015). *Epistemic angst*. Oxford University Press. https://doi.org/10.1515/9781400873913

Scheiner, C. (1612). *De maculis solaribus et stellis circa Iovem errantibus, accuratior disquisition ad Marcum Velserum*. Augsburg Marc Welser.

Schwartz, B. L., & Jemstedt, A. (2021). The role of fluency and dysfluency in metacognitive experiences. In D. Moraitou and P. Metallidou (Eds.). *Trends and prospects in metacognition research across the life span* (25-40). Springer. https://doi.org/10.1007/978-3-030-51673-4_2

Seidel, M. (2014). *Epistemic relativism: A constructive critique*. Palgrave MacMillan.

Stanford, P. K. (2006). *Exceeding our grasp: Science, history, and the problem of unconceived alternatives*. Oxford University Press.

Sudduth, M. (2008). Defeaters in epistemology. *Internet Encyclopedia of Philosophy*. https://iep.utm.edu/defeaters-in-epistemology van Helden, A. (1996). Galileo and Scheiner on sunspots: A case study in the visual language of astronomy. *Proceedings of the American Philosophical Society,* 140(3), 358–396. http://www.jstor.org/stable/987314

Westfall, R. S. (1985). Science and patronage: Galileo and the telescope. *Isis*, 76(1), 11–30. https://doi.org/10.1086/353735

Westfall, R. S. (1989). Galileo and the Jesuits. In R. S. Woolhouse (Ed.). *Metaphysics and philosophy of science in the seventeenth and eighteenth centuries* (45–72). Springer. https://doi.org/10.1007/978-94-017-0361-1_2

Williams, M. (2015). What's so special about human knowledge? *Episteme*, 12(2), 249-268. https://doi.org/10.1017/epi.2015.14

Williamson, T. (2022). *The philosophy of philosophy (Second edition).* Wiley Blackwell. https://doi.org/10.1002/9781119616702

Dubian Cañas es Doctor en Filosofía de la Ciencia de la Universidad Nacional Autónoma de México (UNAM). Ha sido académico visitante en el Departamento de Filosofía de la Universidad de Viena y en el Departamento de Historia y Filosofía de la Ciencia de la Universidad de Cambridge. Se especializa en historia y filosofía de la ciencia, epistemología y metafilosofía.

Dirección: Universidad Nacional Autónoma de México, Edificio E - Unidad de Posgrado, Circuito de Posgrados, Ciudad Universitaria, 04510, Ciudad de México, México. E-mail: dubiandrescanas@gmail.com – ORCID**:** 0000-0002-2134-1045

Theoria, 2025, 40(2), 154-171
https://doi.org/10.1387/theoria.27141

THEORIA
ISSN 0495-4548 – eISSN 2171-679X

The metaphysics of symmetries: a map

(La metafísica de las simetrías: un mapa)

Cristian López*
University of Lausanne
University of Buenos Aires / CONICET
https://orcid.org/0000-0002-2883-4037

Keywords

Keywords: Symmetries
Metaphysics
Fundamentalism
Physics
Realism
Anti-realism

ABSTRACT: My main aim in this paper is to lay out a metaphysical map of the possible metaphysical attitudes that philosophers may hold with respect to physical symmetries. Even though it is customary to distinguish between realism vs. anti-realism, I show that the metaphysical landscape looks a bit more complex. To show this, I differentiate at least nine different views that have been held, or might be held, concerning which place physical symmetries occupy in one's ontology.

Palabras clave

Metafísica
Fundamentalismo
Física
Realismo
Anti-realismo

RESUMEN: Mi objetivo en este artículo es desarrollar un mapa metafísico de las posibles actitudes metafísicas que los filósofos pueden adoptar con respecto a las simetrías en física. Aunque es usual distinguir entre realismo y anti-realismo, en este artículo muestro que el panorama metafísico es un poco más complejo que eso. Para ello, distingo al menos nueve posiciones diferentes que han sido adoptadas, o podrían ser adoptadas, en relación al lugar que las simetrías pueden ocupar en la ontología que uno conciba.

1. *Introduction*

Modern physics has been built upon symmetries. From Galileo's ship to fermion-boson symmetry, different symmetry-based arguments have guided not only theory construction but also empirical research. Examples abound. Consider non-relativistic quantum mechanics. In standard presentations of the theory, Galilean invariances permit us to identify abstract operators with measurable dynamical variables, which introduces physical content into the theory (see Ballentine, 1998, sections 3.3-3.4). Thus, Galilean invariances play an active role in theory construction. Or think of the discovery of the omega-negative baryon (Ω^-), which has been a milestone in particle physics. In this case, the "Eightfold

* ***Correspondence to:*** Cristian López. University of Lausanne, Anthropole, Section de Philosophie, Of. 5155, CH-1015, Switzerland – cristian.lopez@unil.ch – https://orcid.org/0000-0002-2883-4037

How to cite: López, Cristian (2025). «The metaphysics of symmetries: a map»; *Theoria. An International Journal for Theory, History and Foundations of Science,* 40(2), 154-171. (https://doi.org/10.1387/theoria.27141).

Received: 01 December, 2024; Final version: 18 June, 2025.
ISSN 0495-4548 - eISSN 2171-679X / © 2025 UPV/EHU Press

Way" (or SU(3)-symmetry) has strongly guided empirical research during the 1960s, culminating in February 1964 with the discovery of the particle at the Brookhaven AGS.

In the last decades, philosophers have increasingly drawn their attention to the pervasive occurrences of symmetry claims in physics. Why do symmetries work so stupendously well? Do they reveal profound features of the world? Or are they a clever way to represent the natural world? The interest has been multiplefold, ranging from methodological to mathematical as well as ontological issues. What concerns me here is the metaphysical facet of the debate. On this, the attention was mainly drawn towards space-time symmetries (since they may reveal the nature of space-time), local gauge, and permutation symmetries (since they may reveal the aspects of the world's fundamental ontology). Symmetry claims, then, became not only crucial premises to infer the fundamental ontology (see North, 2009, 2021; see López, 2023a for criticisms) but they were also meant to refer to aspects of the fundamental ontology, reifying the concept (see Heisenberg, 1975; Weinberg, 1987; Baker, 2010; French, 2014, Schroeren, 2020, among others; see López, 2025 for criticisms). In opposition to these views, various forms of anti-realist or epistemic views have also been defended, gravitating around notions like "heuristic value", "epistemic principles", or even "prescindible postulates".

The main aim of this paper is to clarify the metaphysical landscape by laying out a map of the different metaphysical attitudes that may *possibly* be held on physical symmetries, whether or not they are represented in the literature. My aim is not to defend or attack any of these views, but to show their relations, differences, and common bases where they exist. I will refer to existing literature defending or attacking the different views when pertinent. The idea of laying out a metaphysical map of symmetries is to show the different metaphysical attitudes that address the problem of which place symmetries occupy in one's ontology; and while the division between realist–anti–realist views might be operationally useful, it falls short of capturing all the nuances that diverging metaphysical attitudes may have. I start off by distinguishing between *Symmetry Dispensabilism* and *Symmetry Indispensabilism*. Within the dispensability camp, at least four different views can be identified: the *No Symmetry View, Symmetry Agentialism, Symmetry Conventionalism, and Symmetry Expressivism*. To a good extent, these views represent anti-realist attitudes in different ways. In turn, within the indispensability camp, five positions can be distinguished: *Symmetry Realism, Symmetry Fundamentalism, Symmetry Inferentialism, Symmetry Epistemicism and Symmetry Normativism*. Even though many views represent realist attitudes, there is no linear association since some views, such as Symmetry Epistemicism and Symmetry Normativism, may well be viewed as anti-realist.

2. The Philosophical Problem(s) of Symmetries

Although physical symmetries come in many flavors and shapes (internal vs. external, local vs. global, theoretical vs. observational, geometrical vs. dynamical, and so on), all of them are, for the most part, formal notions that apply to mathematical structures. From an abstract viewpoint, symmetries are transformations that keep some relevant structure unaltered. In physics, most mathematical structures of interest are sets of differential equations that relate to other mathematical structures (e.g., topological and differential spaces). In consequence, physical symmetries are transformations that preserve the space of solutions of such sets of differential equations. In this precise sense, physical symmetries are said to be structure-preserving functions that map solutions to solutions.

This definition is, however, insufficient to obtain a *physical* symmetry. If this were so, it would always be possible to define a transformation that maps solutions to solutions, augmenting the symmetries of a theory on demand. Thus, as Gordon Belot mentions (2013), symmetries are hard to come by, so their physical definition should not be too liberal to capture that fact. This, in general, amounts to imposing further constraints on the abstract definition (see Read and Møller-Nielsen, 2020, for further constraints). Some of them may be purely formal —e.g., for Lie transformations, they must be continuous or smooth; for classical symmetries, the infinitesimal generators must only depend on the independent and dependent variables of the theory, etc. Others may be physical —e.g., some symmetries are required to not only preserve the geometrical structure of the phase space, but also specific objects (e.g., the Hamiltonian). And others may be interpretative —e.g., physical symmetries are required to preserve the observational content of a physical theory (see, for instance, Dasgupta, 2016), or to identify surplus structure (see Redhead, 1975; Dewar, 2019. See Hall and Murgueitio Ramirez, 2024, for discussion and an overview).

There are, of course, plenty of problems around the theoretical status and role of symmetries in physics. Christopher Martin (2002) has, for instance, centered on the role of local gauge symmetries in relation to other theoretical principles (e.g., renormalizability). Michael Redhead has tackled the problem of the role of symmetries for assessing redundant structures and inter-theoretical relations (Redhead, 1975). Katherine Brading and Elena Castellani have focused on classifying symmetry principles and symmetry arguments in different theories (Brading and Castellani, 2003, 2007). Others have focused on the methodological role of space-time symmetries in elucidating problems around the debate between relationalism and substantivalism (Earman 1989). Lately, the focus has been put on the role of symmetries in the interpretation of physical theories, distinguishing between interpretational and motivational views (Møller-Nielsen, 2018; Luc, 2022). In a similar line, many have concerned the role of symmetries in the invariance principle (Dasgupta, 2016). All this shows to what extent symmetries provide a fertile terrain for philosophical debate. I, however, focus here on the *metaphysical* facet of the problem, concerning the place that symmetries occupy in the ontology.

In metaphysics, this is "the problem of accommodating symmetries" (see French, 2014). Modern physics involves propositions that putatively refer to some transformations that keep some mathematical-physical structure invariant. How should we, philosophers, stand with respect to them? An analogy to laws can be instructive. Physical theories involve propositions that refer to differential equations of motion that intend to represent dynamical laws. But *what* are the dynamical laws? Philosophical landscape, as is well known, is rather leafy. Nonetheless, in one way or another, the problem boils down to accommodating dynamical laws in one's worldview –for some, they represent genuine laws of nature, for others, they do not; for some, they represent laws of nature that govern phenomena, for others, they represent just regularities, etc. So, the problem of accommodating symmetries may be analogously couched. Propositions involving symmetries refer, first and foremost, to *properties* of differential equations within a physical theory's formalism,[1] but the metaphysical question is whether they represent some entity, property, structure, or relation in the ontology.

In the literature, it is customary to distinguish between those who think that symmetry claims refer to aspects of the world (realists) and those who think that they do not (anti-realists) (see, for instance, Brading and Castellani, 2007; Livanios, 2010). Such a division can shed light on the alternatives, but it is still a bit coarse-grained when we pay attention to the details. When we say things such as "symmetries are epistemic", do we mean that they are just conventions of some sort and that we might sooner or later dispense with them? Or when it is defended that symmetry claims refer to aspects of the world, does it mean that symmetries themselves are aspects of reality, or that they can assist us in discovering reality? And if they are part of reality, are they fundamental or derivative in the ontology? This article aims to clarify all these questions.

When the metaphysical problem of symmetries is viewed as a problem of accommodating symmetries in one's worldview, or as a problem of "locating" symmetries in one's ontology, there are at least three options: primitivism, where symmetries are fundamental (irreducible) parts of one's ontology; conservative reductionism, where symmetries can be reduced to something else more primitive, but still conserved as parts of the ontology (but not as parts of the fundamental ontology); or eliminative reductionism, where symmetries are not even part of the ontology (nor fundamen-

---

1 To be clear, the symmetries that have lately drawn philosophers' interests (as space-time symmetries, local gauge symmetries, and permutation symmetries) are primordially *known* in dynamical contexts and employed to simplify and handle dynamical problems. The respective symmetry transformations act upon terms appearing in dynamical equations and the symmetry is hence a property instantiated by the whole formal structure. It is true that we may, a fortiori, predicate the symmetry with respect to the structure of space-time, for instance. But, as I will argue shortly, this already supposes a philosophical position to take with respect to symmetries.

tal, nor derivative), but parts of the epistemic-methodological machinery to account for the ontology. It can be thought that eliminative reductionism, in this sense, it is not a metaphysical view of symmetries. I disagree. Eliminative reductionism implies a very clear statement on one's metaphysics and ontology *—there are no* symmetries in any relevant sense of "there are". Naturally, all these options have their advantage and difficulties as I will show in due course.

A final comment is relevant. Since symmetries come in many shapes and flavors, it would be tempting to think that some of them are real, whereas others are epistemic, leading to a kind of 'symmetry pluralism'. Methodologically, this approach could work (see Read and Møller-Nielsen 2020 for a methodological pluralist view of symmetries). But as I am interested in a metaphysical problem (which place symmetries occupy in one's ontology), the step from methodological pluralism to metaphysical pluralism is not so straightforward. It seems prima facie reasonable to think that, as different symmetries serve different purposes in physical theories, they will have different metaphysical status (i.e., some of them will be ontologically fundamental, others supervenient; others matter of convention, and others epistemically essential). This is a hard problem that deserves a deeper engagement, but the following warning is noteworthy. Even though symmetries are used differently in physical theories, an independent metaphysical argument in favor of such a metaphysical pluralism is needed. In other words, the fact that symmetries are used in different ways does not per se imply that they are all metaphysically different (promoting metaphysical pluralism).

An analogy with the metaphysical debates on law is instructive. Laws are quite varied (e.g., some of them are deterministic, some indeterministic, some probabilistic, some are first-order, others second-order, and so on). Humeanism about laws does not defend that some laws are regularities, whereas others govern phenomena. Neither do defenders of the governing view accept that some laws govern, whereas others supervene upon dispositional properties. My view is that symmetries should be treated similarly: while some form of methodological, or epistemic pluralism about symmetries can be endorsed (which allegedly follows from scientific practice), the plurality of symmetry kinds shouldn't be necessarily translated into the ontology automatically —scientific practice is insufficient for it, independent *metaphysical* reasons must be provided. In any case, this problem doesn't really affect the metaphysical map I lay out: all the metaphysical views I will distinguish can perfectly fine cohabitate in a metaphysically pluralist framework; or they could be alternative metaphysical stances in a globally monist metaphysical view.

3. *Symmetry Dispensabilism and Symmetry Indispensabilism*

In the Introduction, I briefly described two different kinds of symmetries that physics presents to us, but there are many. So, let us begin by clearly stating a blatant fact, the 'symmetry fact'.

Symmetry Fact Modern physics presents us with symmetry claims, that is, modern physics' discourse involves propositions p_i stating that in a theoretical framework a symmetry σ holds. Such propositions have contributed in the empirical success of physical theories.

Two comments are in order. The **symmetry fact**, as stated, is completely descriptive —it just happens that today's physics involves certain propositions referring to symmetries. As such, it is descriptively true, and everyone agrees on it. What is philosophically interesting about this, then? To my mind, the general problem is *whether* putative references to symmetries in p_i can be removed from our physical theories while preserving their empirical success[2]. I take that those that believe that we can*not* do this endorse what I call *Symmetry Indispensabilism* (**SInd**), while those that believe the opposite fall under what I call *Symmetry Dispensabilism* (**SDis**). It might be thought that this is just a convoluted way to rephrase the distinction between realism vs. anti-realism. But I do not think so, since **SInd**, as I argue later, does not

[2] It is worth noting that such a formulation of the problem is not an ontological problem yet. As it will become clearer later on in the paper, I take the ontological problem to be derivative from the general problem as just stated.

https://doi.org/10.1387/theoria.27141

necessarily imply a realist view of symmetries. In the same vein, I believe that an epistemic view of symmetries does not imply either the endorsement of **SDis**.

SInd Putative references to symmetries appearing in the **symmetry fact** cannot be dispensed with while preserving physics' empirical success.

SDis Putative references to symmetries appearing in the **symmetry fact** can be dispensed with while preserving physics' empirical success.

4. Symmetry Dispensabilism

Let us start with **SDis**. According to it, it is true that current physics employs propositions that refer to symmetries (**symmetry fact**), but we should not be misled into thinking that physics must formulate physical theories with such propositions. Future physics might well do it without symmetry claims. There may be scientific as well as philosophical reasons to hope for such a scenario. Think of renormalization in quantum electrodynamics (QED). For instance, it is a fact that in calculating some loop Feynman diagrams in QED, the integrals diverge. Hence, it is a fact that to calculate finite and testable expectation values, renormalization techniques are required. Yet, it is considered that they may be dispensable, in the sense that we can regard QED as an effective field theory, that is, a low-energy approximation of a still-unknown underlying high-energy physical theory. Then, physics could be dispensed with renormalization techniques. An analogous reasoning can, mutatis mutandis, be applied to symmetries —they can, for instance, be symptomatic of redundant structure that we should be dispensed with in future physics.[3]

Indeed, many have taken symmetries as symptomatic of redundant structure, suggesting that they guide new formulations of theories where such redundant structures are expunged entirely (see Dewar, 2015). This instrumental role of symmetries has, however, been the focus of controversy lately. Concerning the specific role of symmetries in symmetry-related models, two approaches have been distinguished: the *interpretational* and the *motivational* (see Møller-Nielsen, 2018; Luc, 2022). The interpretational approach holds that symmetries "allow us to interpret theories as being committed solely to the existence of invariant quantities, even in the absence of a metaphysically perspicuous characterization of the reality that is alleged to underlie symmetry related models" (Møller-Nielsen, 2018, p. 1256). In turn, the motivational view asserts that symmetries "only motivate us to find a metaphysically perspicuous characterization of the reality that is alleged to underlie symmetry related-models, but they do not allow us to interpret that theory as being solely committed to the existence of invariant quantities in the absence of any such characterization" (Møller-Nielsen, 2018, p. 1256).

Although this debate does not directly address the metaphysical problem of symmetries (its target is different) and it is in some way orthogonal to it, it is important to note that it might have some important metaphysical consequences. After all, symmetries are indeed employed as means to identify surplus structure. And, from it, some metaphysical lessons could be drawn. In due course, I will stress whether interpretational and motivational approaches favor one metaphysical view or the other.

Coming back to **SDis**, the symmetry discourse in the **symmetry fact** can also be dispensable based on purely metaphysical reasons, that is, our metaphysics could lead us to reject the symmetry discourse as referential. Think of dispositional

[3] It might be argued that the analogy doesn't really run because the elimination of symmetries will have ontological consequences, but not the elimination of renormalization techniques. But the argument already presupposes that symmetries might have ontological consequences, something that **SDis** denies. If symmetries were eliminated, the consequences would be purely methodological, as the elimination of other techniques. For the argument to go through, it should first be shown that the elimination of symmetries does have ontological consequences, which needs to be proved independently.

monism (e.g., Bird, 2007). Conforming to it, the basic ontology is given by dispositional properties from which laws are obtained. But, if this is so, then the laws can be only metaphysically constrained by the dispositions. Where should symmetries be placed? Are they genuine meta-laws that also constrain the laws? If they constrain the laws, then laws become over-constrained, and dispositions' fundamentality is contested. Bird's dispositional monism then takes symmetries to be pseudo-laws that form invariant structures that ultimately must be regarded as background structures; but "a desirable feature of physical theories is that they should eliminate background structures" (Bird, 2007: 213).

Bird would well represent what I call the "No Symmetry View" (**NSV**), a radical form of **SDis**. This is in close analogy with Nancy Cartwright's view of laws (Cartwright, 1983), mutatis mutandis. It can thus be argued that it is not an important fact of the scientific enterprise "to discover" anything like symmetries, but they are just useful devices that are found in highly idealized models. In that sense, physical symmetries would also 'lie' as they fail to refer to something external; or they 'lie' as they do not constrain phenomena, but idealized laws that can only be obtained in "shielded" models. In this sense, they also fail to represent an essential element in physics theorizing.

NSV Putative references to symmetries appearing in the **symmetry fact** fail to refer to entities, properties, or relations of the world's fundamental ontology. Nor do they capture an epistemically essential element in physics theorizing.

The most obvious problem for this view, one that will be recurrent in metaphysically light views, is that it doesn't really explain why symmetry claims have been so empirically successful. **NSV** seems to dispense with symmetries for metaphysical reasons (which are well-grounded), but it doesn't really explain why physical theories feature symmetries. It is nevertheless possible to find other views that are more systematic and explanatory, still within the **SDis** camp: *Symmetry Conventionalism* (**SC**) and *Symmetry Agentialism* (**SA**).

SC regards symmetries as conventions, that is, as a regular phenomenon observed in a determined community that does not depend on Nature or universal features found across communities. It rather depends on our choices (scientists' choices), which can be driven by numerous facts. Conventions are, then, "up to us", and thereby, the endorsement of symmetry claims that appear in the **symmetry fact** (as previously described) is also up to us —symmetries work like conventions among physicists that for pragmatic reasons became established in a community that share some epistemic standards and canons for theory acceptance. However, a crucial feature of conventions is that they can always be changed, and, to different degrees, are arbitrary. Then, if physical symmetries are conventions, whatever role they may play, it may be played by a choice. Therefore, they are dispensable. Christopher Martin (2002), Alexandre Guay (at least for Yang-Mills theories, 2004, 2008), and Sabine Hossenfelder (2018, where symmetries are almost aesthetic values) seem to uphold a form of **SC**.

An example of **SC** can be found in the so-called gauge symmetries. Alexandre Guay holds that gauge symmetries (i.e., the gauge structure) of Yang-Mills theories play a role in the quantization process as tools to obtain a field theory, being essentially pragmatic (Guay, 2008, p. 353). The fact that there are no alternative tools for quantization at sight, as Guay affirms, doesn't mean they could not exist. Nor does he argue that the gauge principle is *fundamental* for quantization. It just happens that physical theories *pragmatically* need the gauge principle, but it is not strictly necessary. Christopher Martin, in turn, argues that, in the case of the gauge principle, there exist different approaches to the gauge structure with different consequences for the "logic of nature", downplaying the role of the gauge principle –it might even not exist! (Martin, 2002, p. S233-S234).

SC Putative references to symmetries appearing in the **symmetry fact** refer to community-relative conventions that became established by various (epistemic, aesthetic, or pragmatic) reasons among the members of the community.

Poincaré's conventionalism could easily be extended to symmetries, making a case for **SC.** According to him, the adoption of a certain geometrical framework –and by extension, the symmetries inherent in them—is a matter of convention, selected on the basis of simplicity, unificatory power, and effectiveness in describing phenomena. **SC** is stronger than **NSV** because it explains what symmetries are (i.e., conventions), while **NSV** does not say anything about it. It does not explain the empirical success of physical theories, but because symmetries are not responsible for it (as the employment of the imperial metric is not responsible for the empirical adequacy of theories that use it). Nonetheless, although stronger than the **NSV**, conventionalism is still too weak, substantively diminishing the relevance of the **symmetry fact** for physical theories: there are good reasons to think that symmetries are *not* like metric units, for instance.

There is a more substantive naturalistic approach to epistemology and metaphysics that may be more satisfactory to those who want to avoid conventionalism, but to preserve **SDis** –*Symmetry Agentialism* (**SA**). **SA** can fairly be framed within what is known as the Sydney Plan (in square opposition to the famous Canberra Plan; see Price, 2011; Ismael 2014).

As far as I am aware, nobody in the literature has argued for **SA**, but I think it may be a workable philosophical position towards symmetries within **SDis**. In essence, the Sydney Plan is a systematic program for metaphysics that opposes viewing (scientific) concepts as representations (or 'mirrors') of reality. It is rather an anti-representationalist understanding of our scientific concepts that emphasizes the stories about agents and their relation to the world and how such concepts facilitate the interaction (Ismael, 2014). When facing a scientific concept, say *x,* we should not ask what *x* represents or how we can locate *x* in the scientific ontology, but what facts about our scientific understanding and the world jointly support the formation of *x,* and what role does it play in our scientific practice. In this way, scientific concepts and practices constitute "façades" (or user interfaces) to cope with the world: they do not represent external properties, objects, or structures, but "conceptual tools" to deal with the world.[4] The Sydney Plan then provides a fertile terrain for a naturalistic analysis of symmetries (as scientific concepts in the user interfaces), where the right question to raise is what facts about theory construction and experimental practice may require symmetries to cope with phenomena.

Two things are noteworthy. First, the relation of scientific-philosophical concepts to *agents*, and their purposes in the knowledge enterprise, is crucial (hence the name). Second, nothing in **SA** commits us to take symmetries as indispensable. After all, the basis is (at least partially) pragmatic, though more robust than conventionalism.

SA Putative references to symmetries appearing in the **symmetry fact** refer to elements in scientists' user interfaces[5] to navigate the natural world and to understand scientific practice and theorizing.

One of **SA**'s advantages is that it explains why physicists (as agents in active and pragmatic relations to the external world) employ symmetry claims in formulating physical theories. The potential problem is that it is essentially an anti-representationalist philosophy that is very critical of metaphysical projects as usually conducted. So, if SA is an alternative, it is a very radical one that doesn't even address the metaphysical problem of symmetries as usually presented.

Finally, it is worth mentioning a more radical, non-representationalist view of symmetries that can be framed within what I call *Symmetry Expressivism* (**SEx**). Expressivism is a well-known philosophical current in philosophy of language and metaethics claiming that sentences (moral sentences, for instance) are devices for expressing non-cognitivist positive or negative attitudes towards their objects. In moral language, expressivists would say that moral terms function much like "boo to ban abortion" or "hurrah to death penalty" (Ayer, 1974; Sinclair 2009). Therefore, semantic or moral sentences merely convey a non-cognitive attitude of approval or disapproval. It can be argued that in the physics community, physicists analogously employ symmetries. The **symmetry fact** expresses a non-cognitive attitude of physicists towards physical theories: they approve physical symmetries (or express a "pro attitude" towards them) in their physical theories for various non-cognitive reasons (e.g., an example would be "aesthetic pleasure"). This is clearly an eliminativist, metaphysically radical view of symmetries since it places physical symmetries on the side of scientists' non-

4 I believe that the anti-representationalist spirit of SA is the reason why it cannot qualify as a realist view, at least in the traditional sense.

5 In the Sydney plan, representations of all kinds are users' interfaces that facilitate interaction between the agent (as an active system) and an open environment. This is slightly different from mere conventions since scientific-philosophical concepts as interfaces play a specific and more active role in helping human beings navigate the world.

cognitive states of mind, introducing a good dose of irrationalism in physics. But it is a coherent view and can well be part of the metaphysical landscape.

SEx Putative references to symmetries appearing in the **symmetry fact** refer to scientists' non-cognitive states of mind that convey positive attitudes towards positing symmetries in physical theories.

Even though expressivism is a well-established philosophy in many brands of philosophy, it is alien to the philosophy of science. However, some types of expressivism have been lately defended in the philosophy of science for different scientific concepts. For instance, Josh Hunt has defended an expressivist account of explanation and relevance in science (2024). In the same line, he has proposed *nomological expressivism* for the case of laws of nature (Hunt work-in-progress, personal communication). Although nobody has thus far defended expressivism in the case of symmetries, symmetries involve concepts like explanations, laws of nature, counterfactuals, among others, which have been accounted for on expressivist grounds.

All these views make sense of physical symmetries as dispensable posits in physical theory. Contemporary physics indeed posits symmetries in physical theories, but it might as well stop doing it. The reason is that the apparent reference to symmetries in scientific discourse fails to refer to some external structure, relation, or entity. The problem is then to explain the specific role that symmetries play in knowledge, despite being dispensable. They may merely refer to contingent community-relative conventions, to strategies to cope with an external world, or to scientists' non-cognitive positive attitudes. Be that as it may, all of them do not only claim that symmetries are not part of one's ontology, but that even in our way of describing the ontology, they play an accessory role. Whoever thinks that this falls quite short of making sense of symmetries would rather endorse some form of *indispensabilism*, **SInd**.

5. Symmetry Indispensabilism: Realist Views

For many, **SDis** categorically fails to make sense of physics' success. The **symmetry fact** captures something crucial of physics that cannot be dispensed with if its success is to be preserved. In this way, **SInd** endorses the opposite view: not even in principle can all the putative references to symmetries in the **symmetry fact** be removed. Symmetries are nothing like renormalization techniques, nor are they just conventions. Even less do they express scientists' pro attitudes. We do not even have a good reason to expect their future elimination, but quite the opposite, since the trend in the last decades has moved towards a "symmetry-first-approach". Also, if a philosophical position cannot accommodate symmetries' crucial role, then this is per se a form of reductio against it (see Steven French's argument against Bird's dispositional monism, French, 2014, p. 249). This could suggest that **SInd** entails some form of realism. Nonetheless, as I previously mentioned, **SInd** does *not* directly imply a realist attitude towards the symmetry discourse. Indeed, I submit that **SInd** can be divided into five different views, not all of them realist: *Symmetry Realism* (**SR**), *Symmetry Fundamentalism* (**SF**), *Symmetry Inferentialism* (**SI**), *Symmetry Epistemicism* (**SE**), and *Symmetry Normativism* (**SN**).

Let's start with *Symmetry Realism* (**SR**). **SR** simply claims that the justification of the **symmetry fact** is that symmetries do refer to mind-independent structures, relations, or entities in the external world. Physics cannot dispense with symmetries without stopping to capture something important about the world that is within the physics domain (physical entities, properties, relations, or structures). It just happens that the world has such a structure that makes our theoretical statements involving symmetries (approximately) true, which is just a wordy way to say that symmetries are indispensable because they are (somehow) out there in the world. Two arguments uphold **SR** as stated –a type of the No Miracle argument and a type of the Quine-Putnam Indispensability argument. According to the former, from the **symmetry fact** we can infer that the empirical success of the introduction of symmetries in physics can only be explained by stating that symmetries must be 'out there' in the world; otherwise, the empirical success of modern physics would be a miracle. Since we do not believe in miracles, the best explanation of the **symmetry fact** is that symmetries must be 'out there' in the world. A No Miracle type argument, therefore, forces us to take symmetries as real.

A type of Quine-Putnam Indispensability Argument places them in one's ontology:

> P1. We ought to be ontologically committed to only those theoretical posits that are indispensable to our (best) physical theories,
>
> P2. Symmetries are properties of laws of nature that are indispensable to our best physical theories,
>
> C. Therefore, we ought to be ontologically committed to symmetries.

The argument is Quinean in spirit: physical theories existentially quantify over symmetries. Physical theories cannot be formulated without symmetries and involve true propositions (or, at least, it is far from clear how physical theories can dispense with symmetries without losing empirical adequacy). Therefore, symmetries exist, and we should be committed ontologically to them. Nonetheless, **SR** is weaker than expected. The reason is that 'existence' can be a generous word and, paraphrasing Armstrong, symmetries could enter the ontology as 'an ontological free-lunch' (Armstrong, 1997, p. 12; Schaffner, 2009, p. 353). Then, physical symmetries, according to **SR**, do exist but they are not necessarily part of the fundamental ontology; they can rather be supervenient (or derivative) structures, relations, or entities that depend on more fundamental structures, relations, or entities.

SR Putative references to symmetries appearing in the **symmetry fact** refer to entities, properties, or relations of the world's ontology, although they are not necessarily fundamental.

SR in this sense could be endorsed by some versions of contemporary Humeanism. Physical symmetries play the role of meta-laws in the best systems, constraining first-order facts. This is the view of Michael Hicks (2019), where symmetries provide constraints on first-order facts, that is, the regularities that are find in the Humean mosaic. In this sense, as constraints of first-order facts, symmetries cannot be purely epistemic: facts about symmetries can be located in the Humean mosaic, although they are not fundamental structures, relations, or entities in the Humean mosaic. In this sense, physical symmetries may exist as constraints as constraints, but not as fundamental, primitive structures in the Humean mosaic. Toby Friend (2024), in turn, has argued that symmetries are not merely meta-laws, but have something to do with the world's structure, in particular, with the world-making relations —they are grounded in the space-time structure, therefore, revealing fundamental aspects of it. I then take this view as implying that symmetries supervene upon the Humean mosaic, qualifying as a realist view of symmetries.

It is important to note that **SR** can employ against **SDis** the same argument that scientific realists employ against instrumentalist views. Indeed, if symmetries are in some sense real, then **SDis** is not a viable option. Of course, **SDis** of any sort can argue that empirical success is not enough for truth; or, in other words, that empirical success is not a good guide to metaphysics. The dialectic can then be viewed like the following. **SDis** regards that symmetry claims appearing in the symmetry facts are dispensable. The challenge is then to explain *why* they are there despite being dispensable. I think that there are ways to do this, as I have shown before. Yet, some could claim that this is not enough, that it is still mysterious why symmetry claims have been so (empirically) successful in physics. This prima facie suggests some form of **SR** that makes sense of symmetry claims in virtue of their reference to something physical out there. Nonetheless, empirical adequacy might not be the best guide to ontology. In addition, indispensability doesn't necessarily imply realism, as I will argue later.

More popular is the view that physical symmetries *are* fundamental. This is what I call here *Symmetry Fundamentalism* (**SF**), which takes **SR**'s arguments a step forward. There seems to be an additional step from the Quine-Putnam Indispensability Argument to the claim that symmetries are therefore metaphysically fundamental. The Quine-Putnam Indispensability Argument introduces symmetries into the ontology, but it doesn't say anything about the role they play in the ontology. It could then be argued, following Armstrong, that we might be ontologically committed to symmetries, but they are an 'ontological free-lunch'; that is, they are derivative structure, relations, or entities that are "no addition to being". Therefore, they might well be dispensable when determining the *fundamental* ontology (see Armstrong, 1997; and Schaffer, 2009, for a general argument in the same line). To put it differently, while the Quine-Putnam Indispensability argument deals with existence questions (what there is), the metaphysically important question is not about existence but about modes of existence (i.e., as fundamental or derivative). After all, if some theoretical posits are introduced in one's ontology as an ontological free lunch, then one's ontology is no less sparse for containing them than it is for containing the entities which ground them.

SF avoids this by promoting symmetries to the fundamental ontology, both for scientific and philosophical reasons. When the **symmetry fact** is closely examined, symmetries do not merely 'appear' in our physical theories, but they also play a (theoretical) grounding role –for instance, a theory's dynamics is seen as relying on symmetry statements; symmetries are said to 'dictate' the fundamental interactions or, even, the kind of particle that intervene. So, it is very plausible to read this symmetry-way-of-talking as representing an ontological relation. In the theory, symmetry facts ground, for instance, nomological facts or property facts (e.g., facts about elementary particle masses or energies). On this basis, there is also an ontological relation that places symmetries as fundamental, and the rest as derivative (see McKenzie 2014 for a similar argument).

One way to understand **SF** is then that it justifies **SInd** on a metaphysical basis —symmetries are indispensable because they are (or directly represent) aspects of the fundamental ontology. **SF** can then be viewed as **SInd** plus a strong realist attitude towards symmetries and their promotion as part of the fundamental ontology. It then goes beyond empirical adequacy. Symmetries are thus not only part of the ontology in general, but they are entities, properties, structures, or relations of the building blocks of the physical world, upon which the derivative ontology supervenes

SF Putative references to symmetries appearing in the **symmetry fact** refer to entities, properties, or relations of the world's fundamental ontology.

Some advocates of **SF** are Steven Weinberg (1987, 1993), Abdus Salam (1989), Richard Feynman (1987), and Werner Heisenberg (1975) on the physicists' camp. On the philosophers' camp, Steven French (2014) and David Schroeren (2020, 2021) have endorsed **SF**, though on different grounds. As an example, this view has been explicitly defended by, for instance, David Schroeren (2021). Within wave-function monism, state-space symmetries (i.e., automorphisms) are established as fundamental relations that correspond to automorphisms in the (fundamental) ontology. In this sense, then, **SF** amounts to two claims: (a) state-space symmetries correspond to fundamental aspects (relations) of physical reality and (b) particles and fields are ontologically derivative from these fundamental aspects.

The natural advantage of **SF** is that it does justice to the exceptional role of symmetry claims in physics over the last decades. The **symmetry fact** just reflects how physics has discovered what's fundamental. The natural criticism of **SF** is that it overloads the fundamental ontology by adding symmetries as fundamental (i.e., primitives). Of course, any ontology has to decide which entities, properties, relations, and structures are primitive and which ones are derivative. It can therefore be argued that it is not necessary to regard symmetries as fundamental (i.e., primitives) since they can be reduced to something else. It can also be argued that **SF** implies a series of ontological commitments that are too onerous, such as being realist about idealizations and laws of *nature* (see López, 2025).

6. Symmetry Inferentialism: an intermediate position

I have said that **SInd** does not per se force us to adopt a realist attitude towards symmetries. The challenge for any anti-realist position is to show, first, why physical theories have been so successful without making symmetries look like a miracle; and second, in which way physical theories may take symmetries claims as indispensable, without adopting ontological commitments to them. I think that *Symmetry Epistemicism* (**SE**) and *Symmetry Normativism* (**SN**) are two possible metaphysical views that are successful in this respect. But a third, intermediate variation of **SInd** can also be identified. One can embrace **SInd** but resist placing symmetries in the ontology (either derivative or fundamental). Yet, symmetries are not absolutely disconnected from the fundamental ontology, since they serve as "indicators" of it. This view, *Symmetry Inferentialism* (**SI**), can be defined as follows:

SI Putative references to symmetries appearing in the **symmetry fact** do not refer to aspects of reality but allow us to *infer* aspects of it.

According to **SI**, propositions referring to symmetries enter as premises into symmetry-based arguments (see, for instance, Shamik Dasgupta's version of the 'symmetry-to-reality inference', Dasgupta, 2016; also, North, 2009, 2021) to draw metaphysical lessons about the world. Then, symmetries are indispensable but not as theoretical posits that latch onto the world, but as theoretical, heuristic tools for probing what is fundamental. Jill North (2021) seems to adopt **SI**:

> There is a reason for formulating things in terms of structure rather than symmetries, though. Structure is what we are ultimately after (both mathematical structure in the formalism and physical structure in the world), and symmetries are simply an (important) guide to that structure. As mentioned in Chapter 2, symmetries are an indicator of structure, not the structure itself. More importantly, there can be more to the requisite structure than what seems to be indicated by dynamical symmetries. (North, 2021, p. 73)

Just to provide some further examples, Eddy K. Chen's nominalization of non-relativistic quantum mechanics also relies on symmetries as a guide to what is fundamental (Chen, 2019, PhD Dissertation, Ch. 1), from which his quantum state realism derives. Paul Horwich's formulation of the problem of the arrow of time accords with **SI**: since time-reversal invariance is a nomological property, it allows us *to infer* the isotropy of time itself (Horwich, 1987, p. 41). Finally, it has been repeatedly claimed that permutation invariance in quantum theories *entails* that quantum systems are non-individuals (Post, 1963; Fortín and Lombardi, 2021). Therefore, from permutation invariance, we can infer a relevant feature of the world's fundamental ontology —it does not comprehend individuals.

Three comments are in order. First, **SI** might also be regarded as construing symmetries epistemically, and that's why I place it as an intermediate position between realist and epistemic views. As I mentioned before, 'epistemic' usually refers to conditions regarding observational indistinguishability (such as observability, measurability, detectability, etc.). This is one of the senses of 'epistemic' that, for instance, Shamik Dasgupta refers to when assessing symmetries in the context of symmetry-to-reality inferences (Dasgupta, 2016, p. 871). In the same line, James Read and Thomas Møller-Nielsen (2020) defend an epistemic view of symmetry transformations, according to which epistemic symmetries "render the general notion of a symmetry transformation redundant as a tool for metaphysical theorizing about scientific theories" (2020, p. 97). Even though they don't directly concern the metaphysical problem of symmetries, their epistemic approach seems to be at odds with, for instance, **SR**, **SF** and **SI**. This would suggest a metaphysically epistemic view of symmetries, but it is not clear of which sort exactly (for alternative epistemic approaches to symmetries in this sense, see Ismael and van Fraassen, 2003, and Caulton, 2015).

Second, that **SI** is an instance of **SInd** can also be disputed. In Section 3, I have referred to the distinction between interpretational and motivational approaches to symmetries. While the interpretational approach can endorse **SI** and **SInd** since symmetries guide our interpretation of theories as *solely* committed to the existence of invariant quantities, the motivational approach would probably endorse **SI** but under **SDis**. Nonetheless, interpretational and motivational approaches to symmetries are insufficient to deliver a clear metaphysics of symmetries —they are mainly concerned with the role that symmetry transformations play in interpreting physical theories rather than with the task of placing symmetries in one's ontology.

Finally, it has been argued that **SI** is not a good guide to metaphysics since symmetries only appear in highly idealized models and are stipulated in theory construction and modelling for theoretical- and pragmatic-driven reasons (see López, 2023). This criticism goes along with Read and Møller-Nielsen's view of symmetries as epistemic, although the emphasis is placed on the conditions under which symmetries are obtained rather than on the interpretation of symmetry transformations.

7. *Symmetry Indispensabilism: Epistemic-Deflationist Views*

In the previous section, I have said that in the literature on symmetries, 'epistemic' usually refers to the conditions regarding indistinguishability between symmetry-related models or to the role of symmetries in the invariance principle. However, there is a different sense in which symmetries can be regarded as epistemic, closely related to the metaphysical problem of symmetries. This sense treats symmetries as 'epistemic' in the sense of not being part of one's physical ontology, but "general conditions of physical knowledge" or "epistemic constraints." Vassilios Livanios, when contrasting 'ontic' and 'epistemic' approaches, says:

> Philosophers, who try to explain the success of symmetry considerations in science, have followed two different approaches. According to one of them (the epistemic viewpoint), the presence of symmetries in physical theories is related to general conditions of physical knowledge or to some limits inherent in our way of describing the physical world. According to the other (the ontological viewpoint), symmetries are real aspects of the world, usually taken to be properties of the world (or of its structure) or second order laws concerning the form of physical first order laws. (Livanios, 2010, p. 296)

Livanios is not here referring to conditions of observational indistinguishability. Nor is he referring to the role that symmetries play in the invariance principle. He is rather referring to symmetries more generally, replying to the metaphysics-oriented question of what *they are.* When I refer to symmetries as 'epistemic', I will do it in this sense.

I have affirmed that **SInd** does *not* necessarily entail a connection between the **symmetry fact** and what the world is like. This is so because I see a distinction between *epistemic* and *ontological* indispensability. Some theoretical posits can be indispensable to, for instance, systematizing and providing unity to a theory, but it does not mean that we must then be ontologically committed to them. For instance, in classical electromagnetism the electric potential is indispensable to formulate and solve Maxwell's equation, but it doesn't seem to be *therefore ontologically* indispensable (See Maudlin, 2018, for discussion)[6] Rigid bodies are equally indispensable in classical mechanics to formulate the kinematics and dynamics, but they don't *therefore* seem to be *ontologically* indispensable.

The distinction between epistemic and ontological indispensability, hence seems valuable and important. Kerry McKenzie runs a similar distinction between being "fundamental to the fabric of the world" and "fundamental in methodological respects" (McKenzie, 2014, p. 1091). According to her, when symmetries are said to have a relation of ontological dependence between them and, say, elementary particles, they are ontologically fundamental; when symmetries are said to be fundamental as heuristic device for theory construction (as in the case of the prediction and the detection of the Higgs boson), they are methodologically fundamental. Be that as it may, as was mentioned before, the word 'fundamental' has been extensively used by metaphysicians to refer to those entities, properties, structures, or relations on which everything depends, already bearing a strong realist commitment. To avoid confusion, I draw a similar distinction but in terms of indispensability. In my case, ontological indispensability refers to the fact that symmetries enter physical theories because, without them, they will miss relevant descriptions of what the world is physically like, either fundamentally (as in **SF**) or derivatively (as in **SR**).

The distinction is also useful because it allows overcoming No-Miracle-type arguments, avoiding realism (**SR**, more specifically). The empirical success of symmetries in physics can be granted, favoring their indispensability. But real-

[6] Maudlin discusses how the different formulations of the basic structure of a physical theory (in this case, classical electromagnetism) entails that the posits of the theory can be either derivative or fundamental. For instance, when the theory is cast into the language of fiber bundles, the fields are derivative, contrary to the canonical formulation. Although Maudlin's point is to show the need of ontological clarity to avoid conceptual problems, the message to drive home here is the following —theories don't per se imply an ontology as indispensable since what it looks indispensable for the canonical formulation might not be so in alternative formulations.

https://doi.org/10.1387/theoria.27141

ism needs an additional step *if* indispensability can be made sense of in purely epistemic terms. **SInd** can thus branch out further, opening an entire family of philosophical positions towards symmetries that do regard them as indispensable, but as *epistemically* indispensable. This reveals two remarkable aspects. First, **SInd** is compatible with epistemic-normative interpretations of the **symmetry fact**. Second, some philosophical frameworks seem to embrace **SInd** but are reluctant to adopt **SR**, **SF,** or **SI**, but without making symmetries look like a miracle. In this light, these deflationary-epistemic views claim that putative references to symmetries appearing in the **symmetry fact** do not refer to aspects of reality but have an epistemic justification in terms of giving some "systematic epistemic gain" for physical theories (e.g., by making physical theories simpler).

One of these views is what I call *Symmetry Epistemicism* (**SE**). According to it, symmetries play an epistemic-heuristic role in physics theorizing or modelling that is crucial for physics. They can, for instance, serve to an ideal of optimization between simplicity and explanatory power. An easy example is the unification they confer —instead of formulating different laws for different directions in space, it is more economical (epistemically speaking) to formulate a rotation-invariant law (see Rosen, 2018). Under this perspective, it can be argued that it is not enough to formulate empirically adequate lawlike generalizations about phenomena. It is also highly convenient to formulate them most simply without losing explanatory power; and positing symmetries as epistemic-heuristic constraints can play that role.

Although I have before associated Humeanism with **SR** (but not with **SF**), I think that **SE** can also be viewed as compatible with Humeanism, but emphasizing the epistemic nature of symmetries. In this case, Humeans could rather highlight symmetry claims' epistemic-heuristic nature: symmetries are primarily grounded in the formal relations defined over generalizations about the facts of the Humean Mosaic. Or, even stronger, they play a crucial role in the systematization and unification of the true generalizations within the best system. In this line, symmetries are precisely the kind of theoretical resources that allow the best systems to substantially boost their simplicity and informativeness. Symmetries are therefore epistemically indispensable in scientific knowledge since they confer upon the best systems a substantial gain in simplicity and informativeness in simplicity and informativeness; to the same extent to which our knowledge of the natural world benefits from building generalizations about phenomena, it benefits from imposing symmetries upon how such generalizations must be formulated. To what extent are so-construed symmetries *indispensable*? To the extent that physicists *must* strive to formulate better physical theories, in the sense of the simplest and most informative ones. Or, to put it differently, if physical theories are interpreted as best systems, what makes best systems *best* is their epistemic virtues in terms of simplicity, unification, economy of thought, etc. In that sense, it is indispensable for physical theories to include symmetries since, otherwise, they would not really be the *best* systems (see López and Esfeld, 2023, for an argument in the same line).

SE Putative references to symmetries appearing in the **symmetry fact** do not refer to aspects of reality but have an epistemic justification in terms of the optimal systematization of the first-order generalizations that supervene upon regularities. This optimal systematization is crucial to balance simplicity and informativeness.

Metaphysical frameworks that can easily adopt **SE** are Humean Structuralism (Lyre 2010), Super-Humeanism (Esfeld, 2015), or David Albert and Barry Loewer's *Mentaculus* (Loewer, 2012; Albert, 2015). A similar view has been defended by López and Esfeld (2023) in relation to time-reversal symmetry, although their views can be interpreted as a mixing of **SR** and **SE**. In all of them, symmetries might be regarded as real insofar as they supervene on the Humean Mosaic, but they are also an "ontological free-lunch". Therefore, their indispensability does not derive from their being somehow real. Contrarily, symmetries are then primarily assessed in terms of their epistemic contribution to the construction of physical theories, from which their indispensability does derive. For the case of time-reversal symmetry, López and Esfeld argue that the reason for having time-reversal symmetry in our physical theories is not because the Humean mosaic is temporally directionless, but because it makes the formulation of first-order generalizations simpler: the laws of the theory do not depend on the direction of time, gaining modal scope.

Finally, I would like to introduce a philosophical position that goes farther than **SE** but can still be viewed as epistemic. It also makes **the symmetry fact** completely natural and even necessary, avoiding any form of realism. In fact, the view is not new, though it has lost its luster over the last decades. It can be traced back to early works on symmetries by David Hilbert (1921), Eugene Wigner (1949, 1963) and Hermann Weyl (1952), to the neo-Kantian philosophy of Ernst Cassirer (1923 [1910], 1954 [1936]), and to John Nozick's insights than symmetries are closely connected with the notion of objectivity (Nozick, 2001; for some antecedents, see Dirac, 1947 [1930] and Weyl, 1952). This view accepts that all (or most of) our symmetry claims are indispensable for modern physics, but because they play a *normative* role in making physics (both as a discipline and as an object for such discipline) possible by underpinning the notion of physical objectivity. I call this view *Symmetry Normativism* (**SN**).

SN rather turns to a *critical* understanding of objectivity as the necessary unity of knowledge (in the Kantian sense). Neo-Kantians have exhaustively worked on the notion of objectivity and distinguish two types of theories: the copy (or substance) theory and the critical (or functional) theory (Cassirer, 1923, p. 321-323). According to the former, the primitive concept is that of 'object' (or 'structure'); truth is explained in terms of *adequatio*: a proposition is true iff it mirrors the properties of the object or the structure. According to the critical theory, the copy theory overlooks the epistemic preconditions for having objects (in this case, physical, scientific objects). In this theory, the right question to raise is not what physical objects are like, but what makes them possible. The notion of objectivity, hence, must be understood transcendentally —what are the conditions of possibility for physics theorizing and for empirical research to be about physical objects at all (Cassirer, 1923; Heis 2014).

Which role does the **symmetry fact** play as a result? **SN** holds, first and foremost, that propositions involving symmetry statements are a priori *norms* that are the transcendental conditions of objectivity in modern physics. **SN** regards laws and experimental results as facts of modern physics, which are possible in virtue of some preconditions that cement their objectivity, where the references to symmetries in the **symmetry fact** play a crucial role. What are they crucial for? Eugene Wigner has shed light on this in relating symmetries to the necessary invariance of initial conditions, and the very notion of lawhood (Wigner, 1963). Similar views can be found in David Hilbert (see Ryckman, 2008) and his idea of "des-anthropomorphization" of physics through invariances.

SN Putative references to symmetries appearing in the **symmetry fact** do not refer to aspects of reality but refer to transcendental structures (in the normative-Kantian sense) within which physical objects are possible. For physics (both in the sense of knowledge about physical objects and of the possibility of physical objects themselves) to be possible at all, symmetries are required to give the necessary unity, systematicity and permanency that physical objectivity requires.

It is hard to show that *all* symmetries play a normative role in this specific sense. But think of space-time symmetries in classical mechanics as a clearer example. Nozick thought that one of the conditions for physical objectivity was the invariance of a presumed object under various transformations (Nozick, 2001, p. 78). Objectivity, he says, means that some features are independent from the angle from which we see them. In physics, this matches well with various space-time transformations. It can therefore be argued that objects (or quantities) that remain invariant under space-time transformation are physically objective as they do not depend on the reference frame from which they are described. Then, symmetries would play the role of establishing what is objective within a physical theory: those quantities and objects that are invariant under spatial rotation, spatial translation, time translation, time reversal, boosts, etc. Nozick's view per se is closer to **SR** or **SI**, but on this basis, it can be further argued that the objectivity that symmetries confer is not found in nature but imposed by us. In other words, such invariances are not out there in the world and represented in our physical theories, but normatively imposed by us in order to have objective, physical knowledge.

It is important to stress the scope and limits of **SN** (and also of **SE**). First, **SN** can explain the indispensability of symmetries without appealing to realism: symmetries are necessary as epistemic preconditions for having knowledge of physical objects, so the **symmetry fact** comes out as a natural requirement for physics. Second, when saying that physical symmetries are epistemically indispensable, it is not meant that knowledge in general indispensably requires symmetry claims. We can very well obtain valuable knowledge of the world without even acknowledging the existence of the **symmetry fact**. The claim circumscribes itself to knowledge that is obtained from physics. That is, the knowledge that physics provides us indispensably requires symmetry claims, but other forms of knowledge do not necessarily. **SN**, for instance, sets the conditions for having *physical* objects (that is, objects as known by physics), not objects in general.

8. Conclusions

In recent years, the problem of accommodating symmetries in one's ontology has drawn the attention of many philosophers of physics and metaphysicians. In general, it has been customary to distinguish between realist and anti-realist views. Although this distinction is useful, it is perhaps not sufficient to capture all the possible metaphysical attitudes that can be adopted towards symmetries. In this article, I have laid out alternative views that configure a map of the metaphysical attitudes towards physical symmetries. Although not aiming for completeness, it goes beyond the realist *vs.* anti-realist opposition, highlighting the subtleties of similar, but different views, and paving the way for yet unexplored views:

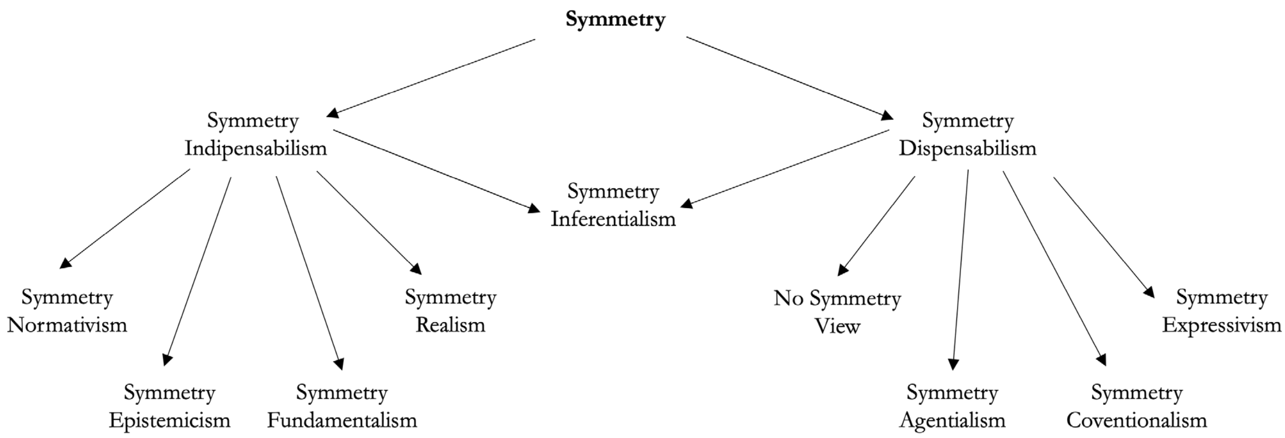

Figure 1

A coarse-grained map of possible metaphysical attitudes towards symmetries.
Dotted arrows represent possible connections

Acknowledgments

I am grateful to Olimpia Lombardi, Michael Esfeld, Alexander Guay, the *Buenos Aires Group of Philosophy of Particular Sciences*, and the *Buenos Aires Group of Metaphysics of Science* for comments and discussion on earlier versions of this article. Some sections of this article have been presented at EPSA 2021, Foundations of Physics 2021, and the BSPS 2022 conference, so my thanks to their audiences for valuable comments. I also grateful to two anonymous reviewers for their valuable feedback, comments, and suggestions. This research was supported by the Swiss National Science Foundation (SNSF, grant no. 105212B_200464), by the Université de Lausanne, and by the National Scientific and Technical Research Council of Argentina (CONICET).

REFERENCES

Albert, D. (2015). *After Physics*. Hardvard University Press.

Armstrong, D. M. (1997). *A World of States-of-Affairs*. Cambridge University Press.

Ayer, A. (1936/1974). *Language, Truth and Logic*. Penguin.

Baker, D. (2010). Symmetry and the metaphysics of physics. *Philosophy Compass* 5, 1157-1166.

Ballentine, L. (1998). *Quantum Mechanics. A modern development*. World Scientific

Belot, G. (2013). Symmetry and equivalence. In R. Batterman (ed.). *The Oxford Handbook of Philosophy of Physics* (pp. XXX), Oxford University Press.

Bird, A. (2007). *Nature's metaphysics.* Oxford University Press.

Brading, K. and E. Castellani. (2007). "Symmetries and Invariances in Classical Physics". In J. Butterfield and J. Earman (eds.). *Handbook of the Philosophy of Science, Philosophy of Physics, Part B* (pp. 1331-1367). Elsevier.

Cartwright, N. (1983). *How the laws of physics lie*. Clarendon Press.

Cassirer, E. (1923). *Substance and Function*. Open Court. [1910]

Cassirer, E. (1954). *Determinism and indeterminism in modern physics*. Yale University Press [1936]

Caulton, A. (2015). The role of symmetry in the interpretation of physical theories. *Studies in History and Philosophy of Modern Physics*, 52, 153-162.

Dasgupta, S. (2016). Symmetry as an epistemic notion (twice over). *British Journal for Philosophy of Science*, 67, 837-878.

Dirac, P. (1947). *The principles of quantum mechanics*. Oxford University Press. [1930]

Earman, J. (1989). *World enough and space-time. Absolute versus relational theories of space-time*. MIT Press.

Esfeld, M. (2015). Quantum Humeanism, or: physicalism without properties. *Philosophical Quarterly*, 64 (256), 453-470.

Feynman, R., and Weinberg, S. (1987). *Elementary Particles and the Laws of Physics: the 1986 Dirac Memorial Lectures*. Cambridge University Press.

French, S. (2014). *The structure of the world.* Oxford University Press.

Friend, T. (2024). How to be Humean about symmetries. *British Journal for the Philosophy of Science*. Forthcoming.

Guay, A. (2008). A partial elucidation of the gauge principle. *Studies in History and Philosophy of Modern Physics*, 39, 346-363.

Hall, G. and Murgueitio Ramírez, S. (2024). Symmetries and Representation. *Philosophy Compass*, https://doi.org/10.1111/phc3.12971

Heis, J. (2014). Realism, functions and the a priori: Ernst Cassirer's philosophy of science. *Studies in History and Philosophy of Science*, 48, 10-19.

Hicks, M. (2019). What everyone should say about symmetries (and how Humeans get to say it). *Philosophy of Science,* 86, 1284-1294.

Heisenberg, W (1975). Development of concepts in the history of quantum theory. *American Journal of Physics,* 43, 389:394

Horwich, P. (1987). *Asymmetries in time*. MIT Press.

Hossenfelder, S. (2018). *Lost in math*. Basic Books

Hunt, J. (2024). Expressivism about explanatory relevance. *Philosophical Studies*, 181, 2063-2089.

Ismael, J. (2014). Naturalism on the Sydney Plan. in M. Haug (ed.). *The Armchair or the Laboratory?* (pp. 86-103), Routledge.

Ismael, J. and van Fraassen, B. (2003). Symmetry as a guide to superfluous theoretical structure. In E. Brading and E. Castellani (eds), *Symmetries in Physics: Philosophical Reflections* (pp. 371-392), Cambridge University Press.

Livanios, V. (2010). Symmetries, dispositions and essences. *Philosophical Studies*, 148, 295-305.

Loewer, B (2012). Two accounts of laws and time. *Philosophical Studies,* 160, 115-137.

López, C. (2021). The philosophy and physics of time reversal in standard quantum mechanics. *Synthese*, 199, 14267-14292.

López, C. (2023). Should physical symmetries guide metaphysics? Two reasons why they should maybe not. *European Journal for the Philosophy of Science,* 13, 23

López, C. (2025). Against Symmetry Fundamentalism. *Erkenntnis*, vol. 90, 1821-1845.

López, C. and Esfeld, M. (2023). Humean Time Reversal. *Synthese,* 202, 31

Luc, J. (2022). Motivationalism vs. interpretationalism about symmetries: some options overlooked in the debate about the relationship between symmetries and physical equivalence. *European Journal for Philosophy of Science*, 13, 40.

Lyre, H. (2010). Humean Perspectives on Structural Realism. In F. Stadler (ed.). *The Present Situation in the Philosophy of Science* (pp. 381-97), Springer.

Martin, C. (2002). Gauge principle, gauge arguments and the logic of nature. *Philosophy of Science,* **69**, S221-234.

Maudlin, T. (2018). Ontological clarity via canonical presentation: Electromagnetism and the Aharanov-Bohm effect. *Entropy*, 20(6), 465.

McKenzie, K. (2014). On the fundamentality of symmetries. *Philosophy of Science,* 81, 1090-1102.

Møller-Nielsen, T. (2018). Invariance, interpretation and motivation. *Philosophy of Science*, 84, 1253-1264.

North, J. (2009). The structure of Physical theories. *Journal of Philosophy*, 106, 57-88.

North, J. (2021). *Physics, structure, and reality*. Oxford University Press.

Nozick, R. (2001). *Invariances: the structure of the objective world*. Harvard University Press.

Post, H., (1963). Individuality and Physics. *The Listener*, 70, 534–537.

Price, H. (2011). *Naturalism without mirrors*. Oxford University Press.

Read, J. and Møller-Nielsen, T. (2020). Redundant epistemic symmetries. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, vol. 70, 88-97.

Redhead, M. (1975). Symmetry in intertheory relations. *Synthese,* 32, v. 1/2, 77-112.

Ryckman, T. (2008). Invariance Principles as regulative ideals: from Wigner to Hilbert. *Royal Institute of Philosophy Supplement*, 63, 63-80.

Salam, A. (1989). Gauge Unification of Fundamental Forces. *Nobel lecture* on December 8, 1979.

Schaffer, J. (2009). On what grounds what. In D. Chalmers, D. Manley and R. Wasserman (eds.). *Metametaphysics* (pp. 347-383), Oxford University Press.

Schroeren, D. (2020). Symmetry fundamentalism: a case study from classical physics. *Philosophical Quarterly*, 71 (2), 308-333.

Schroeren, D. (2021). Symmetry fundamentalism in quantum mechanics. *Philosophical Studies*, 178, 3995-4024.

Sinclair, N. (2009). Recent Work in Expressivism. *Analysis* 69 (1), 136-47.

Weinberg, S. (1987). Towards the Final Laws of Physics. In R. Mackenzie and P. Durst (eds), *Elementary particles and the laws of physics – The 1986 Dirac Memorial Lectures* (pp. 61–110) Cambridge University Press.

Weinberg, S. (1993). *Dreams of a final theory*. Vintage.

Weyl, H. (1952) Symmetry. Princeton: Princeton University Press; extract in E. Brading and E. Castellani (eds), *Symmetries in Physics: Philosophical Reflections* (pp. 21-22.) Cambridge University Press, 2010.

Wigner, E. (1949). Invariance in physical theory. *Proceedings of the American Philosophical Society*, 93 (7), 521-526.

Wigner, E. (1963). Events, laws of nature, and invariance principles. *Nobel Lecture*, December 12.

Cristian López is a SNF Senior Researcher at the University of Lausanne (Switzerland) and a CONICET Associate Researcher at the University of Buenos Aires (Argentina). His areas of specialization are the philosophy of physics, the philosophy of time, the metaphysics of science. He has also worked on the philosophy of economics and the political philosophy.

Address: University of Lausanne, Anthropole, Section de Philosophie, Of. 5155, CH-1015, Switzerland. E-mail: cristian.lopez@unil.ch – ORCID: 0000-0002-2883-4037

Theoria, 2025, 40(2), 172-186
https://doi.org/10.1387/theoria.27509

THEORIA
ISSN 0495-4548 – eISSN 2171-679X

Models of data and the representation of phenomena: a pattern-based inference

(Modelos de datos y la representación de fenómenos: una inferencia basada en patrones)

José V. Hernández-Conde*
University of Valladolid
https://orcid.org/0000-0002-8502-6570

María Caamaño-Alegre*
University of Valladolid
https://orcid.org/0000-0002-7005-9257

Keywords

Pattern
Phenomenon
Statistical inference
Nested modeling
Isomorphism

ABSTRACT: Since the 1960s, the distinction between data and phenomena has fueled debates in the philosophy of science, with scholars arguing that data must be modeled in order to serve as evidence for phenomena. We claim that the modeling of data to obtain evidence for phenomena involves four levels: data, sample structure, population structure and phenomena. Our analysis suggests that the notion of pattern is essential to fully grasp the inferential capacity of data models, where representation occurs through nested surrogative reasoning —typically in the form of an isomorphism that holds at different layers. We also explain how our taxonomy of pattern-based inferential steps could shed light on various aspects of nested data modeling, such as the risk of theoretical bias. To illustrate our proposal, we examine the Eddington experiment —which tested general relativity by observing the deflection of starlight near the Sun—, and show how patterns at different levels of the data modeling provide the basis for nested surrogative reasoning in this case. Transforming the data points identified on photographic plates into a representation of light deflection requires a multi-layered search for patterns, where each pattern takes us one step further in data modeling and one step closer to the target phenomenon.

Palabras clave

Patrón
Fenómeno
Inferencia estadística
Modelización anidada
Isomorfismo

RESUMEN: Desde la década de 1960, la distinción entre datos y fenómenos ha dado lugar a amplios debates en la filosofía de la ciencia, habiendo un claro acuerdo en que los datos deben modelarse a fin de que sirvan de evidencia para determinar la ocurrencia de fenómenos. Sostenemos que la modelización de datos, con el fin de obtener evidencia sobre fenómenos, involucra cuatro niveles: datos, estructura de la muestra, estructura de la población y fenómenos. Nuestro análisis sugiere que la noción de patrón es esencial para entender la capacidad inferencial de los modelos de datos, donde la representación se produce mediante razonamientos subrogatorios anidados —habitualmente en forma de isomorfismo entre diferentes capas. También explicamos cómo nuestra taxonomía de pasos inferenciales basados en patrones podría arrojar luz sobre diversos aspectos de la modelización de datos anidada, como aquellos relativos al riesgo de sesgo teórico. Para ilustrar nuestra propuesta, examinamos el experimento de Eddington —que puso a prueba la relatividad general observando la desviación de la luz de las estrellas cerca del Sol—, y mostramos cómo los patrones en diferentes niveles del modelado de datos proporcionan la base para el razonamiento subrogatorio anidado en este caso. Transformar los puntos/datos identificados en las placas fotográficas en una representación de la desviación de la luz requiere una búsqueda de patrones en varios niveles, en donde cada patrón constituye un paso adelante en el modelado de datos y nos acerca un paso más al fenómeno objetivo.

* ***Correspondence to:*** José V. Hernández-Conde. Department of Philosophy, Faculty of Philosophy and Arts, University of Valladolid, Plaza del Campus s/n, 47011, Valladolid, Spain – jhercon@uva.es – https://orcid.org/0000-0002-8502-6570

* ***Correspondence to:*** María Caamaño-Alegre. Department of Philosophy, Faculty of Philosophy and Arts, University of Valladolid, Plaza del Campus s/n, 47011, Valladolid, Spain – mariaconcepcion.caamano@uva.es – https://orcid.org/0000-0002-7005-9257

How to cite: Hernández-Conde, José V.; Caamaño-Alegre, María (2025). «Models of data and the representation of phenomena: a pattern-based inference»; *Theoria. An International Journal for Theory, History and Foundations of Science,* 40(2), 172-186. (https://doi.org/10.1387/theoria.27509).

Received: 28 April, 2025; Final version: 24 July, 2025.

ISSN 0495-4548 - eISSN 2171-679X / © 2025 UPV/EHU Press

1. *Introduction*

Scientific representation encompasses a wide and diverse range of practices. This paper narrows its focus to a specific domain within that practice: one in which models of data play the key representational role, by enabling inference from data to phenomena.

Since at least the 1960s, the distinction between data and phenomena has given rise to important debates in the philosophy of science. Both Suppes (1962) and, later, Leonelli (2009, 2015, 2019) argue that data can only constitute evidence for phenomena once they have been modeled, that is, assembled and arranged in a particular way. In a similar vein, Suárez (2024) has recently characterized the models of data, within certain domains of inquiry, as low-level mathematical models that represent those abstract features of the data that are relevant for the purposes at hand. His inferential conception of scientific representation emphasizes how mathematical models —such as statistical models of data— involve nested surrogative reasoning between their different layers (Suárez, 2024, p. 167).

Our purpose is to show the importance of the notion of pattern for understanding the inferential capacity of data models. We contend that, despite the fact that patterns play a crucial a role in the inference from data to phenomena, the notion of "pattern" has not been deeply investigated from an analytical point of view. This has often led to an ambiguous use of this notion, as noted by Bruce Glymour (2000), who proposes replacing the distinction between data and phenomenon with a classical distinction in statistics, namely, that between sample and population structure. We will suggest a revision of Glymour's proposal, advocating the need to distinguish between four distinct levels, i.e., data, sample structure, population structure and phenomena, and will explain the character and nature of each step necessary to move from one level to the next. We argue that the inferential, representational capacity of data models can be better understood as a three-step inferential process involving the above four levels. Our claim is that this multilayered account of data models can be useful to account for the kind of nested surrogative reasoning invoked by Suárez (2024) in his characterization of models of data.

From what has been said, it is clear that our concept of pattern is statistical-mathematical. The present analysis of the notion of pattern is limited to cases where the term "pattern" is applied to something that can be derived from a set of data. We stress the importance of patterns in identifying regularities in mathematical representations. These regularities can be expressed numerically. We distinguish three different senses in which the term 'pattern' is be used: (a) mathematical, as a distinction between signal and noise;[1] (b) statistical, which includes both descriptive and inferential uses of statistics; and (c) geometric or informational, as found in DNA chains, the growth patterns of shells and Brussels sprouts. The main sense in our proposal combines (a) and (b). Patterns that are neither mathematical nor statistical in nature —sense (c)—, such as those associated with the identification of the structure of DNA fall outside the scope of our article, since they determine an entity rather than a phenomenon. Consider X-ray crystallography, where evidence famously led Rosalind Franklin and her collaborators to attribute an antihelical structure to the 'dry' form of DNA. In this case, the properties of an entity are established based on the patterns seen in the photographs. However, no specific phenomenon is identified, unlike the melting point of lead, the deflection of light, or linear learning.[2]

We will illustrate the pattern-based inferential steps from data to phenomenon by using a well-known example (Laymon, 1982; Mayo, 1996; Kennefick, 2009, 2019), namely, the so-called "Eddington experiment", which was intended to empirically test the theory of general relativity by confirming the prediction about the deflection of light. In this case, the small sample size entails special difficulties that can only be overcome by introducing local

---

1 This is the prevailing sense in McAllister's (1997) discussion of the inference from data to phenomena. While we share some of his concerns about the indeterminacy of this process, we also believe that our approach can also help to address this issue. However, we will leave this for another paper.

2 See Schindler (2008; 2013, pp. 95-96).

https://doi.org/10.1387/theoria.27509

checks and plausible assumptions about how the instruments function. This made it possible to determine a sufficiently representative sample pattern, i.e., the deviation of the points of light captured in the photographs from the center of the Sun as it appears in them. Based on this, an inference can be made about the phenomenon of light's path deviating when passing close to the Sun. This final step involves making a causal inference, since phenomena are conceived as the (often unobservable) cause of the observed data. Our example also shows how patterns, in the statistical-mathematical sense, play a crucial role in nested surrogative reasoning.

Our approach in this paper unfolds in three steps. First, we offer a concise overview of our perspective on the inferential process from data to phenomena. The next step is to examine how our framework can be useful in understanding data models that involve nested surrogative reasonings between their different layers. Finally, we provide an example to show how, in models of data such as those at play in Eddington experiment, the surrogative reasoning characteristic of representation manifests as isomorphisms holding between different layers of those data models.

2. *Bridging the gap between data and phenomena*

2.1. The data/phenomenon distinction

To frame the problem of patterns, we first examine the debate over the distinction between data and phenomena, where the concept of "pattern" proves highly relevant. As Woodward (2010, p. 792) points out, the introduction of this distinction was motivated by its relevance in addressing the epistemic problem of using data as evidence. The discussion stems from a question that Antonis Antoniou has recently articulated very clearly:

> How exactly are theoretical hypotheses eventually confronted and tested by experimental results given that the latter are often produced in the form of large datasets and in a language that is not accessible to the theory? (Antoniou, 2021, p. 2)

This question has been central in contemporary philosophy of science, where it is widely accepted that empirical phenomena emerge from a complex collection of data (Bogen & Woodward, 1988; Woodward, 1989, 2000, 2010; McAllister, 1997, 2011; Glymour, 2000; Massimi, 2007; Leonelli, 2015, 2019; Bokulich, 2018, 2020). Since the early characterization of the notion of data models by Patrick Suppes (1960, 1962), significant efforts over recent decades have explored how data models are applied in specific research contexts to infer phenomena from data (Leonelli, 2009; Antoniou, 2021; Bokulich & Parker, 2021).

All of the aforementioned authors agree that raw, unprocessed data do not serve as evidence. This debate already has a long history, which we do not aim to examine in detail. Rather, we assume the widely accepted view that scientific observations and experiments rarely allow for the direct observation or detection of phenomena. Instead, they detect certain effects that are captured in the form of data, which can, in principle, be attributed to the target phenomena. The epistemic problem lies in determining the occurrence of stable phenomena from a heterogeneous and unstable set of data, which almost always occurs in combination with noise and whose understanding is also limited (Woodward, 2010, p. 794).

Suppes addressed data models in the 1960s, conceptualizing them as statistical constructs that distill complex experimental processes, whose purpose is to facilitate comparisons between empirical results and theoretical predictions. Although Suppes did not explicitly focus on patterns, his emphasis on homogeneity, stationarity, and order tests suggests an implicit recognition of their importance in data modeling. Under his view, it is the patterns, not the raw data set or unmodeled experimental results, what can be interpreted as phenomena, whether consistent with predictions or not.

Later, Bogen and Woodward reinforced the distinction between data and phenomena. Arguing against van Fraassen's (1980, 1989) view, they held that theories are primarily concerned with unobservable phenomena, rather than observed data (Bogen & Woodward, 1988; Woodward, 1989). According to Bogen and Woodward, data are unstable, observable reports from experiments, unlike phenomena, which are stable, theory-explained features of the world inferred from reliable data patterns. In contrast to data, phenomena are unobservable but serve as evidence for theories because they are the underlying cause of observed data.

More recently, discussions have been focused on how, within different fields, scientists generate and process data to identify empirical phenomena, as exemplified by Leonelli's work in biology (2015, 2016, 2019) and Bokulich's (2018, 2020) in paleontology. Leonelli (2019) views data models, unlike unmodeled data, as actual, ready-to-use evidence, designed to systematically organize data for representing phenomena and supporting evidential claims. Similarly, Bokulich (2020) contends that data models surpass "raw data" in accuracy and reliability, and thus are more appropriate tools to science's epistemic goals. In this vein, Leonelli (2009, 2015, 2019), and Bokulich and Parker (2021) highlight the pragmatic nature of data models, as specifically designed tools to yield particular evidence to support a certain theory.

In McAllister's 1997 and 2011 works, the notion of pattern is central to his analysis of reasoning from data to phenomena, in an approach that is focused on the issue of the extreme underdetermination in this inference. As a result, he does not explicitly define patterns, beyond equating them with the notion of signal in a conception where data is split into signal and nose, and phenomena are identified with signals. In McAllister's view, the identification of patterns —or signals— depends on how data are decomposed into signal and noise, which allows for multiple interpretations of the same original data set. This leads to the concern that inference from data to patterns —and ultimately to phenomena— is fundamentally undermined by the ambiguity of the data, since any data set can yield an indefinite number of patterns.

Particularly interesting for our proposal is Glymour's (2000) critique of the vague use of the term in the data-to-phenomena inference debate, and his suggestion of replacing the philosophical distinction between data and phenomena with the statistical concepts of sample structure (that is, patterns recognized in the data sample) and population structure (that is, population model identified with the phenomenon). According to Glymour, the latter distinction offers greater precision, with sample structure consisting of statistical measures derived from the collected data, and population structure representing the generalization of these measures to the entire population. As a result, the inference from data to phenomena is seen as a two-step process: first, a pattern is identified within the sample, and second, the identified pattern is extrapolated to the entire population and used to define the phenomenon under study.

However, while Glymour's focus on statistical distinctions is valuable, his analysis conflates data with sample structure and population structure with phenomena, leaving the two senses of pattern present in the discussion unclear and mixed with other separate notions. In this regard, our analysis emphasizes the mediating role of patterns as essential for inference from data to phenomena, and the transformation of raw data into interpretable structures that, once generalized, provide a basis for characterizing the underlying phenomenon.

2.2. Patterns as mediators between data and phenomena

Patterns are essential to both statistics and contemporary scientific research. From a philosophical and statistical point of view, patterns are sets of properties in the form of regularities that characterize the members of a particular group. Recognizing patterns is the goal of several statistical methods, including cluster analysis, discriminant analysis, decision trees, and artificial neural networks. These approaches help distinguish between categories and uncover relationships between indirect measures and the phenomenon under investigation. Examples include discharge patterns in electronic particle detectors and reaction times in psychological studies (Bogen & Woodward, 1988).

https://doi.org/10.1387/theoria.27509

We will now show how to distinguish data from patterns and patterns from phenomena by defining specific meanings of "pattern" and organizing the analysis into three sequential steps.

Step (a): From data to patterns in the data (sample statistics)

The initial step involves transforming raw empirical data into structured and interpretable sample statistics. This step involves processing raw empirical data collected in an experiment to identify statistical patterns, which includes computing descriptive statistics such as means, medians, standard deviations, correlations, and visual representations (e.g., histograms), together with data cleaning, variable selection, and outlier removal are performed to refine the data set.

This process is both statistical-methodological and partly epistemological. The raw data, often messy and unstructured, is subjected to cleaning (e.g., removing outliers, handling anomalies) and summarization (e.g., means, medians, correlations) to reveal trends or patterns. These patterns are not yet general claims about the world, nor claims about the target phenomenon; they are descriptive abstractions of the sample at hand. Techniques like histograms or preliminary clustering help distinguish signal from noise, making this a foundational act of pattern recognition. As a result, this step helps differentiate meaningful patterns from noise and provides a summarized representation of the sample behavior.

This first step has an interesting dual character. On the one hand, it is technical (relying on statistical-mathematical tools) but also it is interpretive, as decisions about which variables to prioritize or how to handle inconsistencies unavoidably shape the resulting patterns. This dual character is the norm in contemporary scientific practice, where data preprocessing is as critical as the analysis itself.

Step (b): From sample to population patterns (population statistics)

The second step generalizes the patterns found in the sample to the whole population. The transition from sample statistics to population statistics introduces a leap of generalization, which has a statistical-epistemological character. It requires assumptions about sample representativeness and usually involves techniques such as statistical inference, confidence intervals, and hypothesis testing. On this basis, this step hinges on inference —extrapolating patterns observed in a limited sample to a broader population. Statistical tools such as confidence intervals and hypothesis testing quantify the uncertainty of this jump, despite the fact that the inference is based on an assumption that cannot be conclusively proven, such as the representativeness of the sample.

As a result, the transition from sample statistics to population statistics introduces an epistemological gap, since it relies on probability-based estimations and, more importantly, on extrapolations. Unlike sample statistics, which are directly derived from data, population statistics remain theoretical and require careful validation. This point is where the epistemological weight becomes evident. Unlike step (a), which is grounded in the concrete data collected, step (b) moves into the realm of the unknown, estimating population parameters that cannot be directly observed. Despite this or, precisely, because of it, this step is crucial for science's aim to produce universal assertions and laws, and ultimately causal explanations that go beyond the limits of empirical knowledge.

Step (c): From population patterns to phenomena

The third step interprets population patterns as empirical representations of particular phenomena. Thus, this step moves from population statistics to the empirical characterization of the phenomenon itself, which entails a causal inference from data to phenomena (i.e., the phenomenon is conceived as the —often unobservable— cause of the observed data).

As a result, patterns attributed to the population are interpreted as manifestations of an underlying phenomenon. This is a distinctly epistemological move, as it involves hypothesizing a causal relationship between the observed patterns and the "hidden" reality they reflect. Researchers may use decision trees, cluster analysis, or dimensional reduction techniques —such as factor analysis or principal component analysis (PCA)— to identify key explanatory factors. However, the interpretive role of the researcher is critical —deciding which variables best capture the phenomenon.

Two-step or three-step approach?

Finally, depending on the analytical method adopted, the transition from data to phenomenon can follow either a three-step approach —progressing through sample statistics, population statistics, and empirical characterization of the phenomenon— or a two-step approach, where dimensional reduction could allow a direct identification of the phenomenon at the sample level.

The two-step approach —using techniques such as factor analysis or principal component analysis (PCA) directly on the cleaned data— constitutes a significant alternative. By identifying underlying factors or components early on, this approach collapses the three-step process into two: (1) from the data to the sample-level phenomenon, via dimensional reduction, and (2) from the sample-level phenomenon to the population-level phenomenon, via statistical validation. This approach bypasses the explicit construction of sample statistics as an intermediate step, and instead searches for the phenomenon in the underlying structure of the data.

3. Inferential steps, patterns and nested modeling

In the following sections, we will examine how our approach can reveal the nested nature of data models and the crucial role of patterns in the inferences that lead to the representation of phenomena. To this end, we will rely on Suárez's view of models of data, which explicitly incorporates the concept of nested reasoning —a notion that we also deem to be important.

3.1. Nested modeling in Suarez's view of data models

Although Suárez's inferential account of representation excludes isomorphism between source and target structures as an essential condition for representation, he nevertheless acknowledges its significant role as a means of scientific representation (Suárez, 2024, pp. 167-168). He emphasizes that surrogate reasoning, when based on theoretical and mathematical models, is often facilitated by isomorphisms between source and target. Among the various domains of modeling where isomorphism is most likely to play this role, Suárez draws attention to the domain of data models. Thus, despite his rejection of isomorphism as a constitutive element of representation, he regards data models as a type of model in which isomorphism is most plausibly the means of representation.

Suárez (2023, pp. 46-54; 2024, p. 167) characterizes models of data as low-level mathematical constructs that represent abstract and relevant features of the data, often through isomorphisms. In his view, mathematical models of data are multilayered and involve nested surrogate reasoning between their different layers. Consequently, he refers to both isomorphism and nested modeling as features of data modeling, which together help to enable surrogative reasoning from source to target. In this context, he states the following:

> We saw that, while mathematical models —such as the Lotka-Volterra model or statistical models of data— are prima facie candidates for isomorphic relations, they turn out in more complex ways to involve nested surrogative reasonings between their different layers. These models are again no objection to [inf] since any means to promote surrogative reasoning from source to target is compatible with the source's inferential capacities. (Suárez, 2024, p. 167)

In an earlier paper, Suárez (2023) examined an example of nested modeling in the field of asteroseismology —also discussed by Castellani & Schettino (2023). In this case, data models involve the identification of fluctuation patterns in the luminosity of stars which, supported by multiple modeling assumptions, allow for the inference of seismic oscillations within them. These seismic oscillations, in turn, provide highly valuable evidence for models of stellar structure. The data models that enable the identification of seismic oscillations are nested in a complex background of models, which Suárez analyzes in detail (2023, pp. 120-123). The expression 'nested models' intuitively refers to the idea that, in some contexts of inquiry, "(...) all models contain assumptions that involve or result from other models" (Suárez, 2023, p. 112). Suárez characterizes nested modeling not only as a presuppositional relationship between models, but also as one that involves the use of real (collected) data to enrich the empirical interpretation of theoretical models and to develop them further —thereby allowing the data to be inferred or estimated through these enriched, more sophisticated theoretical models. More specifically:

> It concerns the nested nature of modelling; the fact that most models operate against a background that incorporates further models, and where data is routinely tested backwards, as it were —by deducing the values of theoretical parameters from the data given some background assumptions, and where those parameters are then fed into new models that can appropriately generate the data. (Suárez, 2023, p. 112)

It is worth noting a peculiarity of the case of asteroseismology that is shared with other cases from geology and engineering, namely the use of forward or inverse modeling to explore how different modeling alternatives yield different estimates of the data. Stellar structure models serve as theoretical building blocks that provide preliminary and partial guidance for interpreting brightness patterns in terms of seismic oscillations. This guidance is subject to revision if the data do not agree with the predictions of the theoretical stellar structure models (*ibid.*, pp. 125-126). In the example analyzed by Suárez, the evidential —or "observational"— basis of stellar astrophysics is established by means of a complex combination of nested models required to obtain and interpret data on stellar seismic oscillations. These data are derived from the fundamental phenomenon of regular pulsations in the brightness of a star, which are caused by gravitational or acoustic oscillations generated by rotational or convective forces. Further modeling assumptions are required to go from the data on seismic oscillations to the oscillation patterns of nodes (radial and non-radial) for both pressure and gravity modes (*ibid.*, p. 121). The overall oscillatory profile of a star provides highly valuable evidence for evaluating not only stellar structure models, but also traditional modeling assumptions in the form of idealizations. Suárez emphasizes the need to re-examine the nested models once the data modeling is complete, especially if the results suggest that certain assumptions should be reconsidered. In the case of asteroseismology, assumptions such as blackbody radioactive equilibrium, spherical symmetry and uniform composition may be challenged by the empirical evidence gathered.

Suárez's description of nested models operating in the above case suggests a kind of circularity; however, he claims that this does not amount to an experimenter's regress in the sense described by Collins (1992). The latter would involve the assumption of theoretical models in data modeling, thereby (partly) determining the very evidence used to test those same theoretical models. As is well known, Collins' classic example is the detection of gravitational waves. In his view, theoretical assumptions —such as those about stellar structure— may be useful for establishing some tentative constraints on how data should be derived or explained, but this would be part of a wider procedure aimed at comparing actual data, obtained by independent means, with data inferred under certain theoretical assumptions. In Collin's example, theoretically predicted oscillation frequencies are compared with actual observed oscillations, and discrepancies lead to adjustments in the theoretical parameters or background models used to make the predictions. As Suárez argues, rather than introducing a theoretical bias, the nested structure of data models allows for a more thorough assessment of how evidence is constructed and evaluated.

3.2. Nested modeling as a pattern-based inference

We believe that our approach, which is in line with Suárez's view in several ways, can provide valuable insights into how data models are characterized. Although our approach is not committed to a particular theory of representation, it reinforces the idea that isomorphism plays a major role in how data models represent phenomena. In particular, we emphasize the importance of patterns in capturing regularities in mathematical representations, which makes them particularly well-suited for representing phenomena using isomorphisms. In our view, the notion of pattern is essential to understanding the inferential capacity of data models. Furthermore, we fully back the idea that, in typical cases of data modeling, surrogate reasoning —characteristic of scientific representation— relies on isomorphisms that hold across different layers of the model.

The Eddington experiment is also useful to see how, in principle, the nested modeling required for the development of data models is not necessarily subject to the problem of experimenter's regress. For the reasons set out by Kennefick (2019) and emphasized by Suárez (2025), we believe that the concerns raised by Earman and Glymour (1980) regarding Eddington's experiment are misguided. As convincingly argued by Kennefick in a chapter entitled "Not only because of theory" (2019, pp. 181-225), the background models supporting the light deflection inference were independent of the theoretical models from general relativity that led to Einstein's prediction. Although general relativity and Einstein's predictions were taken into account in drawing conclusions from the data, the data themselves were generated independently and stood on their own. Neither the generation of the data nor their selection and modeling depended on theoretical assumptions from Einstein's theory. Kennefick provides historical evidence to show that even if both Eddington and Dyson admired the theory as a theoretical achievement, neither of them assumed its adequacy in advance. It is therefore inaccurate to attribute to them a confirmation bias in data selection, as Earman and Glymour do, because neither of them assumed the adequacy of the theory. Other reasons given by Kennefick regarding the methodology used to select data further support the claim that the selection process was unbiased. An analysis of the nested models involved in this case sheds light on the quality of the data selected and the methodological soundness of their modeling. The discarded data —those more consistent with Newtonian predictions— came from photographic plates where interfering factors, known to affect the operation of the instruments, had clearly degraded the image quality. If assumptions about confounding factors were critical for selecting reliable data, assumptions about confounding factors were equally important for modeling the data to identify a pattern in the shift of starlight. As we will see in the next section, Einstein's prediction was only confirmed when the effect of the main confounding factor —the scale shift— was calculated and subtracted from the original observed pattern.

In the next section, we show what this inferential practice involving different layers would look like if our characterization of data models were supplemented with some insights from Suárez's approach. In applying our taxonomy of inferential steps, we show how the presuppositional relations between layers in nested modeling can be seen as corresponding to the presuppositional relations between layers in inference from data to phenomena via patterns. In both cases, models are built on a background of other models, and the quality of evidence can only be properly assessed by tracing the lineage of modeling assumptions involved in the nested structure.

4. *Application Example: Eddington Experiment*

4.1. Experiment presentation

An important example of inference from data to phenomena is the "Eddington experiment", designed to empirically test Einstein's general relativity by confirming the predicted deflection of starlight near the Sun during the 1919 solar eclipse starlight (Laymon, 1982, pp. 108-113; Mayo, 1996, pp. 132-139, chapter 8; Kennefick, 2009, 2019; Suárez, 2025). Woodward (2000, p. 164; 2010, pp. 792-797) frequently mentions this case to illustrate how data from photographic plates mix the target variable (light deflection) with confounding factors such as telescope peculiarities, thermal effects on instruments and chemical influences during plate development. As Woodward (2010, pp. 793, 797) argues, understanding these additional variables is essential to assess the reliability of inferences from data to phenomena. In our analysis, this knowledge aids in identifying patterns that serve as evidence for the phenomenon under study, and applying the concept of "pattern" enhances the precision of such inferences.

As is well known, British astronomers Frank Watson Dyson and Arthur Stanley Eddington organized two expeditions to gather evidence from the 1919 eclipse, in order to test Einstein's prediction of light bending near a massive body (Laymon, 1982; Mayo, 1996). The method relied on discrete data, namely, star positions on photographic plates taken during the eclipse and compared with night-time plates of the same field, captured months later when the Sun had shifted from the Hyades. Three data sets were produced by the two expeditions, the first was to Sobral in Brazil and the second was to the island Príncipe off the West African coast. Data sets 1 and 2 were generated using the 4-inch Sobral telescope and the Sobral astrograph, respectively. Data set 3 was produced with the Príncipe astrographic telescope. The data were cleaned and selected, excluding data set 2, as well as light point positions from photographic plates with images considered defective due to astronomical or atmospheric effects.[3]

In this experiment, the inference process involves a key pattern: the sample pattern, that is, the observed displacement of star positions —i.e., light points— in the eclipse photographs relative to the Sun's center. This pattern was used to infer the phenomenon of starlight deflection near the Sun. It took decades to expand the sample size to obtain a more robust, generalizable population pattern.[4] However, thanks to the efforts made in controlling secondary factors and analyzing sources of error, the sample pattern was found to be a good representation of the phenomenon. Although the similarity between the Príncipe and Sobral observational patterns did not in itself justify their selection, they were chosen (on the basis of other factors) for the subsequent extrapolation of their results to a broader population —namely, all instances where starlight passes near a massive body. In this case, the epistemological factor had much more weight than the statistical one, since the sample size (i.e., the number of locations from which the eclipse observation was made) was small, making the statistical dimension less relevant than it would have been had there been more locations. Although the photographic plates were considered representative, the astronomers recognized the need for further validation and confirmed the pattern's consistency through additional observations (Kennefick 2009: 38). This is why there is a two-step pattern-based inference in this example instead of a three-step one.

The first step involved identifying the sample pattern by superimposing eclipse and comparison plates, cleaning defective data (e.g., due to atmospheric effects), and isolating light deflection from confounding factors like scale shift —a radial displacement mimicking deflection. Unlike deflection, which peaks near the Sun's limb and diminishes with distance, scale shift affects stars farther from the image center most strongly. Hence, a key distinction exists between the two effects. Light deflection is most pronounced for stars near the Sun's limb and diminishes with distance, whereas scale shift has the greatest impact on stars far from the image center, with minimal effect near the center.

To accurately isolate the pattern tied to light deflection and rule out scale shift as the cause, calculating the scale shift effect is essential. The Cambridge team addressed this by taking "check plates" of a different star field at night, both in the UK and Príncipe. These plates served to detect significant scale changes between the eclipse and comparison photographs, enabling Eddington to account for variations arising from differences in timing, location, or

[3] As data sets 2 and 3 were both of poor quality, there has been a heated debate as to whether excluding only data set 2 —which was more in agreement with Newton's prediction— was a biased decision. While some argue that it was (Earman & Glymour, 1980), others oppose this idea (Kennefick, 2009, 2019; Schindler 2013), albeit for different reasons. For a brief overview of our position on this matter, see our comments in Section 3.2 above.

[4] For a survey of light deflection measurements from 1919 to 1960 see von Kluber (1969, pp. 58-63).

equipment setup. Using the positions on the check plates (white positions in the image), Eddington quantified the scale shift and subtracted its effect from the light point displacements observed when superimposing the eclipse and comparison plates. This subtraction isolated the light deflection effect, represented as the deviation α between the black and white positions of the same stars shown in Figure 1. Thus, in this experiment selecting the appropriate data and controlling the primary confounding factor (i.e., scale shift) are pivotal in identifying the sample pattern associated with light deflection.

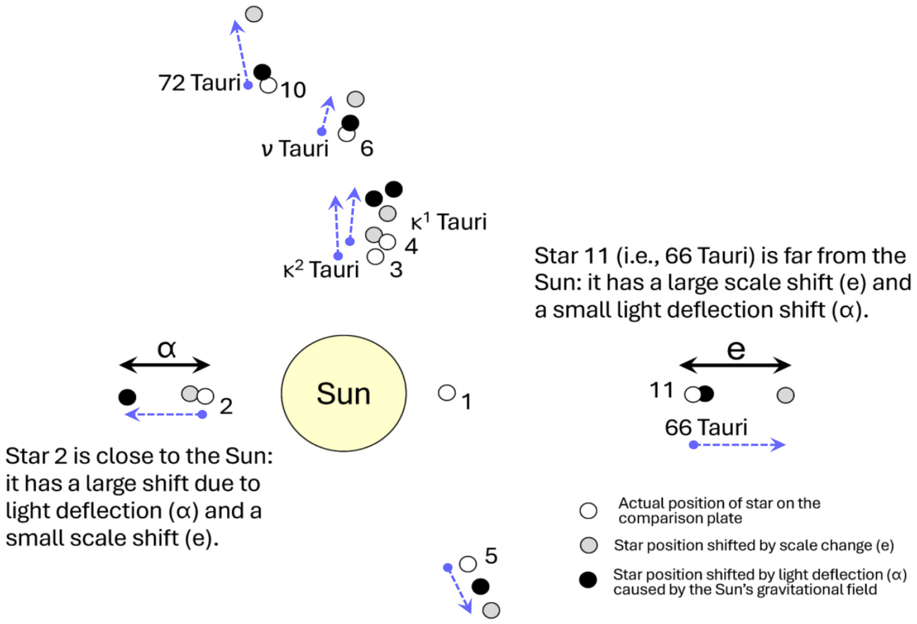

Figure 1

Position of the stars closest to the Sun during the 1919 eclipse.
[Diagram adapted from Kennefick (2019, p. 191)]

The second step entails inferring the phenomenon from the established population pattern via causal inference. This step confirmed Einstein's prediction of light bending by interpreting the generalized pattern in step two as evidence of the phenomenon. Using standard astronomical statistical methods, the light point displacements were analyzed as deviations. The results revealed Sobral's deviation angle as 1.98 ± 0.12 arcsec and Príncipe's as 1.61 ± 0.30 arcsec. Given the Newtonian prediction of 0.87 arcsec for light grazing the Sun's surface, these findings aligned with Einstein's predicted value of 1.75 arcsec, confirming the phenomenon of light deflection.

4.2. Interpretation in terms of nested modeling and isomorphisms

Let us now examine how Eddington's experiment can be interpreted in terms of nested modeling and isomorphisms. Eddington's experiment exhibits the typical presuppositional relationships between models that characterize nested modeling. Rather than focusing on the theoretical model as the final result, we concentrate on the empirical model that underpins the observation of light deflection. This empirical phenomenon is established on the basis of other models, with the model representing the former being nested within those supporting the latter. In what follows, we will schematically outline the source-target relation at each level of modeling, along with the relation that enables the representation of the target by the source. To fully account for the role of isomorphism, we introduce an initial step that was ignored in our previous analysis.

The starting point of data handling in the experiment is the selection of photographic plates, from which the sample data are selected. In this case, the SOURCE$_{(0)}$ is the relative positions of light points in selected photographic plates, the TARGET$_{(0)}$ is the relative positions of stars in Hyades clusters, and the ISOMORPHISM$_{(0)}$ taking place between SOURCE$_{(0)}$ and TARGET$_{(0)}$ is the means to infer that points in the plates correspond to light coming from the stars (see Figure 2).

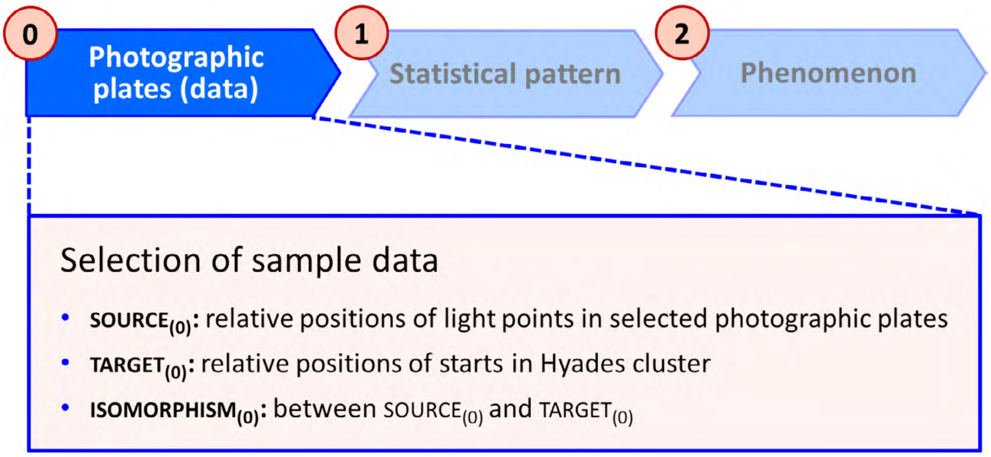

Figure 2

Isomorphism in the selection of sample data. [Initial step]

Then, researchers identified a statistical pattern in a sample of photographic plates. In this initial step, the SAMPLE$_{(1)}$ are the photographic plates (i.e., the relative positions of light points in selected plates, which was SOURCE$_{(0)}$), the SOURCE$_{(1)}$ is the statistical patterns of displacement of the light points in the photographic plates relative to the center of the Sun —as depicted in Figure 1—, and the TARGET$_{(1)}$ is the light's behaviour displacement in the sample depending on whether it passes or not close to the Sun. ISOMORPHISM$_{(1)}$ takes places between the SOURCE$_{(1)}$ (i.e., sample statistics from the places) and the TARGET$_{(1)}$ (i.e., trajectory of light in the observational scene), enabling the inference from source to target (see Figure 3). However, no isomorphism takes place between the SOURCE$_{(1)}$ (i.e., sample pattern) and the SAMPLE$_{(1)}$ (i.e., the photographic plates).

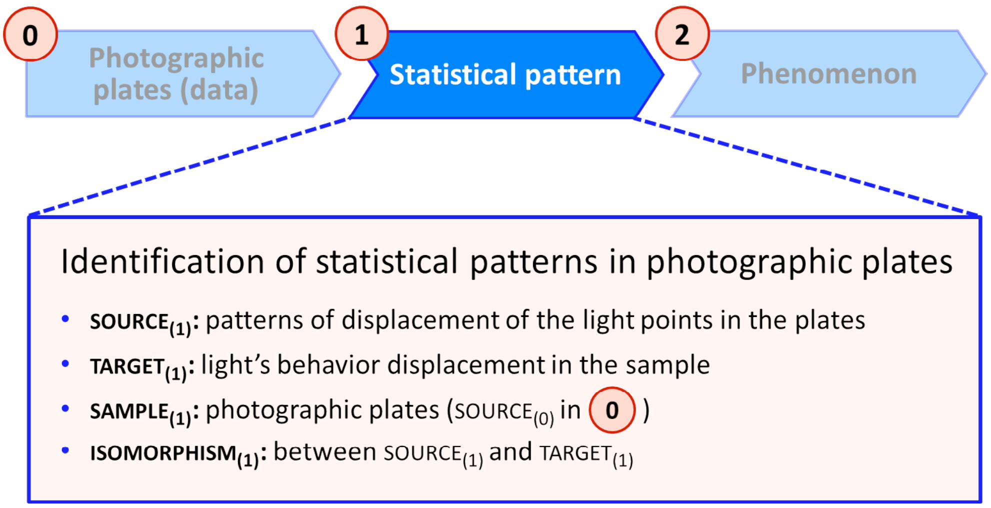

Figure 3

Isomorphism in the identification of statistical patterns. [Step 1]

Steps b) and c) in our characterization of the inference from data to phenomenon (see Section 2.2) have no direct equivalents in the case of Eddington's experiment, since no additional instances of light behavior during eclipses were available at the time. In this historical example, population patterns do not come into play to infer the phenomenon —namely, the deflection of light by the Sun. However, the extrapolation of the result is indeed pursued. In this case, the full generalizability of the detected phenomenon requires steps (b) and (c) to be carried out subsequently, once new samples become available.[5] This would allow the inference of a population pattern generalizable across cases of the same type, by extending the statistical patterns of displacement observed in the photographic plates to patterns of displacement expected in the entire population. In this second step, an isomorphism holds between the population patterns and the general phenomenon of light deflection beyond Eddington's specific observational context. Moreover, a homomorphism[6] is present between the statistical patterns and the population patterns, and it is this homomorphic relation that licenses the inference from source to target.

Returning to the actual development of the inference in our example, the next step involves inferring the cause of the sample pattern —namely, the phenomenon of light deflection as it travels near the Sun. In this case, the SOURCE$_{(2)}$ is a particular value of starlight deviation inferred from the statistical pattern (see Figure 4), while the TARGET$_{(2)}$ remains the same as TARGET$_{(1)}$— namely, the deflection of starlight as it travels near the Sun.

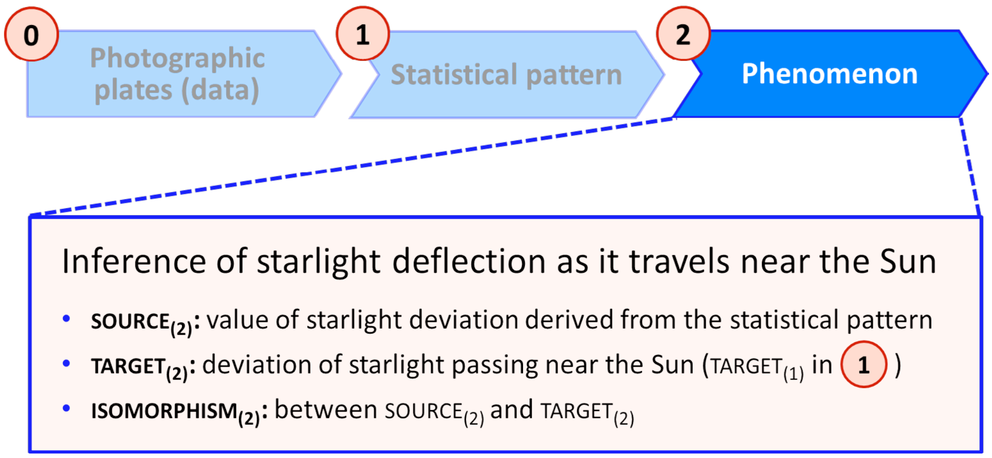

Figure 4

Isomorphism in the inference of starlight deflection when travelling near the Sun. [Step 2]

[5] The frequency of this type of case depends largely on the extent to which sample collection is constrained (i.e., on whether there are restrictions on extending the sample). When samples are readily available and affordable, the steps involving population patterns proceed as outlined in Section 2.2, that is, population patterns are inferred from sample patterns and serve as the basis for inferring the phenomenon. But even in the present case, the inferred particular phenomenon is considered as an instance of a general phenomenon that could be explained by general relativity.

[6] A homomorphism of systems is a weaker mapping relationship between two systems that preserves the structure of the source system in the target system. More formally, there is a homomorphism between two systems *S* and *T*, if $f: S \rightarrow T$ is a function that respects the relationships and operations defining the behavior and structure of system *S* within system *T*. On the other hand, an isomorphism is a homomorphism *f* that is injective (each component in *S* maps to a unique component in *T*), and surjective (each component in *T* has a preimage in *S*).

5. Conclusions

The preceding discussion emphasized the role of patterns as mediating in the inference from data to phenomenon. Our analysis reinforces the idea that data models are nested within background models that significantly influence data selection and processing —from assumptions about boundary conditions and control of hidden variables, to the statistical treatment required for pattern identification. We conclude that patterns play a crucial role in bridging data and the identification of phenomena, enabling surrogative inference through isomorphic relationships between source and target at each level of modeling.

Acknowledgments

Our work has greatly benefited from the helpful discussions we held in our reading group on Models of Data about 2024, especially those with Mauricio Suárez and Adán Sus. We would like to thank M. Suárez also for inviting us to contribute to the Symposium on Inference and Scientific Representation, at the *XI Congress of the Society for Logic, Methodology and Philosophy of Science in Spain*, where we got valuable feedback on our paper. Thanks also to Gui Sanches de Oliveira, as well as Aníbal M. Astobiza, for sharing interesting ideas on the subject of scientific representation.

This research was financially supported by the research project *The Philosophy of Statistical Modeling Science*, (PID2021-126416NB-I00, Spanish Ministry of Science and Innovation).

REFERENCES

Antoniou, A. (2021). What is a data model? *European Journal Philosophy of Science, 11*(4), 1-33. https://doi.org/10.1007/s13194-021-00412-2

Bogen, J. & Woodward, J. (1988). Saving the phenomena. *Philosophical Review, 97*(3), 303-352. https://doi.org/10.2307/2185445

Bokulich, A. (2018). Using models to correct data: Paleodiversity and the fossil record. *Synthese 198* (Supl. 24), 5919-5940. https://doi.org/10.1007/s11229-018-1820-x

Bokulich, A. (2020). Towards a taxonomy of the model-ladenness of data. *Philosophy of Science, 87*(5), 793-806. https://doi.org/10.1086/710516

Bokulich, A. & Parker, W. (2021). Data models, representation, and adequacy-for-purpose. *European Journal for Philosophy of Science, 11*(1), 31. https://doi.org/10.1007/s13194-020-00345-2

Castellani, E. & Schettino, G. (2023). Nested modalities in astrophysical modelling. *European Journal for Philosophy of Science, 13*, 11. https://doi.org/10.1007/s13194-023-00511-2

Collins, H. (1992). *Changing order: Replication and induction in scientific practice*. University of Chicago Press.

Earman, J. & Glymour, C. (1980). Relativity and eclipses: The British eclipse expeditions of 1919 and their predecessors. *Historical Studies in the Physical Sciences, 11*(1), 49-85. https://doi.org/10.2307/27757471

Glymour, B. (2000). Data and phenomena: A distinction reconsidered. *Midwest Studies in Philosophy, 31*(1), 184-201. https://doi.org/10.1023/A:1005499609332

Kennefick, D. (2009). Testing relativity from the 1919 eclipse —a question of bias. *Physics Today, 62*(3), 37-42. https://doi.org/10.1063/1.3099578

Kennefick, D. (2019). *No shadow of a doubt: The 1919 eclipse that confirmed Einstein's theory of relativity*. Princeton University Press. https://doi.org/10.1119/1.5123052

Laymon, R. (1982). Scientific realism and the hierarchical counterfactual path from data to theory. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association, 1*, 107-121. https://doi.org/10.1086/psaprocbienmeetp.1982.1.192660

Leonelli, S. (2009). On the locality of data and claims about phenomena. *Philosophy of Science, 76*(5), 737-749. https://doi.org/10.1086/605804

Leonelli, S. (2015). What count as scientific data? A relational framework. *Philosophy of Science, 82*(5), 810-821. https://doi.org/10.1086/684083

Leonelli, S. (2016). *Data-centric biology: A philosophical study*. University of Chicago Press. https://doi.org/10.7208/chicago/9780226416502.001.0001

Leonelli, S. (2019). What distinguishes data from models? *European Journal for Philosophy of Science, 9*, 22. https://doi.org/10.1007/s13194-018-0246-0

Massimi, M. (2007). Saving unobservable phenomena. *British Journal for the Philosophy of Science, 58*(2), 235-262. https://doi.org/10.1093/bjps/axm013

Mayo, D. G. (1996). *Error and the growth of experimental knowledge*. University of Chicago Press.

McAllister, J. W. (1997). Phenomena and patterns in data sets. *Erkenntnis, 47*(2), 217-228. https://doi.org/10.1023/A:1005387021520

McAllister, J. W. (2011). What do patterns in empirical data tell us about the structure of the world? *Synthese, 182*(1), 73-87. https://doi.org/10.1007/s11229-009-9613-x

Schindler, S. (2008). Model, theory, and evidence in the discovery of the DNA structure. *The British Journal for the Philosophy of Science, 59*(4), 619-658. https://doi.org/10.1093/bjps/axn030

Schindler, S. (2013). Theory-laden experimentation. *Studies in History and Philosophy of Science Part A, 44*(1), 89-101. https://doi.org/10.1016/j.shpsa.2012.07.010

Suárez, M. (2023). Stellar structure models revisited: Evidence and data in asteroseismology. In N. Boyd, S. de Baerdemaeker, K. Heng & V. Matarese (Eds.), *Philosophy of Astrophysics: Stars, Simulations and the Struggle to Determine What is Out There* (pp. 111-129). Springer. https://doi.org/10.1007/978-3-031-26618-8_7

Suárez, M. (2024). *Inference and representation: A study in modeling science*. University of Chicago Press. https://doi.org/10.7208/chicago/9780226830032

Suárez, M. (2025). Daniel Kennefick's No Shadow of a Doubt, *BJPS Review of Books*, 2025. https://doi.org/10.59350/thebsps.13234

van Fraassen, B. C. (1980). *The scientific image*. Clarendon Press. https://doi.org/10.1093/0198244274.001.0001

van Fraassen, B. C. (1989). *Laws and symmetry*. Oxford University Press. https://doi.org/10.1093/0198248601.001.0001

von Kluber, H. (1960). The determination of Einstein's light-deflection in the gravitational field of the sun. *Vistas in Astronomy, 3*, 47-77. https://doi.org/10.1016/0083-6656(60)90005-2

Woodward, J. (1989). Data and phenomena. *Synthese 79*(3), 393-472. https://doi.org/10.1007/BF00869282

Woodward, J. (2000). Data, phenomena, and reliability. *Philosophy of Science 67*, S163-S179. https://doi.org/10.1086/392817

Woodward, J. (2010). Data, phenomena, signal, and noise. *Philosophy of Science 77*(5), 792-803. https://doi.org/10.1086/656554

José V. Hernández-Conde is Associate Professor of Philosophy of Language at the University of Valladolid (Spain). In 2022, he was awarded a Leonardo Grant by the BBVA Foundation. His research focuses on the philosophy of language and the mind, as well as the philosophy of data science and artificial intelligence.

Address: Department of Philosophy, Faculty of Philosophy and Arts, University of Valladolid, Plaza del Campus s/n, 47011, Valladolid, Spain. E-mail: jhercon@uva.es – ORCID: 0000-0002-8502-6570

María Caamaño-Alegre is Associate Professor of Philosophy of Science at the University of Valladolid (Spain). She specializes in philosophy of science, with her research interests lying at the intersection of general philosophy of science, the methodology of science, the philosophy of language and epistemology.

Address: Department of Philosophy, Faculty of Philosophy and Arts, University of Valladolid, Plaza del Campus s/n, 47011, Valladolid, Spain. E-mail: mariaconcepcion.caamano@uva.es – ORCID: 0000-0002-7005-9257

Theoria, 2025, 40(2), 187-209
https://doi.org/10.1387/theoria.26061

THEORIA
ISSN 0495-4548 – eISSN 2171-679X

The role played by scientific realism in the measurement problem of quantum mechanics

(El rol que desempeña el realismo científico en el problema de la medición de la mecánica cuántica)

Juan Guzmán
Universidad Nacional Autónoma de México
https://orcid.org/0000-0001-7142-284X

Jorge Manero*
Universidad Nacional Autónoma de México
https://orcid.org/0000-0003-4588-5132

Keywords
Philosophy of quantum mechanics
Measurement problem
Semantic realism
Non-collapse interpretations

ABSTRACT: Many attempts have been made to characterise and solve the infamous measurement problem of quantum mechanics by advocating, implicitly or explicitly, different realist perspectives. As a result, we are still uncertain where this problem and its corresponding solution are to be located in the realism-antirealism debate. On the basis of a well-known characterisation of scientific realism, this paper intends to fill this gap by arguing that the quantum description of the processes involved in typical measurements is problematic from the standpoint of *semantic realism*, which is a necessary but not sufficient condition for fully-fledged scientific realism.

Palabras clave
Problema de la medición
Realismo semántico
Interpretaciones de no-colapso
Filosofía de la mecánica cuántica

RESUMEN: Se han hecho muchos intentos de caracterizar y resolver el problema de la medición de la mecánica cuántica adoptando, implícita o explícitamente, distintas perspectivas realistas. En consecuencia, aún no sabemos con certeza dónde situar este problema y su correspondiente solución en el debate realismo-antirrealismo. Con base en una conocida caracterización del realismo científico, este trabajo pretende llenar este vacío argumentando que la descripción cuántica de los procesos implicados en las mediciones es problemática desde el punto de vista del realismo semántico, el cual es una condición necesaria pero no suficiente para un realismo científico.

1. Introduction

It is well known that the feasibility of interpreting *quantum mechanics* (QM) realistically is controversial, in part because this theory faces numerous conceptual issues that cannot be addressed without directly participating in the realism-antirealism debate. Among these issues we find the *measurement problem*. Although alternative quantum theories and interpretations have been suggested in order to solve this problem, we do not yet have a definite picture of what ex-

* **Correspondence to:** Jorge Manero. Instituto de Investigaciones Filosóficas, Universidad Nacional Autónoma de México, Circuito, Mario de La Cueva s/n, Ciudad Universitaria, Coyoacán, 04510, Mexico City, Mexico – jorge.manero@filosoficas.unam.mx – https://orcid.org/0000-0003-4588-5132

How to cite: Guzmán, Juan; Manero, Jorge (2025). «The role played by scientific realism in the measurement problem of quantum mechanics»; *Theoria. An International Journal for Theory, History and Foundations of Science,* 40(2), 187-209. (https://doi.org/10.1387/theoria.26061).

Received: 19 February, 2024; Final version: 12 November, 2024.

ISSN 0495-4548 - eISSN 2171-679X / © 2025 UPV/EHU Press

actly might reconcile these alternatives with scientific realism. For example, beyond the standard way to formulate QM, including *unitary quantum mechanics* (UQM) and *standard quantum mechanics* (SQM), some of these alternative interpretations (to be addressed in this contribution) comprise *Frauchiger-Renner formulation* (FRQM), some *non-collapse interpretations* (NCQM), the *Bohmian approach* (BQM) and *objective collapse interpretations* (GRW). By UQM we shall mean the unitary theory that excludes from SQM (the one usually reviewed in textbooks) any reference to the collapse postulate but retains the completeness assumption with respect to the wave function. By FRQM we shall mean the theory associated with a particular interpretation of (von Neumann, 1932) that, to the best of our knowledge, forms the basis of the extended Wigner's friend experiment proposed by (Frauchiger and Renner, 2018). Finally, by NCQM we shall mean the set of non-collapse theories that are both unitary and include some interpretation of QM, such as the many-worlds interpretation, relational quantum mechanics, the modal interpretation and the consistent histories approach. Note that these interpretations share with UQM the completeness assumption and the wave function unitary evolution, but they are different in the sense that they deny the obtaining of definite and single outcomes by means of different methods (to be described in this contribution).

Through a critical revision of how these quantum theories and interpretations have characterised the measurement problem and/or supposedly solved it, our main contribution consists in understanding in what precise realist sense this problem associated with QM is a problem for quantum philosophers. Considering that UQM, SQM, FRQM and NCQM entail serious concerns regarding objectivity, understanding, and predictive reliability in our theoretical dealings with phenomena, we seek to demonstrate that, to put it crudely, *the measurement problem is a problem for semantic realism*, which is a necessary but not sufficient condition for fully fledged scientific realism. Since semantic realism is compatible with (but not reducible to) realism and (or) antirealism, we conclude that the measurement problem is neutral with respect to these philosophical positions. This does not mean, however, that the measurement problem is independent of the realism-antirealism debate; rather, this problem presupposes certain realist commitments, such as semantic realism, that cannot be appreciated if one insists in endorsing the endpoints of a more complex, continuum realist spectrum.

Our methodological procedure goes as follows: in Sections 2 and 3 we lay down an appropriate characterisation of scientific realism based on physics. Considering a well-known description of the measurement process, in Section 4 we proceed to identify some interpretations compatible with QM without having any a priori preference towards one of them. Once we have exhausted the most reasonable possibilities, in Section 5 we analyze the sense in which they are problematic. In so doing, we support the claim that the measurement problem is a semantic realist struggle by identifying the underlying realist criteria that are at stake in each of these interpretations. Finally, in Section 6, we conclude that these interpretations are problematic in the sense that they fail the test of instrumental empirical adequacy or only/and ontological clarity.

2. *The scientific realism triad*

Following the standard literature in the field (van Fraassen, 1980; Psillos, 1999; Chakravartty, 2017), physics-based scientific realism may be generally defined in terms of the conjunction of three necessary components.[1] In addition to the well-known mind-independent condition (i.e., a necessary condition for fully fledged scientific realism that shall be taken for granted), the semantic component (hereinafter *semantic realism*) is, put simply, the view concerning how physical theories are to be interpreted and what they are supposed to be saying without assuming that they are correct, whilst the epistemic component (hereinafter *epistemic realism*) states, in addition to this semantic correspondence, that what these theories say about the world is approximately correct.

1 NB: We focus only on the way these authors define scientific realism, regardless of whether or not they endorse it.

As introduced by (Feigl, 1950; Horwich, 1982; van Fraassen, 1980) and later defined by (Psillos, 1999, p. 10), semantic realism is associated with the realist's desire to read physical theories literally, in the sense that the truth conditions of theoretical assertions are specified, not to be confused with epistemic realism, which consists in specifying their truth values (i.e., to assert whether or not the truth conditions obtain) (Psillos, 1999, p. 10).[2] From the point of view of object-oriented realism, the semantic component stipulates that the theoretical terms of physical theories have the sole function of denoting putative objects, i.e, informing us what the world would be like according to these theories, *as if they were correct*, whilst the epistemic component states, in addition to this semantic correspondence, that theoretical terms approximately refer to the actual objects that constitute the world.

Although we accept that the use of the word 'semantic realism' could generate a linguistic confusion, it should be stressed that this semantic component is compatible with anti-realism. This is because this component is a necessary but non-sufficient condition for fully-fledged scientific realism. Let us note that, by definition, antirealism denies fully-fledged scientific realism. Since fully-fledged scientific realism necessarily assumes the metaphysical, semantic and epistemic components, antirealism must deny one (or more) of these components. Granted this, someone of antirealist spirit might embrace certain realist commitments, such as the semantic component, without assuming the epistemic component, and thus without collapsing to fully-fledged scientific realism.[3] As a result, the semantic component of scientific realism can be endorsed irrespective of whether we are realists or antirealists.

In the following section, we shall expand the definition of semantic realism by identifying additional semantic criteria against which our best physical theories (QM in particular) are to be evaluated. Since semantic realism has not been a point of special focus for philosophers of physics, we believe that they will find this way of expanding its definition both appropriate and fruitful.

3. *Scrutinising semantic realism*

According to our own reading of the above definition, we propose to identify two criteria that align with the semantic component of scientific realism: *ontological clarity* and *empirical adequacy*.

3.1. Ontological clarity

Ontological clarity means that the realist must specify what exactly the theory in question says (i.e., what its theoretical terms designate), in the sense that the realist must describe the scientific world depicted by the theory in terms of an ontology that has no considerable traces of semantic ambiguity. The association of this criterion with the semantic component of realism follows from conventional literature in the field, and in particular, from a proper reading of (van Fraassen, 1980). Saying, in van Fraassen's words, that "a literal construal of a theory can elaborate by identifying what theoretical terms designate", provided that "on a literal construal, the apparent statements of science really are statements, capable of being true or false" (van Fraassen, 1980, p. 10) is another way of saying that ontological clarity

2 It must be stressed that without the semantic component, there cannot be an epistemic component. Indeed, although literal construals of theories are not the same as truth-value construals, it is necessary to have a literal interpretation of the theory just to specify the conditions that make its assertions true (van Fraassen, 1980).

3 Note that this is precisely what van Fraassen does in order to motivate his constructive empiricism. Indeed, he insists that "not every philosophical position concerning science which insists on a literal construal of the language of science is a realist position" (van Fraassen, 1980, p. 11). To arrive at this conclusion, he divides the anti-realists into two sorts: "The first sort holds that science is or aims to be true, properly (but not literally) construed. The second holds that the language of science should be literally construed, but its theories need not be true to be good." Based on this distinction, he makes clear his position by saying "The anti-realism I shall advocate belongs to the second sort" (van Fraassen, 1980, p. 9-10).

https://doi.org/10.1387/theoria.26061

is, in part, a necessary condition for semantic realism. In particular, it is a condition that makes semantic realism different from epistemic realism because a literal construal is not related to "our epistemic attitudes towards theories, nor to the aim we pursue in constructing theories, but only to the correct understanding of what a theory says" (van Fraassen, 1980, p. 11). Of course, as is the case for any kind of theoretical virtue, how to specify in any particular situation the precise conditions for ontological clarity to occur, is controversial. However, if we are intending to define ontological clarity from the perspective of physics, it can be conceived as providing a minimal characterisation of the ontological profile of the physical states of the theory, or what is frequently called the 'ontic state' of the theory. More specifically, we should clarify whether or not the mathematical objects defining the physical states denote something in the physical world, and if they do, what kind of entities they are supposed to be denoting.

3.2. Empirical adequacy

Empirical adequacy means that the semantic realist also expects that the relevant empirical data, confined to the domain in which a physical theory can be effectively applied, can be explained in terms of the ontology in question. As was the case for ontological clarity, the latter definition of empirical adequacy is also prone to further concerns. More specifically, we should try to clarify the notions of 'empirical data' and 'explanation' involved in this definition.

Firstly, note that the primary set of empirical data that scientists usually read from their complicated measuring devices before any kind of systematisation process takes place is a regularity occurring in the manifested world. Although this set of data represents information about the degrees of freedom of the system under analysis, this information is ultimately displayed in the three-dimensional macroscopic world in the form of some physical object and property, such as a 'pointer' heading in a certain direction. This presupposes a broader conception of empirical data that comprises not only quantum phenomena manifested at the microscopic (and perhaps the mesoscopic) level, but also macroscopic phenomena of everyday experience (including the macroscopic properties of measurement devices). Under this broader conception, we can redefine empirical adequacy as a capability to explain the macroscopic manifestations of the relevant experimental data successfully predicted by the theory in terms of its posited (or elaborated) ontology. Furthermore, empirical adequacy also presupposes a nomological conception of empirical data, which comprises not only phenomena that have been experimentally detected, but also physically possible phenomena impossible to be detected in actual epistemic conditions (including thought experiments allowed by the physical theory).

Secondly, the notion of explanation involved in the definition of empirical adequacy has to do with what is known as the 'macro-object problem', namely, the problem of accounting for how the qualities of macroscopic or observable objects (such as the colours of observable tables) are nothing but manifestations of the qualities of objects (such as electromagnetic waves interacting with tables made of atoms) that are ultimate and fundamental, out of which the actual structure of the world is constituted. To address this problem in the context of physics we have to fulfil two requirements, one technical, the other (metaphysically) explanatory.

In the first place, we must show that any physical theory, if it deserves to be called a scientific theory, satisfies what has been called the 'consistency constraint' (Lam and Wüthrich, 2018). This technical requirement consists in 'saving the phenomena', namely, in identifying the theoretical basis required to make the predictions open to confirmation by a process of measurement. So, we should be capable of representing the underlying physical processes involved in typical measurements (and the outcomes produced by such measurements) in terms of the mathematical language of the theory. Considering the consistency constraint as an instrumental form of empirical adequacy (hereinafter, *instrumental empirical adequacy*), its satisfaction is just a matter of modelling a mathematical entailment between a list of data and a set of theoretical principles and postulates. Since this requirement bears no resemblance to the nature and characterisation of the underlying ontology of the theory, it is necessary but not sufficient to satisfy the empirical adequacy criterion.

On the other hand, we must satisfy an additional, different requirement: with the help of certain theoretical resources —and perhaps an ingenious metaphysical toolbox (e.g., supervenience, grounding and/or metaphysical dependence)—, we have to tell a consistent story of how the macroscopic world, such as tables and chairs, can be recovered from the fundamental ontology of the theory (Ney, 2015a,b; Lam and Wüthrich, 2018). This means that the macroscopic counterpart of the relevant experimental data successfully predicted by the theory must be cashed out in terms of such an ontology. Let us finally make two important remarks concerning this broader conception of empirical adequacy.

As regards the first remark, our definition of empirical adequacy is pragmatic and involves a non-factive notion of explanation. As required by semantic realism, scientific theories have to tell a literally construed story of what the world would be like according to these theories, as if they were correct. It follows that this story is conceived as 'purportedly' explanatory by virtue of the fact that it solves the macro-object problem in a way that exhibits some theoretical virtues, such as ontological clarity, although it could be wrong (and therefore might not be the actual explanation for the phenomena). This means that we are understanding purported explanations as propositions that are not only capable to successfully describe phenomena and generate novel and successful predictions —in compliance with the first instrumental requirement. As emphasised by (Boge, 2023), they are also capable of ensuring predictive success on account of the ontology of the theory— in compliance with the second ontological requirement.[4] This pragmatic notion of explanation contrasts with a factive one by virtue of the fact that semantic realism is associated with the former, whilst fully-fledged scientific realism is associated with the latter. Indeed, once we assume that scientific theories should tell a literally construed, approximately true story (i.e., if we assume the epistemic component), what we mean by explanation changes from being pragmatic to being factual, in the sense that this story involves true propositions, all of which are (part of) the reason why the predicted phenomena occur.

As regards the second remark, this definition is interpretative-dependent. Note that the second requirement is related to the clarity criterion because we have to specify a clear ontology before specifying how this ontology can comprehensively recover the observed phenomena. Without the specification of a precise ontology, we would not be able to tell how the physical world is according to this theory. And since this theory would be imprecise as regards its ontology, it also would have nothing to say about let alone how the macroscopic world can be recovered from the ontology of the theory. In this way, the clarity criterion is a necessary condition for the empirical adequacy criterion (understood in the broad metaphysical sense).

We shall now proceed to briefly sketch the formal-interpretative structure that lies behind the measurement problem irrespective of the philosophical position endorsed. After so doing, we shall investigate whether or not all the realist components are satisfied in the face of different QM interpretations, with the aim of articulating the exact connection between scientific realism and the measurement problem.

4. *The measurement process in quantum mechanics*

Based on a syntactic conception of scientific theories, many scholars make a sharp distinction between the mathematical formalism of a physical theory and its empirical and/or ontological interpretation. In the particular case of QM, this distinction is drawn by identifying a 'purely Hilbert space formalism', which is merely extensional or uninterpreted. However, we think that this way of characterising QM is misleading because even its 'axiomatisation', originally introduced by (von Neumann, 1932) but later refined by (Jammer, 1974), is already interpreted in terms of primitive notions, such as 'physical states' and 'observables'. Furthermore, based on (Maudlin, 2019), we could try to speak of this axiomatised theory as an operational recipe solely designed for predictive purposes, but this conception already presupposes that it contains an empirical interpretation impossible to be disassociated from its mathematical formalism. Con-

[4] Note that Boge emphasises the indispensable role of this second ontological requirement in the context of factive explanations —to justify the claim that scientific realism is only appropriate if science delivers true explanations. However, we can extrapolate his statement and claim that such a requirement also arises from non-factive, purported explanations in the context of semantic realism.

https://doi.org/10.1387/theoria.26061

sidering that it seems difficult to draw a line between a purely Hilbert space formalism and its corresponding empirical and/or ontological interpretation, we shall conceive of QM as a theory which comprises an interpreted mathematical apparatus and a mathematically-based interpretation. As a result, theory and interpretation are regarded as indistinguishable terms.

Granted this, we shall start our discussion by carefully identifying some possible quantum interpretations (without having any a priori preference towards one of them). In so doing, we shall consider Maudlin's method of characterising each of these interpretations depending on the acceptance or denial of certain interpretative assumptions associated with the measurement process. The form of these assumptions shall be expressed by means of truth-values assigned to the following claims:

(1) The wave function of a system is complete (i.e., specifies all of the physical properties of a system).

(2) The wave function always evolves in accordance with a linear, unitary dynamical equation (i.e., the Schrödinger equation).

(3) The measurement of a physical property always has definite and single outcomes.[5]

There are many ways one might accept of deny these claims, but not all combinations are acceptable. According to the the well-known account of the measurement problem proposed by (Maudlin, 1995), the formal-interpretative structure of the measurement process in QM can be presented in the form of a trilemma (hereinafter *Maudlin's trilemma*). This trilemma consists in the conjunction of the above (1)-(3) claims, all of which can be demonstrated to be mutually inconsistent. The proof of their mutual inconsistency will not be revised here as it is well known in the literature. The important point for us, however, is that Maudlin's trilemma consists in a series of inferences that allow us to identify certain interpretations compatible with QM, all of which arise from denying at least one of the above claims. For example, as we shall see below, both UQM and NCQM accept (1) and (2), but reject (3) by means of different methods.[6]

Understood in this way, Maudlin's trilemma cannot be considered problematic with the exception of evaluating the consistency of its formal-interpretative structure. This allows us to avoid the widespread mistake of interpreting the inconsistency between the above (1)-(3) claims as representing the measurement problem without even providing the conceptual basis of what really counts as problematic. If there is something like a problem within this trilemma beyond the alleged consistency, we should pay attention to the philosophical consequences of advocating any of the possible interpretations arising from fixing the truth-values of the above claims. In more precise terms, we need a philosophical standard from which one can evaluate such interpretations. It is precisely the identification of this standard with semantic realism that allows us to know in exact terms how exactly the measurement problem is associated with scientific realism. This will be the main task of the next section.

5. *The problem of the measurement process in the face of scientific realism*

Returning to our discussion, we shall now articulate the philosophical basis of the explanation of why and in what sense the incompatibility between completeness, unitary evolution, and definite-single values for measurement outcomes really constitutes a problem for QM.

[5] NB: Do not confuse the claim that measurements have single outcomes with the claim that they have definite outcomes. As we shall see below, the first claim is related to the problem of outcomes, whilst the second one with the problem of interference.

[6] Note that any pair of QM interpretations shall be differentiated, not only based on the denial of one of the (1)-(3) claims, but also depending on the methods used to deny such a claim. Furthermore, we shall see that there are interpretations, such as FRQM, that accept different combinations of the (1)-(3) claims by introducing different observers (each one which assigns different truth-values to these claims).

First and foremost, note that irrespective of Maudlin's own interpretation, claim (3) can be construed *empirically* as stating that the measurement of a physical property is always observed to have definite and single outcomes, or *ontologically* as implying that, after the process of a measurement, there always exist objective definite and single outcomes for such a physical property. Considering that the observability of definite and single outcomes upon a measurement is a necessary (albeit not sufficient) condition for their existence, we shall interpret claim (3) ontologically. The reason for this particular interpretation (and the drawing of this ontological/empirical distinction) is mainly that the denial of the existence of definite and single outcomes upon measurement is necessary if incorporating NCQM into Maudlin's framework, a set of interpretations that presumably solve the measurement problem and that we would like to critically discuss in this contribution.

As we shall see next, we believe that the incompatibility of claims (1)-(3) is problematic if we are willing to interpret QM realistically on the basis of the criteria associated with the semantic component of realism (i.e., empirical adequacy and clarity with respect to the ontic state of the system in question). Let us now justify this claim by analysing the interpretative consequences of abandoning each one of the three claims (1)-(3), whilst retaining some of the others. For illustrative purposes we shall consider each of the relevant combinations of denying or accepting the above claims without following a predefined order of exposition. Although this analysis might seem a recitation of previous work, we believe that the association of Maudlin's account of the measurement process with semantic realism has not been articulated. This association, as we shall see, shall permit us to characterise and address the corresponding problem from a broader, unambiguous standpoint.

5.1. The problem of instrumental empirical adequacy

In the following lines, we shall argue that the measurement problem is associated with the lack of instrumental empirical adequacy that prevails in UQM, FRQM and NCQM. Whilst we shall concentrate a little bit more on demonstrating the lack of the instrumental form of empirical adequacy within FRQM and NCQM, it is important to address this task within UQM (which seems to be trivial) as we have to frame the well-known measurement process associated with QM in terms of our definition of instrumental empirical adequacy.

5.1.1. The instrumental empirical inadequacy of unitary quantum mechanics

We shall demonstrate that if (3) is denied, whilst (1) and (2) are accepted, without embracing NCQM, we end up with a problem of instrumental empirical adequacy. This shall lead to the conclusion that UQM is empirically inadequate in the instrumental form.

Since by accepting (1) and (2) we are starting our analysis from the completeness assumption and the Schrödinger equation, it would be wise to ask what exactly these claims have to do with the empirical adequacy of the theory, and therefore, with the particular physical situation of a typical measurement. To answer this question, let us characterise the notion of empirical adequacy in the particular context of QM.

Assuming that QM aims to describe the macroscopic world in terms of the way it describes the microscopic world (including the experimental devices in measurement situations), empirical adequacy states that the theory should have an operational technical procedure required to account for the underlying physical, macroscopic processes involved in typical measurements (i.e., instrumental empirical adequacy); and that the unobservable, microscopic ontology posited by the theory (or elaborated by the realist) should explain the empirical observations successfully predicted by it. These two basic criteria might be satisfied using different strategies. While the first one just would need a set of operational rules that dictate how vectors and unitary operators in Hilbert space (and the way they evolve in time) make successful predictions about measurement outcomes (manifested in some way by the experimental device), the second one would require to figure out an appropriately grounded explanation (either physical or metaphysical) of how a set of definite and single outcomes of the observable properties associated with macroscopic objects (such as the direction of the pointer on the experimental device) correspond to the definite and single values of the properties associated with unobservable, microscopic objects (such as the spin of a particle, or the position associated with a system of particles, etc.).

However, the perfect demonstration that the instrumental criterion of empirical adequacy cannot be satisfied by UQM corresponds to the famous Schrödinger's cat experiment: UQM predicts that the quantum state of the whole system (represented by the wave function in a complete form), including the system under analysis (e.g., a radioactive atom) and the experimental device (e.g., a cat), evolves unitarily according to the Schrödinger equation to an irreducible superposition of macroscopic states, where each non-realised term correlates, in a post-measurement situation, a definite and single value of the microscopic properties of the system under analysis (e.g., the atom decaying or not decaying) with the definite and single value of the corresponding macroscopic properties of the measurement device (e.g., the cat being dead or alive). This means that whatever the unobservable objects of the system, the predicted measured outcome recorded by the experimental device is undefined because it corresponds to a superposition of macroscopic states. Since every time a physical property is actually measured a definite and single outcome can be read from the experimental device in question, there is just a small step from this empirical fact to the conclusion that UQM is empirically inadequate in the instrumental form, in the sense that it does not have an operational procedure to explain the measurement process of systems in states of superposition. Therefore, since the semantic component of realism requires that UQM satisfies the empirical adequacy criterion in the instrumental form, and this criterion is not satisfied by it, then we are led to conclude that the acceptance of (1) and (2) and the abandonment of (3) is problematic by virtue of the fact that UQM cannot be interpreted realistically, at least in this semantic sense.

5.1.2. The instrumental empirical inadequacy of Frauchiger-Renner formulation

Let us now address the most interesting situation. We shall demonstrate two conditionals at once:

> If claim (2) is denied, whilst (1) and (3) are retained, without embracing SQM (assuming the projection postulate) or GRW (adding non-linear terms to the dynamics), we end up with a problem of instrumental empirical adequacy.
>
> If claim (1) is denied, whilst (2) and (3) are retained, without embracing BQM (adding hidden variables to the physical state), we end up with a problem of instrumental empirical adequacy.

Based on ideas originally developed by (von Neumann, 1932) and (London and Bauer, 1939), we shall define Frauchiger-Renner formulation (FRQM) as the theory that introduces into QM a distinction between two different points of view from which one describes the measurement process (hereafter *Neumann's observer distinction*). As originally argued by (von Neumann, 1932), there is either the *external* descriptive view, according to which the observer or the experimenter is treated in the same way as the system under analysis, evolving both by the Schrödinger equation —and inducing macroscopic superpositions due to its unitary evolution—, or the observer's descriptive *internal* view, in which the system in question suddenly collapses, after a measurement is performed, to only one single term in the superposition —and obtains definite and single outcomes as a consequence.[7] As we shall elaborate in this section, this means that FRQM implicitly presupposes that there are external observers that deny (1) and accept both (2) and (3), but also there are internal

7 Note that when a measurement is performed inside of a closed lab, this premise is sometimes expressed in terms of a distinction between internal and external observers, provided such a measurement leads to a collapse for the former observers but does not lead to a collapse for the latter observers. However, this way of making the aforementioned distinction is misguided in general: there can be internal observers experiencing unitary evolution (inevitably leading to macroscopic superpositions) in the same way that external observers experiencing collapses (from which definite and single outcomes arise). Granted this, the difference in notation between 'internal' and 'external observers' shall mean a distinction among observers who differ in relation to whether or not any measurement leads to a collapse, irrespective of whether they are inside or outside of a lab.

observers that deny (2) and accept both (1) and (3).[8] As such, FRQM is different from SQM and NCQM. As we shall see later, SQM only accepts internal observers that deny (2), whilst FRQM also accepts external observers that deny (1). Similarly, NCQM only accepts external observers that deny (3), whilst FRQM accepts internal observers that deny (2) and external observers that deny (1). Granted this, we shall conclude that FRQM is empirically inadequate in the instrumental form. Let us justify this claim by considering internal observers first, and then proceed to introduce external observers (in addition to the internal ones).

First and foremost, internal observers deny claim (2) by introducing an additional postulate into QM, called the *projection postulate*. This collapse mechanism somehow makes the macroscopic superposition "disappear", such that one can effectively predict that the final state of the system after the measurement takes place is represented by the eigenvector corresponding to the measured outcome. At first sight, this strategy seems to be appropriate to satisfy the instrumental empirical adequacy criterion. Although this is true in typical cases, such as the ones with Schrödinger's cats, we believe that the first instrumental requirement of empirical adequacy cannot always be satisfied even with the introduction of the projection postulate. There is a concrete situation in which this scenario occurs: if we cannot say anything about the nature and behaviour of the physical system in question until it is observed then what has happened in the past ends up being beyond the descriptive scope of the theory. This situation, apart from being at odds with classical accounts of causality,[9] drastically reduces the empirical domain of the theory to the present state of affairs. Thus, FRQM can only be interpreted as a predictive recipe effectively applied to measurement situations, implying some temporal limitation in the way FRQM can be empirically adequate in the instrumental form.

Secondly, recent thought experiments have demonstrated that the projection postulate, originally introduced by internal observers, is not always an appropriate tool for prediction because there are concrete physical systems with internal *and* external observers upon which the application of this formal apparatus leads to incorrect or undefined predictions.[10] This shall lead to the conclusion that FRQM is empirically inadequate in the instrumental form. In order to arrive at this conclusion, we shall appeal to a non-standard interpretation of a well-known *extension of the famous Wigner's friend experiment*. In so doing, we shall first introduce the mainstream interpretation of this thought experiment (together with a brief description of its original interpretation). After realising that the outcome of this experiment is irrelevant from the point of view of the mainstream interpretation, we shall try to read this experiment in a different, non-standard way in order to arrive at our desired conclusion.

Originally introduced by (Frauchiger and Renner, 2018), this controversial thought experiment (hereinafter the *FR experiment*) is cast in the form of a no-go theorem (hereinafter the *FR theorem*) that mainly states that no theory that is fully compliant with the predictions of QM can at once: (i) capture the universal validity of this theory (including the macroscopic level); (ii) assign a single outcome to each measurement; and (iii) demand consistency between different observers. The moral of this argument, according to (Frauchiger and Renner, 2018), is that the mutual incompatibility of these assumptions puts forward the claim that QM "rejects a single-world description of physical reality" (Frauchiger and Renner, 2018, p. 3).

---

8 It is important to note that, according to our own reading of (Frauchiger and Renner, 2018), they assign truth-values to (1)-(3) claims without being aware that these claims are implicitly presupposed by them.

9 Some would justify this situation by assuming that the realisation of present facts causes facts in the past to be realised (an effect precedes its cause in time). This unfamiliar account of causality, called 'retrocausality', has been plainly discussed in (Faye, 2018).

10 It is important to point out that this demonstration has been recently objected by (Okon, 2021). Although he agrees that the projection postulate is not an appropriate tool for prediction for reasons concerned with its ambiguous nature, he argues that the conclusion arising from such thought experiments assume an incorrect assumption and this makes any prediction required by this conclusion unreliable. Since this assumption is not correct, according to him, there is no reason to conclude that the application of the projection postulate leads to inconsistent and incorrect predictions; on the contrary, it is because this postulate is so ambiguous that we mistakenly embrace incorrect assumptions and cannot even make predictions among these experiments. Although we think that Okon's objection should be discussed in depth, we shall assume that the projection postulate applied in the context of these thought experiments leads to inconsistent and incorrect predictions.

However, as critically discussed by numerous scholars (Sudbery, 2019; Lazarovici and Hubert, 2019; Tausk, 2019; Muciño and Okon, 2020; Boge, 2019), the FR theorem's conclusion relies on a fourth silent, implicit assumption, which considerably reduces the scope of the theories ruled out by the FR theorem. Although this theorem is still a subject of controversy, there is a shared hypothesis accepted by (almost) the academic community that the alleged fourth silent assumption presupposed by the FR theorem corresponds to the aforementioned *Neumann's observer distinction*. Since this assumption is obviously compatible with FRQM, this interpretation together with any other (still unknown) interpretation that accepts (i)-(iii) together with the fourth silent assumption are ruled out by the FR theorem. In more precise terms, the point emphasised by these scholars is that this distinction presupposed by the FR theorem induces empirical differences between successive predictions obtained by internal and external observers. The reasoning is that if external observers would like to obtain further predictions after internal observers have performed their own measurements, and the former observers evolve the wave function according to Schrödinger's equation and no term ever disappears, surely they must obtain different predictions than the latter observers would obtain applying the projection postulate (in which case some terms disappear).

In sum, according to these scholars, (Frauchiger and Renner, 2018) mistakenly believe that any QM interpretation compatible with premises (i)-(iii) leads to inconsistencies. Consequently, what (Frauchiger and Renner, 2018) are implicitly showing (without knowing it) is that FRQM (i.e., the only interpretation that assumes the (i)-(iv) premises) leads to inconsistencies. From this point of view, it seems that the result of the FR theorem is not so relevant as the only interpretation that is ruled out by it is FRQM, as opposed to any other well-known interpretation that denies some of the (i)-(iv) premises, such as UQM, SQM, NCQM, BQM and GRW. Under these circumstances, we think that the FR theorem can be read in a different, more interesting way, and can proved to be relevant for other purposes. From this alternative perspective, the point of introducing FRQM is not to rewrite the FR theorem as it is interpreted in the mainstream literature, namely, that FRQM leads to inconsistencies. It would make no sense to introduce FRQM in order to repeat an argument against it without having any relevant consequences. On the contrary, our aim is to put forward the claim that the measurement problem is a problem for FRQM too. Indeed, we seek to demonstrate that FRQM fails the test of instrumental empirical adequacy, provided the measurement problem is a problem for any interpretation incompatible with this criterion. To do so, let us describe the FR theorem as it is discussed (Bub, 2017), a version which will help us to better address the point at issue.

Suppose there are two observers at a great distance from each other, Alice A and Bob B, both of whom find themselves inside their respective laboratories, A inside L and B inside L'. Additionally, there are two observers X and Y outside the respective laboratories, whom can perform measurements on them. For ease of notation, the irrelevant degrees of freedom of all the observers are not considered and the experimenters' environment is excluded from the description, such as measurement devices and laboratory equipments. Before the experiment, A prepares a coin C in the initial state $\sqrt{1/3}\,|h\rangle_A + \sqrt{2/3}\,|t\rangle_A$.

Initially, at time $\tau = 0$, A measures C in the $\{|h\rangle_A, |t\rangle_A\}$ basis and records the result "h" or "t". At time $\tau = 1$, she then prepares a qubit Q as follows: if the result at $\tau = 0$ came out to be h, A prepares Q in the state $|0\rangle_B$; if it rather came out to be t, she prepares it in the state $\frac{1}{\sqrt{2}}(|0\rangle_B + |1\rangle_B)$. A subsequently sends Q to B by a communication channel from L to L'.

At time $\tau = 2$, B measures Q in the $\{|0\rangle_B, |1\rangle_B\}$ basis and records the result "0" or "1". At subsequent times, there are two possible scenarios:

The first scenario is such that at $\tau = 3$ the observer X measures A's laboratory L first in the basis:

$$|f\rangle_L = \frac{1}{\sqrt{2}}(|h\rangle_A + |t\rangle_A)$$

$$|o\rangle_L = \frac{1}{\sqrt{2}}(|h\rangle_A - |t\rangle_A)$$

and records the result "o" or "f", whereas at $\tau = 4$, the observer Y measures B's laboratory L' in the basis:

$$|f\rangle_{L'} = \frac{1}{\sqrt{2}}(|0\rangle_B + |1\rangle_B)$$

$$|o\rangle_{L'} = \frac{1}{\sqrt{2}}(|0\rangle_B - |1\rangle_B)$$

and records the result "o" or "f". On the contrary, the second scenario is such that the observer Y measures first and precedes the measurement of the observer X. At the end of the experiment, the possible results of the measurements performed by each observer are compared. It turns out that a contradiction analogous to (Frauchiger and Renner, 2018) arises from this comparison. However, this contradiction is a conclusion derived from a series of premises that need to be explicated. We shall begin presenting a systematic formulation of the problem by enumerating both the explicit and hidden assumptions of the FR theorem in the following manner:

> *Premise (i)*. QM is universally valid and describes both micro and macroscopic systems.
>
> *Premise (ii)*. Measurements have definite and single outcomes.
>
> *Premise (iii)*. if an observer X finds that another observer Y is certain about a measurement outcome, then X must also be certain about such a measurement outcome.
>
> *Premise (iv)*. When a measurement is performed, such a measurement leads to a collapse for some observers, but it does not lead to a collapse for other observers.

The contradiction arises in the following manner. Let us first consider (i) and (ii). The system of both laboratories L and L', composed of A, B, C and Q, can be described by the following quantum state:

$$|\Psi\rangle = \frac{1}{\sqrt{3}}(|h\rangle_A|0\rangle_B + |t\rangle_A|0\rangle_B + |t\rangle_A|1\rangle_B) \tag{1}$$

This expression is the state viewed from the perspective of any observer located outside the laboratories (X or Y) without having described their own measurements. If we consider the measurements performed by X in L and Y in L', (1) can be expressed as follows:

$$|\Psi\rangle = \frac{1}{\sqrt{12}}|o\rangle_L|o\rangle_{L'} - \frac{1}{\sqrt{12}}|o\rangle_L\,|f\rangle_{L'} + \tag{2}$$

$$+\frac{1}{\sqrt{12}}\,|f\rangle_L|o\rangle_{L'} + \frac{1}{\sqrt{12}}\,|f\rangle_L\,|f\rangle_{L'}$$

$$= \sqrt{\frac{2}{3}}\,|f\rangle_L|0\rangle_B + \frac{1}{\sqrt{3}}|t\rangle_A|1\rangle_B \tag{3}$$

$$= \frac{1}{\sqrt{3}}|h\rangle_A|0\rangle_B + \sqrt{\frac{2}{3}}|t\rangle_A\,|f\rangle_{L'} \tag{4}$$

Before any measurement is performed, the purely unitary evolution of the total composite system (i.e., X, Y, A, B, C, and Q) evolves to a superposition expressed by (2), (3), and (4). Note that we can combine the results obtained by different observers at different times in the following way:

https://doi.org/10.1387/theoria.26061

From (3) we can see that the relative state of *X* and *Q* with respect to the observer *X* is $|o>_L\ |1>_B$ with the pair of possible values *{o,* 1*}* (as a result of measuring *L* and obtaining the outcome "*o*"), whereas from (4) we can see that the relative state of *C* and *Y* with respect to the observer *Y* is $|h>_A\ |o>_{L'}$ with the pair of possible values *{h, o}* (as a result of measuring *L'* and obtaining the outcome "*o*"). It follows that the pair of actual outcomes *{h,* 1*}* are found only once the pair *{o, o}* is obtained when *X* and *Y* jointly measure *L* and *L'*, respectively.

In this step we have used (iv), to assume that *X* and *Y* simultaneously obtained definite and single outcomes, so that we can predict future states of affairs conditioned by *XY*'s actual measurement outcome *{o, o}*; and also (iii), to be allowed to juxtapose *X* and *Y* to give us the perspective that consists of both observers *XY*. This latter assumption means that if *Y* [*X*] finds that the *L*[*L'*] measurement outcome is "*o*" from the perspective of the *X* [*Y*] observer, this outcome has also to be "*o*" as judged from the perspective of the *Y* [*X*] observer, allowing us to observe the composed system *LL'* from the *XY* juxtaposed observer. Finally, only if we introduce preceding information of the state (1), even after *A*, *B*, *X* , and *Y* have measured *C* , *Q*, *L*, and *L'*, obtaining as a result definite and single outcomes (e.g., *h* or *t* for A; 1 or 0 for *B*; and "*o*" or "*f*" for *X* and *Y*), we obtain a contradiction: the state (1) shows that the probability of obtaining *{h,* 1*}* is zero.

Note that (iv) played an indispensable role in the previous argument because the final pair of outcomes *{h,* 1*}* is a future prediction that was obtained only when the observer *XY* used the collapsed state $|o>_L\ |o>_{L'}$ corresponding to what she found, not the full, unitary-evolving superposition (2). At the same time, however, we considered the superposition (1), corresponding to the unitary-evolving state of *A*, *B*, *C* , and *Q* before any measurement takes place, even after the observer *XY* performed a measurement and found *{o, o}*.

Since FRQM assumes (iv) and the rest of the above premises, we may conclude that it is ruled out by the FR theorem. As regards the instrumental empirical adequacy of this interpretation, this conclusion brings important consequences. Note that the claim that FRQM leads to inconsistent predictions (as opposed to other interpretations, such as BQM) can be reinterpreted as stating that this collapse approach simply cannot make the correct quantum predictions in this experiment. Examples of these incorrect predictions are precisely those that were followed in the derivation of the contradiction, such as the possibility of predicting the final outcome *{h,* 1*}* if the combined observer *XY* measures the outcome *{o, o}*. In the framework of BQM, however, the outcome "*o*" of *Y* is not only compatible with A's outcome "*h*", as used in the FR theorem. Rather, the final outcome *{t,* 1*}* is also possible. Let me elaborate more on this.

If there is an ontology constituted by space-time events in a non-relativistic context, such as configurations of Bohmian particles corresponding to *X*, *Y*, *A*, *B*, *C*, *Q* and their respective settings and outcomes, the predictions obtained by each observer should, of course, respect the time order of those events. In the case in which observer *X* measures first at time $\tau = 3$ and *Y* performs the second measurement at time $\tau = 4$ (provided *A*'s and *B*'s measurements have been performed at time $\tau = 0$ and $\tau = 2$, respectively), *X* will (locally) influence the subsystem composed of *A* and *C* in a way that *A*'s initial state and the corresponding result (either *h* or *t*) could be changed. Since the state of *A* and *B* are correlated nonlocally, if the result of *X* turns out to be "*o*", it is a fact of the matter that *B*'s result must be 1. This 1 will remain invariant until the second measurement of observer *Y* is performed. This second measurement could (locally) influence and change *B*'s state and its corresponding result (either 1 or 0). Now, in this story it is not true that the result "*o*" of *Y* is only compatible with *A*'s result *h*, as it would be in the case in which observer *Y* measures first and *X* performs the second measurement. In sum, in the case in which *Y* performs her measurement first, *Y*'s outcome is independent of *X*'s setting and outcome; in the case in which *X* performs his measurement first, however, *Y*'s outcome depends on *X*'s setting and outcome.

Therefore, since BQM is not ruled out by the FR theorem and does not lead to inconsistencies, we may reinterpret the inconsistencies arising from FRQM as the failure of this approach to make some correct quantum predictions. This makes FRQM theory empirically inadequate in the instrumental form. That is, FRQM is ruled out by the FR theorem by virtue of the fact that it cannot account for FR experiment's predictions correctly obtained by, for example, BQM.

Considering this line of reasoning, one might wonder whether the hidden assumption (iv) is the ultimate reason FRQM cannot make the correct predictions in the FR experiment. Following a similar argument to (Okon, 2021) (although not the same conclusion), we believe that there is a deeper reason that lies behind (iv) that is responsible for the instrumental empirically inadequacy of FRQM in the context of the FR experiment. This underlying reason is associated with the fact that *X* or *Y* does not possess the resources to predict the complete outcomes obtained by *A* and *B* after the outcomes obtained by the juxtaposed observer *XY* are determined. Since the state (1), as defined from the point of view of *X* or *Y* (without having described their own measurements), evolves according to the linear Schrödinger evolution, and all observers (i.e., *X*, *Y*, *A* and *B*) obtain definite and single outcomes, Maudlin's trilemma tells us that FRQM is incomplete. In other words, the state (1) cannot specify all the physical properties of *A* and *B*. It is expected that the lack of specification by the wave function of the actual physical properties associated with these subsystems induces the alleged incorrect predictions and explains why BQM, which is a complete theory, makes the correct predictions.

This result reflects the fact that in the FR experiment there are not only internal observers that deny (2) and accept (1) and (3), but also there are external observers that deny (1) and accept (2) and (3). In other words, FRQM is empirically inadequate in the instrumental form precisely because there are experimental situations in which it abandons the completeness assumption.

5.1.3. The instrumental empirical inadequacy of non-collapse interpretations

We shall now demonstrate that if (3) is denied, whilst (1) and (2) are retained, we end up with the same problem of instrumental empirical adequacy (as corroborated by the previous discussion) unless we explain the empirical fact that measurement outcomes *seem* to be definite and single-valued even if macroscopic superpositions exist and never really collapse, an explanation which will be shown to be unsuccessful. Considering that claim (3) presupposes the existence of (as opposed to the observation of apparent) definite and single outcomes, its denial leads therefore to the adoption of NCQM.[11]

One might argue that denying claim (3), whilst retaining (1) and (2), is one exception to the conclusion that QM is empirically inadequate in the instrumental form. This exceptional case would derive from simply denying the existence of definite and single outcomes and accepting the existence of macroscopic superpositions, and therefore concluding that the theory is empirically adequate in the instrumental form because it naturally accounts for actual superpositions. However, the instrumental empirical adequacy of a theory cannot be derived from simply assuming that it accounts for something that exists but is impossible to detect in everyday experience, such as macroscopic superpositions. Let us appeal to an example which corroborates this claim.

Based on numerous recent 'Schrödinger cat states' experiments, some philosophers have proffered NCQM according to which macroscopic (in addition to microscopic) superpositions are objective physical states which never really collapse. Examples of NCQM are the many-worlds interpretation, relational quantum mechanics, the modal interpretation and the consistent histories approach. However, although the way to account for the reality of these macroscopic superpositions varies across the board, the common feature among these interpretations is that they believe that measured outcomes are not actually definite and single-valued even if we observe that they seem to be definite and single-valued in everyday experience. That is, the price to be paid for denying claim (3) whilst retaining (1) and (2) is that there should be a way to explain the observable fact that macroscopic superpositions are impossible to detect in everyday experience and, when the properties of a physical system are measured, measurement outcomes *seem* to have definite and

[11] NCQM will be defined here simply as those interpretations which deny the existence of definite and single outcomes but retain (1) and (2). In this way, we shall include neither BQM nor the Hamiltonian modal interpretation as part of what we call NCQM because these interpretations accept the existence of definite and single outcomes through the introduction of a preferred set of commuting observables (i.e., the position and the energy, respectively).

single values. This example demonstrates that even when the existence of macroscopic superpositions is assumed, whilst that of definite and single outcomes is denied, the challenge of instrumental empirical adequacy still prevails. Unfortunately, NCQM are not empirically adequate in the instrumental form, as we shall demonstrate next. We believe that NCQM are empirically inadequate in the instrumental form for at least two possible reasons. Whilst we accept that the first reason might be argued to be ungrounded for certain reasonable intuitions (to be described below), we believe that the second one poses a serious challenge to the instrumental empirical adequacy of these interpretations.

Firstly, one might argue that the FR theorem can also be applied to NCQM based on the possibility that these interpretations comply with all (i)-(iv) assumptions, and introduce internal observers that deny (3) and accept (1) and (2) in addition to external observers that deny (3) and (1) and accept (2). In a similar way to FRQM, this would lead to the conclusion that NCQM are ruled out by the FR theorem, and therefore are empirically inadequate in the instrumental form by virtue of the fact that they yield incorrect predictions. To arrive at this conclusion, one would need to focus on the way assumption (iv) is interpreted. Note that the aforementioned Neumann's observer distinction between different observers that lies behind assumption (iv) might be interpreted epistemically (as opposed to ontologically). Since macroscopic superpositions exist and never really collapse in NCQM, the fact that definite and single outcomes are observed might be interpreted as an apparent *epistemic* manifestation, not necessarily associated with *actual*, definite and single outcomes. This would mean that in the case of NCQM, this hidden assumption is not precisely aimed at reconciling unitary evolution with objective (actual) definite and single outcomes, such as our own reading of (Maudlin, 1995)'s analysis in denying completeness; on the contrary, it interprets both unitary evolution and the epistemic (empirical) fact that definite and single outcomes are seemingly observed as two different, non-reducible points of view. Therefore, the fourth silent assumption of the FR theorem would not only be compatible with particular, idiosyncratic interpretations of QM that believe in the existence of definite and single outcomes, such as FRQM, but also with some NCQM that explain in some way or other the appearance of (apparent) definite and single outcomes, such as relational quantum mechanics, the consistent histories approach and some modal interpretations.

Unfortunately, this conclusion might be argued to be ungrounded after a careful analysis. In fact, one might reply that the compatibility between NCQM and assumption (iv) is not sufficient to derive the conclusion that these interpretations are ruled out by the NCQM theorem. Rather, they should be compatible with the rest of the assumptions (i)-(iii). But we are currently uncertain of whether or not these interpretations accept the other assumptions. For example, based on an intuitive reading of some of these interpretations, one might argue that relational quantum mechanics denies (iii), whilst the consistent histories approach denies (ii), and therefore block the conclusion that they are ruled out by the FR theorem.[12] However, as things currently stand, we cannot undoubtedly conclude, with the exception of FRQM, that this is actually the case.

Secondly, even if we were to block the previous conclusion in solid ground, we would inevitably came to realise that NCQM are empirically inadequate in the instrumental form because they have not succeeded in solving at least one problem, the issue of selecting a single outcome upon the measurement of a physical property (hereinafter the *problem of outcomes*).

As regards the instrumental criterion of empirical adequacy, the principal challenge is to provide an empirical interpretation of the quantum formalism without collapses associating mathematical representations (e.g., vectors and operators) with the empirical data. To have this interpretation, however, the purely unitary evolution of this formalism should not only be able to macroscopically distinguish the mathematical counterparts of the possible independent outcomes associated with the physical properties of the system (hereinafter the *problem of interference*), but also to pick up a single outcome that is actually measured (i.e., the problem of outcomes, as defined above). Whereas the first problem is addressed by providing a recipe to reduce the overlap between the effective wave function packets of the macroscopic

[12] One concrete idea in this direction is perspectivalism endorsed by (Dieks, 2019), which mainly states that NCQM are interpretations that deny (iii) by virtue of the fact that they permit the possibility of ascribing to the same physical system more than one objective state.

superposition (as long as they are allowed to have approximately disjoint support), the second is overcome by providing a selective method to pick up the respective wave packet associated with the measured outcome.

One implicit and essential method to which some NCQM usually appeal is the decoherence mechanism.[13] From the point of view of decoherence, the process of measurement consists in a (purely unitary) quantum-mechanical interaction between the system and a great number of environmental degrees of freedom (including those of the observer and the measurement apparatus) that come into play, such that the interference of macroscopic states exists but cannot be observed. Formally speaking, the decoherence mechanism treats, for all practical purposes, the improper or reduced states of the superposition *as if they were* proper mixed states. This allows the identification of a set of distinguishable wave packets with approximately disjoint support (enough to be macroscopically relevant) associated with the possible values of the physical quantities to be measured (together with their probability of occurrence).

However, it is important to remark that the decoherence mechanism does not (and does not aim to) select a single outcome during the process of measurement. Since decoherence only intends to solve the problem of interference (as opposed to the problem of outcomes), the NCQM that appeal to decoherence need to account for their instrumental empirical adequacy by other means. This does not mean, however, that if decoherence were reliable, it would be useless as regards the instrumental empirical adequacy of the theory. As is correctly pointed out by (Schlosshauer, 2008), there is no way of solving the problem of outcomes without accounting for the non-observability of interference. This means that adherents of decoherence agree that there is a special mechanism described by this framework that is necessary (but not sufficient) for the observer to see definite and single outcomes.[14]

Under these circumstances, proponents of NCQM have figured out other strategies that might solve the problem of interference together with the problem of outcomes without appealing to the decoherence mechanism. For example, let us consider relational quantum mechanics. As discussed in (Rovelli, 1996), the apparent breakdown of the unitary evolution of the physical state results, according to this interpretation, from the fact that the observer is unable to provide a complete description of its interaction with the system. This relational approach is motivated by a theorem due to (Breuer, 1995), which states that if a measuring apparatus is part of a physical system, the state of the apparatus cannot fully encode the state of the whole system (including the apparatus itself). Consequently, we cannot infer the state of the whole system by observing the state of the apparatus. Although we shall not inquire into the details of relational quantum mechanics, it is important to note that this approach has fallen prey to various objections, some of which point towards the lack of ontological clarity and empirical adequacy of this theory.[15] In replying to these critiques, one might suggest introducing the decoherence mechanism to solve this problem of clarity and empirical adequacy (as seems to be suggested in (Biagio and Rovelli, 2021)). However, if we reconsider this strategy, we fall back into the above objections. It seems that NCQM cannot get rid of the decoherence mechanism until they show that it is irrelevant to them without facing serious problems, an approach which, to the best of our knowledge, has not been entirely successful.

Finally, in the particular case of many-worlds interpretations, which is the dominant proposal among the type of theories classified as NCQM, the problem of instrumental empirical inadequacy can also be expressed as the incapability of this interpretation to justify the emergence of probabilities and of the Born's rule. If the time evolution governing the dynamics of the physical state is given by the Schrödinger equation (which is a deterministic law), how does the Born's rule or the concept of probability itself emerge? If one assumes, as Everett argues, that each world is equally real, what does it mean to say that it is more probable to be in one world than in another? Without answers to these questions,

[13] Although the decoherence mechanism is regularly involved in these interpretations, as is argued in (Bacciagaluppi, 2016), it is not necessarily involved, as we shall see below.

[14] Although not generally accepted, there have been serious objections to considering decoherence a reliable strategy not only to select a single outcome (something which is generally accepted as noted above) but even to explain the non-observability of interference. One of these objections was recently proposed by (Okon and Sudarsky, 2016).

[15] One example is due to (Muciño *et al.*, 2022). However, Rovelli's response to this objection (Rovelli, 2021) has been criticised because he does not offer a solid argument that addresses the point at issue.

https://doi.org/10.1387/theoria.26061

we cannot obtain an empirically adequate interpretation. Following (Norsen, 2017), let us look at the problem more closely.

Consider a 1/2 spin particle that is in the ↑ state (up) along the $\hat{n}$ direction. This state can be represented as a superposition (up-down) along the z-direction, that is:

$$\psi_{\uparrow}^{n} = \sqrt{p}\,\psi_{\uparrow}^{z} + \sqrt{q}\,\psi_{\downarrow}^{z}. \tag{5}$$

According to UQM, the probability of finding the state $\psi_{\uparrow}^{z}$ or the state $\psi_{\downarrow}^{z}$ is p or q, respectively (we will assume, without loss of generality, that $p > q$). However, in the Everettian proposal, both worlds are equally real. Granted this, if one wishes to have the same predictive success as UQM, one expects (statistically) that if an observer is in a world corresponding to the $\psi_{\uparrow}^{z}$ state and repeats the experiment many times, he will more often be found in worlds associated with the $\psi_{\uparrow}^{z}$ state. A statistic that respects Bohr's rule must be such that by repeating the experiment N times, the expected number for the $\psi_{\uparrow}^{z}$ state will be Np. The problem is that to recover that statistic one must presuppose Born's rule, and this argument seems to be circular: for $\Psi = \sum c_i \psi_i$, a statistical weight $w_i = |c_i|^2$ must be associated with each branch. The question is, why is it reasonable and what does it mean to demand that each branch should have such a statistical weight considering that each world is "equally real"? As long as these questions are not addressed and the Born's rule is not justified (without appealing to measurements or observers), this proposal is not in a better shape than UQM.[16]

Omitting further details, the most important point to learn from this discussion is the fact that, without providing a reliable answer to the problem of outcomes, the instrumental empirical adequacy of NCQM, which is a necessary condition for scientific realism, is compromised. Specifically, these theories, as they are standardly formulated without any collapse mechanism, do not have an operational technical procedure required to account for definite and single outcomes.

5.1.4. The instrumental empirical inadequacy of epistemic/statistical interpretations

Finally, we shall see that if claim (1) is denied, whereas (2) and (3) are retained, without adding hidden variables to the physical state, then we end up with a problem of instrumental empirical adequacy. As we know from (Maudlin, 1995), denying claim (1) without introducing hidden variables leads to epistemic/statistical interpretations.

According to many scholars, there is a shared intuition that epistemic/statistical interpretations are irrelevant for a quantum realist reading simply because they are only interested in calculating predictions or the lack of knowledge of the physical system as opposed to describing objective states of affairs through the lens of QM. Our opinion is that this intuition is true for statistical but not for epistemic interpretations. The problem with the latter interpretations is that if we are aiming at calculating the lack of knowledge of certain states of affairs, we need to have at least a predetermined standard in physical reality with respect to which we are making such a calculation. Indeed, epistemic interpretations make a silent and unjustified commitment to the existence of some unknown, predetermined aspect of the physical world, whilst they explicitly deny the possibility that QM may describe that aspect of the world objectively.

Let us illustrate this critique with an example provided by (Maudlin, 1995) without inquiring into its details. If the state of the system composed of the spin of an electron and an x-spin measuring device evolves to a macroscopic superposition (as dictated by QM), epistemic interpretations will interpret this superposition as if it *really* were a mixed

[16] As suggested by (Deutsch, 1999; Saunders, 2010), some attempts have been made to justify the Born's rule. They basically assume that every rational agent has no choice but to behave according to this rule. However, as argued by (Norsen, 2017), these proposals can be challenged, and there is still no consensus on whether these strategies solve the problem of instrumental empirical adequacy.

state, meaning that the whole system is in either one or the other term of the superposition, but we are uncertain about which is the actual realised term. However, in the absence of any justification for regarding a pure state as a mixture, the macroscopic superposition alone cannot be interpreted in epistemic terms. The reason is that we are incapable of being ignorant of which individual term is realised if we cannot differentiate and single out any term from the superposition to begin with. In this way, the identification of terms in this macroscopic superposition would be equally regarded as introducing predetermined values for the spin of the electron that are not included in (or described by) QM. This seems not irrelevant but a problem for at least the semantic component of scientific realism. Let us recall that scientific realism in general and semantic realism in particular presuppose the possibility of providing an ontologically clear and empirically adequate description of the physical world *through the lens of QM*. However, although epistemic interpretations are committed to the existence of the aspect of the world of which we are ignorant —which means that they accept the mind-independent condition of scientific realism— they assume that no description of this aspect of the world can be provided by QM. This leads to the conclusion that these interpretations are empirically inadequate in the instrumental form. Let us elaborate more on this.

The first obvious consequence of endorsing the epistemic interpretation is that, following the previous example, it does not provide a solution to the aforementioned problem of outcomes —a sufficient condition for the instrumental empirical adequacy of QM— as this interpretation explicitly denies the possibility of knowing with certainty which is the realised, single outcome of the electron spin. Furthermore, the same example shows that epistemic interpretations do not provide a solution to the aforementioned problem of interference —a necessary condition for the instrumental empirical adequacy of QM— because they presuppose that the physical state of the electron spin and the measuring device represent an array of possible definite outcomes, without providing any justification for this presupposition based on QM. Therefore, the problem of instrumental empirical adequacy remains unsolved.

5.2. The problem of ontological clarity

As we shall see, the measurement problem is also related to the lack of ontological clarity regarding the interpretation of the theory. The reason for this relation can be traced back to the empirical adequacy criterion broadly understood. Recall that ontological clarity is a necessary condition for empirical adequacy. The latter not only involves an instrumental criterion (the consistency constraint), but also the (metaphysically) explanatory criterion, which depends on the clear specification of the ontic state of the theory. Thus, regardless of the satisfaction of the instrumental criterion, the fact that a theory is ontologically unclear implies that it is at odds with the explanatory criterion, and thus with the empirical adequacy criterion. To illustrate this observation, let us investigate each of the interpretations already mentioned.

5.2.1. The lack of ontological clarity of unitary quantum mechanics

We shall demonstrate that if claim (3) is denied but (1) and (2) are retained, without embracing NCQM, we end up with a problem of ontological clarity. This shall lead to the conclusion that UQM is ontologically unclear.

In addition to the deterministic evolution, the root of the instrumental empirical adequacy problem in UQM lies on the basic realist assumption of this theory that the physical state of the system is complete and represents in some way or other all of the objective physical properties of the system. In this way, it is assumed that the macroscopic superposition of states that arises in measurement situations is a physical state in its own right, in the sense that it represents the quantum world as it, in fact, is. However, note that QM does not specify with precision the ontic state of the theory. That the wave function specifies all the physical properties of a system does not imply that it exists as an object in the world. Rather, it means that there are physical properties of unobservable entities (not explicitly specified) that have a factual correspondence with the wave function. Thus, as things stand, UQM is unclear in the ontological sense.

https://doi.org/10.1387/theoria.26061

5.2.2. The lack of ontological clarity of collapse and non-collapse interpretations

We shall demonstrate three conditionals at once:

> If claim (1) is denied, whilst (2) and (3) are retained, without embracing BQM (adding hidden variables to the physical state), we end up with a problem of ontological clarity.
>
> If claim (2) is denied, whilst (1) and (3) are retained, without embracing GRW (adding non-linear terms to the dynamics), we end up with a problem of ontological clarity.
>
> If claim (3) is denied, whilst (1) and (2) are retained, we end up with a problem of ontological clarity.

Since the first two conditionals are embraced by FRQM; the second conditional is embraced by SQM; and the third one by NCQM, we shall conclude that these three theories or interpretations are ontologically unclear.

As regards SQM and FRQM, the introduction of the projection postulate means that what was initially a deterministic and unitary evolution governed by the Schrödinger equation turns out to be an indeterministic jump that has to be introduced by hand from the outside. The lack both of naturalness and a clear and comprehensive explanation for this collapse mechanism is the reason why many philosophers think of the measurement problem in SQM and FRQM as a lack of objectivity with respect to our theoretical dealings with the quantum phenomena. According to this scenario, both theories would be at odds with our objectivity standards because the nature and behaviour of what is measured is observer-dependent. There would not be any kind of explanation of what happens when the observer does not have a human brain capable of rationalising like a PhD student, or when the observer is not present as a conscious witness at a certain place and time (Bell, 1971). In the particular case of FRQM, this problem has been characterised as a conflict (or lack of rules) in establishing the correct interpretation among the two different points of view arising from the aforementioned Neumann's observer distinction. Furthermore, the epistemological or ontological interpretation of this conflict between two different points of view has been taken as the basis for further reasons why this theory must either objectively describe the state of consciousness of the observer (London and Bauer, 1939).

What about NCQM? Although these interpretations do not introduce the projection postulate and the system in NCQM evolves according to a purely unitary evolution, the aforementioned Neumann's observer distinction between different observers is also introduced by means of certain external mechanism (such as decoherence) that makes possible the observability of (epistemic) definite and single outcomes. This means that, apart from the obscure character of this or other proposed non-collapse mechanisms, NCQM preserve the same problems of SQM and FRQM as regards objectivity.

For example, the most advocated many-worlds interpretation is committed to wave function monism, namely, the idea that the wave function is all there is in the world —in proper terms, the physical object mathematically represented by the wave function. Proponents of this interpretation believe that this feature makes the theory parsimonious, and for this reason, they refuse to postulate anything in addition to the wave function. However, if the wave function is all there is in the world, in what way can we extract from this abstract object a picture of the physical world that can explain our empirical experience i.e., an explanation that solves the macro-object problem?

At first glance, Wallace (2010) addresses the above question by arguing that the macroscopic entities of the empirical world should not be considered illusory even if they do not appear in the basic axioms of the theory; rather, they should be understood as patterns or structures arising from the states of the microphysical theory in question. In his own words:

> There are structural facts about many microphysical systems which, although perfectly real and objective (try telling a deer that a nearby tiger is not objectively real) simply cannot be seen if we persist in describing those systems in purely microphysical language (Wallace, 2010, p. 58).

And, in the specific context of many-worlds, Wallace claims:

> Structurally speaking, the dynamical behaviour of each wavepacket [i.e., each decoherent branch of the wave function] is the same as the behaviour of a macroscopic classical system. And if there are multiple wavepackets, the system is dynamically isomorphic to a collection of independent classical systems (Wallace, 2010, p. 64).

This seems odd. As argued by (Norsen, 2017), establishing an isomorphism is not sufficient to solve the macroobject problem in this context. He comments that a wave packet propagating in configuration space is roughly isomorphic to a particle traveling along an (approximate) trajectory through configuration space. This trajectory in turn is isomorphic to N trajectories in 3-dimensional space; however, this does not imply that the existence of a wave packet propagating in $3N$-dimensional space leads to the existence of N particles in physical space. The example he gives is the isomorphism between a billiard ball bouncing on a square table with two beads inside a wire bouncing back and forth: The isomorphism is established by assuming that one bead is small enough to pass through the hole of the other, this to disregard the interaction between the two. A billiard ball, beyond mathematical equivalence, is not a pair of beads in a wire nor does the existence of one imply the existence of the others.

Since the macro-object problem remains unsolved in this context, a many-worlds interpretation with a mass density ontology is often proposed. This interpretation postulates a mass density (as defined by GRW) for each branch and the effect of decoherence is considered to justify why each world cannot realize the existence of the other. Using the Schrödinger's cat experiment, there will be a live cat in one world and a dead cat in another but each with half the mass (the inhabitants of each branch will not realize this fact). We could say that this interpretation does solve the ontological problem. However, it is still an empirically inadequate theory as the emergence of probability and the Born's rule remains unsolved.

Consequently, what lies behind these alternative views or conclusions is the fact that the nature and the dynamics of the physical state of the system alone cannot account for (objective or epistemic) definite and single measurement outcomes and because of that, SQM, FRQM and NCQM must appeal to ambiguous mechanisms whose operation does not depend on the objective constituents, principles and laws of the theory but on the idiosyncrasy of the observer. With the projection postulate or any non-collapse mechanism introduced, however, these theories remain ontologically unclear in the sense that there is no pre-conceived objective ontology that can exhaustively explain the observed phenomena and the measurement process itself as any objective physical interaction between the constituents of the system and the experimental device in question. Rather, the constitutive nature and behaviour of what is measured directly depends on the experimental context and cannot be conceived before the interaction takes place. In this way, the projection postulate or any non-collapse mechanism involve satisfaction of neither the instrumental criterion (as demonstrated above) nor the explanatory criterion, both of which are necessary to make the theory empirically adequate. Therefore, since the semantic component of realism requires that SQM, FRQM and NCQM satisfy the clarity criterion, and this criterion is not satisfied by them, then we are led to conclude that they cannot be interpreted realistically, at least in this semantic sense.

Furthermore, the fact that FRQM is ontologically unclear not only implies that this theory is at odds with our preferred definition of scientific realism, but also that the first instrumental requirement of empirical adequacy can not always be satisfied even with the introduction of the projection postulate. Indeed, the FR theorem described above may be regarded as an illustration of the claim that instrumental empirical adequacy fails as a consequence of the ambiguity of the projection postulate (or any mechanism that explains the observability of definite and single outcomes). The reason is that the acceptance of Neumann's observer distinction by FRQM implements a profound ontological ambiguity upon the description of the quantum phenomena. Thus, in order to avoid incorrect predictions we have to know how to apply the formalism of FRQM in a given physical situation. We suggest that it is through the specification of a clear ontology and the acquisition of knowledge of what exactly is happening in the physical world when a measurement is performed that we are able to apply this formalism without inconsistencies. As is the case for the second metaphysical criterion of empirical adequacy, the FR theorem also suggests that the first instrumental criterion cannot be satisfied without the ontological clarity criterion. Empirical adequacy and ontological clarity are deeply correlated semantic realist criteria.

5.2.3. The lack of ontological clarity of epistemic/statistical interpretations

Finally, we shall see that if claim (1) is denied, whereas (2) and (3) are retained, without embracing BQM (adding hidden variables to the physical state), then we end up with a problem of ontological clarity associated with epistemic/statistical interpretations.

Statistical interpretations are by construction ontologically unclear by virtue of the fact that they are only interested in calculating predictions of the physical system. As for epistemic interpretations, let us recall the famous PBR theorem. According to certain (debatable) assumptions, this theorem shows that models of QM in which pure quantum states are interpreted epistemically (i.e., represent probabilistic or incomplete states of knowledge about physical reality) cannot reproduce quantum predictions. As a result, pure quantum states must be construed ontologically in the sense that they describe objective states of affairs. As is well known, this conclusion is derived by considering that the quantum state can be interpreted ontologically if every complete physical state or ontic state in the theory is consistent with only one pure quantum state; or epistemically if there exist ontic states that are consistent with more than one pure quantum state (Pusey *et al.*, 2012). This distinction lends support to the idea that epistemic quantum states cannot describe ontic states. Furthermore, if we were to interpret the quantum state epistemically, we would be incapable of providing a minimal characterisation of the ontological profile of the ontic state of the theory. Since this minimal characterisation is a prerequisite for the ontological clarity of QM, the epistemic interpretation, according to this distinction, neglects the possibility of QM being ontologically clear and, therefore, does not satisfy the semantic component of scientific realism. We can deduce from this observation that if the premises of the PBR theorem are true, its conclusion reinforces the objection that epistemic interpretations pose against ontological clarity, and provides necessary (albeit not sufficient) conditions to be a semantic realist with respect to QM.

6. *Concluding remarks*

Based on a well-known characterisation of scientific realism, we conclude that the description of the processes involved in typical measurements framed in terms of UQM, SQM, FRQM, NCQM and epistemic and statistical interpretations is problematic from the standpoint of what we have defined as semantic realism. In particular, we have argued that this description is problematic by virtue of the fact that there is no specification of what exactly these interpretations say in terms of their ontologies (i.e., they are ontologically unclear), and that the relevant empirical data, confined to the domain in which they can be effectively applied, cannot be explained in terms of these ontologies and cannot be entailed by their corresponding formulations (i.e., they are empirically inadequate). Since semantic realism is compatible with (but not reducible to) realism and (or) antirealism, we conclude that the measurement problem is neutral with respect to these philosophical positions. As emphasised in the introduction, this does not mean that the measurement problem is independent of the realism-antirealism debate; rather, this problem pressuposes certain realist commitments that cannot be appreciated if one insists in endorsing the endpoints of a more complex, continuum realist spectrum. Let us make a final remark.

As discussed above, the conclusion of this paper locates the measurement problem in a neutral position with respect to the endpoints of the realism-antirealism debate. Indeed, even if it is a problem that troubles someone of realist spirit, the antirealist should also be worried as she embraces certain realist commitments, such as instrumental empirical adequacy, without collapsing to fully-fledged scientific realism. However, another important result implied by our discussion, is that the measurement problem is also neutral with respect to metaphysical debates informed (or justified) by QM. This is because semantic realism is equally compatible with metaphysical of philosophical positions competing against each other. For example, *structural realism* contrasts with standard scientific realism as it reverses the ontological priority of the relata over and above the relations and structures in which they stand. However, this form of realism acknowledges that empirical adequacy and ontological clarity at the level of the relata should be satisfied. As emphasised by (French, 2014, Ch. 7), this is because there is no way to explain the empirical phenomena in terms of relations and structures without having to explain such phenomena in terms of the relata in the first place.

Acknowledgements

We wish to thank the two anonymous referees for this journal for their helpful comments and suggestions. We confirm that this work was supported by CONAHCYT under the scholarship "Estancias posdoctorales por México 2022(1)" [Grant number: 442599]; and by the grant UNAM-DGAPA-PAPIIT IN101724. These funding sources have no involvement in study design; in the collection, analysis and interpretation of data; in the writing of the report; and in the decision to submit the article for publication.

REFERENCES

Bacciagaluppi, G. (2016). The role of decoherence in quantum mechanics. *The Stanford Encyclopedia of Philosophy* (Fall 2016 Edition), E. N. Zalta (ed.), URL=<https://plato.stanford.edu/archives/fall2016/entries/qmdecoherence/>.

Bell, J. S. (1971). Against measurement. In Aspect, A. (ed.) *Speakable and unspeakable in quantum mechanics* (pp. 213-231). Cambridge: Cambridge University Press.

Boge, F. J. (2019). Quantum information versus epistemic logic: An analysis of the Frauchiger–Renner theorem. *Foundations of Physics*, *43*(10), 1143-1165.

Boge, F. J. (2023). The positive argument against scientific realism. *Journal for General Philosophy of Science*, *54*, 535-566.

Breuer, T. (1995). The impossibility of accurate state self-measurements. *Philosophy of Science*, *62*, 197-214.

Bub, J. (2017). Why Bohr was (mostly) right. *arXiv:1711.01604v1[quant-ph]*.

Chakravartty, A. (2017). Scientific realism. *The Stanford Encyclopedia of Philosophy* (Summer 2017 Edition), E. N. Zalta (ed.), URL=<https://plato.stanford.edu/archives/sum2017/entries/scientific-realism/>.

Deutsch D. (1999). Quantum theory of probability and decisions. *Proceedings: Mathematical, Physical and Engineering Sciences*, *455*(1988), 3129-3137.

Di Biagio, A., & Rovelli, C. (2021). Stable facts, relative facts. *Foundations of Physics*, *51*(30), 1-13.

Dieks, D. (2019). Quantum mechanics and perspectivalism. In Lombardi, O., Fortin, S., López, C., & Holik, F. (eds.) *Quantum worlds: Perspectives on the ontology of quantum mechanics* (pp. 51-70). Cambridge: Cambridge University Press.

Everett, H. (1957). 'Relative state' formulation of quantum mechanics. *Reviews of Modern Physics*, *23*(3), 454.

Faye, J. (2018). Backward causation. *The Stanford Encyclopedia of Philosophy* (Summer 2018 Edition), E. N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2018/entries/causation-backwards/>.

Feigl, H. (1950). Existential hypotheses: realistic versus phenomenalistic interpretations. *Philosophy of Science*, *17*, 35-62.

Frauchiger, D. & Renner, R. (2018). Quantum theory cannot consistently describe the use of itself. *Nature Communications*, *3*(3711).

French, S. (2014). *The structure of the world. Metaphysics and representation*. Oxford: Clarendon Press.

Horwich, P. (1982). Three forms of realism. *Synthese*, *51*(2), 181-201.

Jammer, M. (1974). *Philosophy of quantum mechanics. The interpretations of quantum mechanics in historical perspective*. John Wiley & Sons.

Lam, V., & Wüthrich, C. (2018). Spacetime is as spacetime does. *Studies in History and Philosophy of Modern Physics*, *64*, 39-51.

Lazarovici, D., & Hubert, M. (2019). How quantum mechanics can consistently describe the use of itself. *Scientific Reports*, *3*(470).

London, F., & Bauer, E. (1939). The theory of observation in quantum mechanics. In Wheeler, J. & W. Zurek, W. (eds.) *Quantum theory and measurement* (pp. 217-259). New Jersey: Princeton University Press.

Maudlin, T. (1995). Three measurement problems, *Topoi*, *14*, 7-15.

Maudlin, T. (2019). *Philosophy of physics: Quantum theory*. Princeton: Princeton University Press.

Muciño, R., & Okon, E. (2020). Wigner's convoluted friends. *Studies in History and Philosophy of Modern Physics*, *72*, 87-90.

Muciño, R., & Okon, E., & Sudarsky, D. (2022). Assessing relational quantum mechanics. *Synthese*, *200*(399).

Ney, A. (2015). Fundamental physical ontologies and the constraint of empirical coherence: a defense of wave function realism. *Synthese*, *132*(10), 3105-3124.

Ney, A. (2017). Finding the world in the wave function: some strategies for solving the macro-object problem. *Synthese*, *137*, 4227-4249.

Norsen, T. (2017). *Foundations of quantum mechanics*. Cham: Springer International Publishing.

Okon, E. (2021). On the objectivity of measurement outcomes. *The British Journal for the Philosophy of Science, 76*(2), 345-371.

Okon, E., & Sudarsky, D. (2016). Less decoherence and more coherence in quantum gravity, inflationary cosmology, and elsewhere. *Foundations of Physics*, *46* (7), 852-879.

Psillos, S. (1999). *Scientific realism: How science tracks truth*. London: Routledge.

Pusey, M., Barrett, J., & Rudolph, T. (2012). On the reality of the quantum state. *Nature Physics*. *8*, 475-478.

Rovelli, C. (1996). Relational quantum mechanics. *International Journal of Theoretical Physics*. *35*(8), 1637-1678.

Rovelli C. (2021). A response to the Muciño-Okon-Sudarsky's assessment of relational quantum mechanics. ArXiv:2106.03205 [quant-ph].

Saunders S. (2010). Many worlds: an introduction. In Saunders S., Barrett J., Kent A., Wallace D. (eds.). *Many worlds? Everett, quantum theory, and reality* (pp. 1-52). Oxford: Oxford University Press.

Schlosshauer, M. (2008). *Decoherence and the quantum-to-classical transition*. Heidelberg: Springer.

Sudbery, A. (2019). The hidden assumptions of Frauchiger and Renner. *International Journal of Quantum Foundations*. *5*(3), 98-109.

Tausk, D. (2019). A brief introduction to the foundations of quantum theory and an analysis of the Frauchiger-Renner paradox. arXiv:1812.11140v2.

van Fraassen, B. (1980). *The scientific image*. Oxford: Clarendon Press.

von Neumann, J. (1932). *Mathematical foundations of quantum mechanics*. New Jersey: Princeton University Press.

Wallace D. (2010). Decoherence and ontology. In Saunders S., Barrett J., Kent A., Wallace D. (eds.). *Many worlds? Everett, quantum theory, and reality* (pp. 53-72). Oxford: Oxford University Press.

Juan Alberto Guzmán García is an Adjunct Professor at Universidad Nacional Autónoma de México. He works on philosophy of physics and the foundations of quantum mechanics.

Address: Facultad de Ciencias, Universidad Nacional Autónoma de México, Investigación Científica, Ciudad Universitaria, Coyoacán, 04510, Mexico City, Mexico. E-mail: d0n_juan@ciencias.unam.mx – ORCID: https://orcid.org/0000-0001-7142-284X.

Jorge Alberto Manero Orozco is a postdoctoral researcher at Universidad Nacional Autónoma de México. He works on philosophy of science, epistemology/metaphysics of science and philosophy of physics.

Address: Instituto de Investigaciones Filosóficas, Universidad Nacional Autónoma de México, Circuito, Mario de La Cueva s/n, Ciudad Universitaria, Coyoacán, 04510, Mexico City, Mexico. E-mail: jorge.manero@filosoficas.unam.mx – ORCID: https://orcid.org/0000-0003-4588-5132.

Theoria, 2025, 40(2), 210-230
https://doi.org/10.1387/theoria.26291

THEORIA

ISSN 0495-4548 – eISSN 2171-679X

Mapping the communication of science

(Mapeando la comunicación científica)

Vincenzo Crupi*
University of Turin
https://orcid.org/0000-0002-8727-5001

Saúl Pérez-González*
University of Valencia
https://orcid.org/0000-0001-9594-9558

Keywords

Science communication
Framing
Message fatigue
Science denialism

ABSTRACT: Science communication is a broad field and involves very diverse activities. This paper aims to illuminate and partly systematise the diversity of science communication. We focus on three important dimensions: size of the audience, frequency of interaction, and decision-making relevance. Based on them, we introduce a three-dimensional space of science communication wherein particular scenarios can be located. We argue that relevant challenges for science communication are particularly associated with certain areas of this space. Based on the proposed framework, we also address potential strategies and developments in science communication.

Palabras clave

Comunicación científica
Encuadre
Fatiga de mensaje
Negacionismo de la ciencia

RESUMEN: La comunicación científica es un campo amplio que abarca actividades muy diversas. Este artículo tiene como propósito esclarecer y sistematizar parcialmente dicha diversidad. Nos centramos en tres dimensiones clave: el tamaño de la audiencia, la frecuencia de interacción y la relevancia para la toma de decisiones. A partir de estas dimensiones, proponemos un espacio tridimensional de la comunicación científica en el que pueden ubicarse los escenarios particulares. Sostenemos que ciertos desafíos relevantes para la comunicación científica están especialmente asociados a ciertas zonas de este espacio. En base al marco propuesto, también abordamos posibles estrategias y desarrollos en el ámbito de la comunicación científica.

1. *Introduction*

The idea that science communication can be studied with the tools of science itself (e.g., Fischhoff & Scheufele, 2013) is plausible. After all, we can apply the main elements of the scientific approach (developing hypotheses and critically evaluating them on the basis of controlled data) to the most diverse phenomena —why not to science communication? Indeed, results obtained in established scientific fields (cognitive psychology, just to mention one clear example) can already be integrated in analyses of science communication (Mercier, 2020). So it makes sense to

* ***Correspondence to:*** Vincenzo Crupi. Department of Philosophy and Education Sciences, University of Turin, via Sant'Ottavio 20, 10124 Torino, Italia – vincenzo.crupi@unito.it– https://orcid.org/0000-0002-8727-5001

* ***Correspondence to:*** Saúl Pérez-González. Department of Philosophy, University of Valencia, Avd. Blasco Ibáñez 30, 46010 Valencia, Spain – saul.perez@uv.es – https://orcid.org/0000-0001-9594-9558

How to cite: Crupi, Vincenzo; Pérez-González, Saúl (2025). «Mapping the communication of science»; *Theoria. An International Journal for Theory, History and Foundations of Science,* 40(2), 210-230. (https://doi.org/10.1387/theoria.26291).

Received: 01 May, 2024; Final version: 04 June, 2025.
ISSN 0495-4548 - eISSN 2171-679X /

think that there is specific knowledge concerning science communication, with no convincing alternative to systematic, rational, and empirically testable research, i.e., scientific research, to obtain and consolidate such knowledge.

To apply scientific methods to the study of science communication, the science of science communication adopts an empirical approach towards issues such as understanding audiences, designing messages, and evaluating communication strategies. In this sense, one of the central aims of the science of science communication is to address the so-called "science communication paradox" (Kahan, 2015): although human societies have significantly expanded their understanding of the natural and social world, there is little consensus on the scope of their collective knowledge. Other major challenges are related with new media formats, emergent research topics, and polarised political environments (Akin, 2017). The science of science communication is meant to be distinct from research on other forms of communication, e.g., political communication, because of the privileged rhetorical status of science communication and, consequently, the specific standards and norms to be met (Jamieson, 2017).

It should be aknowledged, however, that this remains an emergent field with contributions from a variety of potentially diverging perspectives (Kahan *et al.*, 2017). Under these circumstances, a rather basic effort of organization of the phenomena may be worthwhile. The main aim of this paper goes in this direction: our goal is to outline a scheme for classifying diverse kinds and modes of science communication and to explore the potential of such scheme. In developing our proposal, we will keep the use of theoretical assumptions and tools to a minimum, adopting a descriptive approach potentially providing a basis for a comparative and independent evaluation and/or integration of different theoretical and practical approaches. Our discussion will especially illuminate and partly systematise the diversity of science communication, introducing a framework for assessing an array of scenarios, problems, and strategies.

The structure of the paper is as follows. Section 2 introduces the issue of the diversity of science communication settings and activities. Section 3 discusses three relevant dimensions of science communication. It outlines a three-dimensional space regarding which particular communication scenarios can be located. Section 4 illustrates the three-dimensional space of science communication by means of particular cases. Section 5 discusses important challenges faced by science communication. We argue that some difficulties are particularly associated with certain areas of the spatial representation outlined. Section 6, based on the proposed framework, addresses potential strategies and developments in science communication. Finally, section 7 concludes.

2. *The diversity of science communication*

Science communication is a very broad field. According to the influential characterisation proposed by Burns *et al.* (2003), science communication encompasses the diverse skills, media, activities, and dialogues that aim to produce awareness, enjoyment, interest, opinions, or understanding of science in a certain audience. The target audience of science communication can be (a group of) media, corporations, policy-makers, or the general public.[1] Furthermore, science communication can concern virtually any topic related with science or technology. Still, science communication can be demarcated as usually guided by some general rules including to be ethical, to respect the factual truth, to not omit viable options, and to declare possible conflicts of interest (Jucan & Jucan, 2014).

Science journalism and documentaries that address topics in the natural sciences (e.g., astronomy) are often considered paradigmatic examples of science communication. Certainly, the main role of those media contents, which are

[1] In this paper, we will focus on cross-sectoral communication between scientific communities and other segments of society. Nonetheless, science communication may also involve interactions within or between scientific communities (e.g., scientific conferences), as well as interactions between laypeople (e.g., citizen meetings about scientific issues) (Trench & Bucchi, 2010).

targeted to the general public, is to increase awareness and understanding of important scientific facts. Nevertheless, science communication arguably is involved in many other scenarios such as medical consultation, science and technology museums, public-health campaigns, and even science fiction films. These also partly aim at producing certain personal responses towards science in their audience (enjoyment is surely one example).

In a pioneering study, the UK Office of Science and Technology and the Wellcome Trust (2001) reviewed science communication activities in Britain. The study included nine in-depth face-to-face and fifteen telephone interviews with key players in science communication. Moreover, forty-four web questionnaires and thirty-five telephone interviews were carried out. According to the study, science communication activities in Britain diverge in important aspects such as the objective of the activity, how innovative it is perceived to be, and the geographic area of influence. Those activities include, among others, open days, science clubs, and government-commissioned reports. In the last decades, the diversity of science communication has arguably increased even further. The development of new technologies such as smartphones and the popularisation of social networks such as Facebook or Instagram have given rise to new forms of interaction between science communicators and audiences. Consider for instance YouTube. YouTube has become an important platform for sharing both user-created and professionally generated content about science (Allgaier, 2020). This website allows for a direct, immediate, and interactive form of science communication. Communicators can create and upload their videos autonomously and even interact with the audience in real time.

Nonetheless, the diversity of science communication and the divergences between particular communicative scenarios arguably deserve more attention than is currently acknowledged from scientists, science communicators, and theorists. Lack of attention may be also related to so-called *deficit approach* to science communication. Despite sustained criticism among specialists, many scientists and science communicators still embrace core elements of the deficit approach and tend to identify science communication straight away with the closure of a knowledge gap in the audience (Besley & Tanner, 2011; Cortassa, 2016; Davies, 2008; Suldovsky, 2016).[2] This approach, however, does tend to overlook particularities of the diverse science communication scenarios and to invite a one-size-fits-all approach to practice.

The deficit approach is based on the *deficit model* of the public understanding of science (Ahteensuu, 2012; Bubela *et al.*, 2009; Bucchi & Trench, 2014; Burns *et al.*, 2003). According to this model, novices usually have inadequate or incomplete knowledge or understanding of science. Such ignorance is considered to be the basis of most negative attitudes towards science and technology (e.g., scientific scepticism). For advocates of the deficit approach, the main role of science communication is to provide accurate information about science and fill the deficit in knowledge. It is considered that such (one-way) transfer of information is the fundamental tool for achieving scientific literacy and, as a consequence, public support of science.[3]

In the deficit framework, the diversity of science communication is indeed typically overlooked. It is assumed that both the challenge faced and the adequate intervention are virtually the same in all cases and differences among di-

---

2 In the (academic) debate on science communication, several concerns have been raised against the deficit approach. Science communication scholars have argued that this approach is overly simplistic, empirically inadequate, and largely ineffective (Bubela *et al.*, 2009; Bucchi & Trench, 2014; Nisbet & Mooney, 2007).

3 In recent decades, various alternative approaches to science communication have been proposed. Among these, the communication of scientific consensus has garnered particular attention (Bayes *et al.*, 2023). It has been suggested that consensus messages influence the perceived consensus and, consequently, laypeople's beliefs and attitudes. Existing evidence supports the role of perceived consensus in the formation and acceptance of beliefs (Mercier, 2020). Furthermore, in the context of climate change communication, several empirical studies have found a positive effect of consensus messaging on the acceptance of climate change and support for public action (van der Linden *et al.*, 2015, 2019). However, concerns have also been raised regarding the potential side effects of consensus messages (Chinn & Hart, 2023; Ma *et al.,* 2019). Under certain conditions (e.g., conflicting prior beliefs), such messages may be perceived as manipulative, provoke reactance, and even backfire.

verse contexts of science communication elude thorough consideration. In particular, little attention is paid to the particularities of the audience, the source of information, and the scientific domain. The deficit approach appears to essentially presuppose that the public is homogeneous regarding the relevant aspects (Nisbet & Scheufele, 2009; Simis *et al.*, 2016). It is considered that novices lack important scientific knowledge and that informing them will successfully fill the gap and lead to positive attitudes towards science. Peculiarities related with audience's concerns, values, political views, and pre-existing beliefs are also typically neglected. Also, the deficit approach does not take into account the particular sources involved in science communication and the way in which they relate to the audience. The deficit model focuses on the content of the message (i.e., the relevant scientific information) and hardly takes into account contextual aspects of the communicative process. Finally, in the deficit framework, differences between scientific domains are regarded as irrelevant with virtually no impact on how science communication generally works. In all cases, the main problem is related with a knowledge deficit and should be addressed by providing the missing piece of scientific information.

Although science communication scholars often claim that the deficit approach has been abandoned, its prevalence among scientists and science communicators has been highlighted by different empirical studies. By means of several group interviews and discussions, Davies (2008) examined scientists' ideas on science communication. She highlights the influence of the deficit framework and argues that science communication is widely viewed as a one-way transfer of scientific information. Besley and Tanner (2011) interviewed science communication scholars about their views on scientists and their communicative training. According to their study, scholars broadly agree that the deficit approach remains prevalent among scientists.

There are several factors that seem to contribute to the persistence of the deficit approach. Suldovsky (2016) argues that there are strong links between widespread views of fundamental aspects of science (e.g., scientific knowledge) and science communication (e.g., purpose), and the deficit approach. For example, it is often considered that science communication's main role is to promote science within the public sphere. The communicative relationship between science and the public is thus often understood as unidirectional: science contributes to improve society, but societal phenomena and determinants play no role in science and scientific understanding. Simis *et al.* (2016) also argue that there are practical aspects that foster the persistence of the deficit approach. They argue that one crucial aspect is scientists' (lack of) training. Scientists rarely have formal training and understanding of public communication. They are not taught specifically about how laypeople form their beliefs about science and how to effectively communicate with them. The authors also consider another important factor, namely that the deficit approach leads to easy-to-implement policy designs. According to this approach, policy interventions simply have to address ignorance and promote scientific literacy.

The diversity of science communication is sometimes overlooked even by scholars of science communication among which the deficit model is widely rejected. Many analyses of science communication, although aiming at general conclusions, focus exclusively on a particular kind of science communication practices (Bubela *et al.*, 2009). Special attention is typically given to print media and science journalism. Furthermore, studies on science communication usually address the communication of scientific research and do not pay attention to the particularities of the communication of technological developments (Trench & Bucchi, 2010). Finally, science communication scholars usually focus on the effects of one-time exposure to a single message, which is studied in isolation, and pay little attention to continued exposure and its implications (So *et al.*, 2017).

Overlooking the diversity of science communication is not harmless. Not only does it lead to descriptively inadequate accounts of science communication, but it can also have practical consequences. The challenges faced by science communication and the efficacy of interventions usually depend on the particularities of the communicative context. Not paying attention to the latter may result in ineffective (or even harmful) interventions. Consider, for instance, the audience. Although science communicators care about the audience and for the message to be understood, relevant features of the target population may be overlooked. Previous knowledge about certain scientific issue varies greatly depending on aspects such as interests, social background, and domain. That means, for example, that the previous knowledge of the audience is likely to vary from one communication context to another. This previous knowledge, however, may have a significant influence in the effect of certain science-related messages (Bruine de Bruin & Bostrom, 2013; Fischhoff, 2013). Indeed, humans are more likely to accept those messages that suit their pre-existing views. For example, in two experiments about science communication on carbon-nanotubes and genetically modified foods, Druckman and Bolsen (2011) found that individuals' impressions and reactions to new factual information were biased by their previous views on the topic. Participants in the experiment tended to perceive factual information consistent with their initial opinion of the technology as stronger. Moreover, they also tended to perceive neutral factual information as consistent with their prior opinion. When the previous knowledge of the audience is overlooked, one can see even negative side effects taking place. For example, messages that the audience perceives as inaccurate, oversimplified, or repeating well-known facts may sometimes erode the trust in science communicators (Fischhoff, 2013).

In sum, science communication is a broad area including diverse activities such as science journalism, medical consultation, and operating museums. Nonetheless, the diversity of science communication tends to be overlooked by scientists, science communicators, and even science communication scholars. In some cases, this lack of attention may be related with the (unabated prevalence of the) deficit approach to science communication.

3. Three dimensions of science communication

Considering the breadth of science communication, it is essential to identify, analyse, and categorise the diverse scenarios that may emerge. Such categorisations can enrich theoretical discussions about science communication and also provide a practical framework for communicators. This framework could assist in tasks like selection of strategies, development of materials, and anticipation of potential challenges.

In the recent literature, we find several attempts to systematise science communication (Desmond, 2024; Elliott, 2017; Sánchez-Mora, 2016). Elliott (2017) focuses on engagement activities, i.e., interactions among people or institutions to exchange views, highlight problems, or evaluate proposals. Based on the entities involved and the way of interaction, he distinguishes four main forms of engagement: (1) bottom-up engagement between community groups and researchers, (2) top-down engagement exercises to elicit public participation, (3) interdisciplinary engagement between scholars, and (4) engagement with laws, policies, and institutions that influence scientific research. This categorisation aims to highlight the diverse ways in which values may influence science. On the other hand, Desmond (2024) focuses on expert communication and distinguishes it from other forms of science communication. Expert communication is perceived as authoritative, conveys assertions about the state-of-the-art, and informs decision-making about possible courses of action. Regarding expert communication, Desmond outlines a typology based on the priority given to transparency and actionability. He identifies and characterises, among others, persuasive communication, paternalistic communication, and manipulative communication.

In this section, we introduce a new framework for assessing and characterising particular scenarios of science communication. We take our current proposal as a complement and an extension of others available, with a focus on three relevant dimensions: size of the audience, frequency of interaction, and decision-making relevance. As it has been noted, in science communication, target audiences can diverge for several aspects, e.g., social role. One crucial aspect which may be overlooked is the *size of the audience*. The size of the audience can range from one person to a large number of people. In science communication contexts such as medical consultation or scientific advising, the audience is often small. It is usually limited to one patient (and perhaps some member of their family), in the first case, or some decision-makers in the second case. In public lectures, however, the audience often includes hundreds of people. In those lectures, renowned scientists usually present major research findings to a large group of laypeople. In some contexts, the audience may be even larger. For instance, public health campaigns are usually targeted to the whole population of a region. In fact, given the recent technological developments and the popularisation of social networks, the potential audience of a science-related message can be millions of people worldwide. This is the case, for example, of science-based films and YouTube videos.

The second dimension is the *frequency of interaction* between the communicator and the audience. The recurrence of the communicative interaction between science communicators and audiences can vary greatly. It can range from one unique interaction to regular interactions every few days. Those regular meetings may even span a long period of time. In the context of temporary exhibitions in science and technology museums, for example, the frequency of interaction with the audience is very low. People usually visit a particular exhibition just once before it is dismantled or moved to another venue. However, in other scenarios, there is a high frequency of interaction with the audience. For example, some science communicators become public celebrities and frequently address a loyal audience of laypeople from diverse platforms. This is the case of public figures in science communication such as Carl Sagan, Neil deGrasse Tyson, or Richard Dawkins.

The third dimension is the *decision-making relevance* of the topic addressed. Science communication can be about any topic related with science or technology and the relevance of those topics for decisions of concern of the audience varies greatly. Some scientific issues are connected to important health, economic, or political decisions, while others are highly disconnected from the audience's everyday life and decisions. For example, as popular as it is, science dissemination about astronomy has typically little decision-making relevance for laypeople. In this context, science communication mainly aims at satisfying an intellectual curiosity and increase awareness about objects and phenomena on a cosmic scale. On the other hand, communication on nutrition may have a high decision-making relevance. In this context, the aim of science communication is easily related to dietary habits in the target population. Differences regarding decision-making relevance may be found even among communication activities of the same type. Public lectures, for instance, address topics with diverse degrees of relevance, from mental health to dinosaurs' extinction.

The decision-making relevance of a particular topic may change from one audience to another. For instance, information about the heat conductivity of a newly discovered inorganic material may have little decision-making relevance for the general public, but be relevant, say, for decision-makers in charge of distributing public funds for research or making investment choices. Consequently, the decision-making relevance of a topic should be considered in relation to a certain audience.

One can usefully integrate these three dimensions in a three-dimensional space of science communication. This space can be represented with a figure like the following.

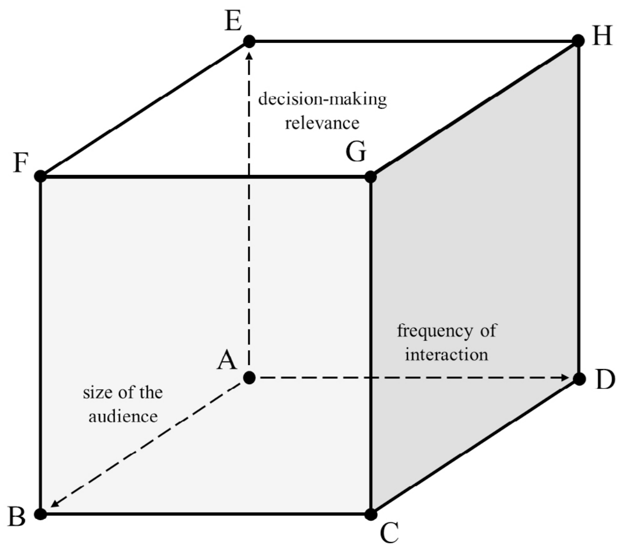

Fig 1

Three dimensions of science communication

https://doi.org/10.1387/theoria.26291

Summing up, science communication includes very diverse cases. Particular scenarios can be characterised regarding the size of the audience, the frequency of interaction, and the decision-making relevance. Those three dimensions build up a three-dimensional space of science communication in which we can locate particular cases. We consider that the proposed dimensions and the resulting framework have important strengths. First, our proposal encompasses many forms of science communication. Any communication from scientific communities to other segments of the society can be situated with respect to the proposed dimensions, while other approaches tend to focus on a particular kind of scenarios (see, for instance, Desmond, 2024). Second, the dimensions considered can often be assessed in advance. Even before a communicative interaction has taken place, science communicators can assess those parameters and roughly locate the communicative scenario in the three-dimensional space. Third, situating a scenario with respect to these dimensions helps to anticipate possible challenges and to assess available strategies in science communication. (We will explore this last aspect in sections 5 and 6.)

4. Particular scenarios of science communication

In order to illustrate diverse areas of the three-dimensional space of science communication, we are going to characterise and locate in this space some paradigmatic scenarios of science communication.

Let us consider, in the first place, science communication in clinical practice. In specialised medical consultation, interactions between doctors and patients are usually situated in the area close to vertex E. The interlocutor of the communicator (doctor) is a single individual (patient) and perhaps their family. Furthermore, there is a low frequency of interaction. The doctor and the patient usually just meet once or few times to talk about the issue of the condition detected and the appropriate treatment.[4] Finally, the scientific content involved is normally quite relevant for patients' decision-making: the doctor provides information to guide decisions directly related to health.

An illustration of this kind of scenario is provided by early-diagnosed prostate cancer (Chen *et al.*, 2017). In this context, the urologist must inform the patient about the condition and the feasible (surgical and non-surgical) interventions, among which one has to choose. The audience of the message is typically a particular patient. Not only is the message targeted to the patient, but it is also tailored to them (or at least it should be; see Tentori *et al.* 2018). The doctor and the patient only meet a few times after the diagnosis to discuss the clinical situation and the available interventions (indeed, the time frame for decision is often substantially limited). Moreover, the issue addressed has a high decision-making relevance for the patient. After the communicative interaction, the patient is supposed to endorse an informed choice among the available courses of action.

Secondly, consider science communication in public health. In public health campaigns, interactions between communicators and the general public are usually in the area close to vertex F. The audience of the communicator is a large group of people (e.g., the whole adult population of a country). Furthermore, there is a low frequency of interaction. People are normally addressed by the message of a particular campaign during a limited period of time. Finally, the topic at issue is usually relevant for decision-making. Public health campaigns aim to inform citizens and guide their behaviour in health-related areas.

In order to illustrate this kind of scenario, consider vaccination against seasonal influenza. Health communication campaigns are crucial to inform about and promote vaccination (Rowlands, 2014). In the context of seasonal influenza, where communication campaigns are usually funded by national or regional health agencies, the target audi-

[4] There are some cases in which specialised doctors and patients meet regularly (e.g., medically supervised diets). In those cases, the context of science communication may be closer to vertex H. Note, however, that only a fraction of the physician-patient interaction concerns the communication of scientific content.

ence of the message is the whole population of a state or a region. Moreover, citizens are addressed by the messages of the campaign for a limited period of time. In particular, they are involved mainly during the weeks preceding the expected annual outbreak of influenza infections. The topic at issue has significant decision-making relevance for the citizens. On the basis of the received information and their pre-existing beliefs, they have to decide whether to get vaccinated or not.

As a final example, let us discuss podcast-based science communication about, say, astronomy. In the context of dissemination podcasting about astronomy, interactions between science communicators and novices are usually in the area close to vertex C. The (potential) audience of the message is a large group of people: podcasts are usually available in open platforms accessible to everyone. Often there is a high frequency of interaction between the communicator and the audience (most podcasts release episodes on a regular basis, e.g., weekly). Also, the issue addressed has a low decision-making relevance for the audience: in their everyday life, their decisions are not meant to be much influenced by the information received about astronomic phenomena and discoveries.

A specific and well-known example of this kind of scenario is the podcast *StarTalk Radio* hosted by Neil deGrasse Tyson. *StarTalk Radio* aims to divulge about astronomy and physics. The podcast is available in several platforms such as Apple Podcasts, Google Play Music, SoundCloud, Spotify, Stitcher, and TuneIn. Furthermore, it is complemented with videos available on YouTube. As a consequence, the potential audience of this podcast amounts to millions of people. There is also a high frequency of interaction with the audience. A new audio episode is released every week. Furthermore, the audience can interact with the team of the podcast by means of their accounts in social media such as Twitter, YouTube, Instagram, Facebook, SnapChat, and TikTok. Some episodes are even devoted to answer questions by the audience. Nonetheless, most of the issues addressed in the podcast are not relevant for audience's decision-making. It usually addresses topics such as the nature of black holes, space exploration, and the Big Bang.

To sum up, representative cases of science communication differ regarding the dimensions considered and, consequently, their location in the three-dimensional space. Medical advising, public health communication, and podcast-based science communication about astronomy provide illustrations of how diverse activities are related.

5. *Challenges in science communication*

In the diverse areas of the space of science communication, different challenges are typically faced. By placing certain activities and initiatives in this space, one will be able to anticipate some of the difficulties to be handled.

5.1. Challenges related with size

As it has been noted (see section 2), recipients of science communication may diverge in many relevant aspects, e.g., political views, education, social background, etc. Moreover, as the target population increases in size, the audience often tends to become more heterogeneous concerning these aspects.[5] For instance, if we direct our message to the population of a city rather than that of a neighbourhood, one can plausibly expect greater divergences in prior scientific knowledge. This audience heterogeneity may result in the message having different effects on different members of the target population. Let us discuss this in some detail.

---

5 From a logical standpoint, population size and heterogeneity are independent factors. Surely enough, an increase in size does not necessarily entail an increase in heterogeneity. We take the dependence of heterogeneity on population size as a tendency which will be modulated by other factors and is likely to emerge in at least some common circumstances. As a consequence, the illustrations that we discuss here are meant to apply on the provisos that such ordinary circumstances hold. (We thank an anonymous reviewer for prompting this remark.)

Individuals assess the messages that they receive guided by mechanisms of "open vigilance" (Mercier, 2020). Some of these mechanisms assess the content of the message, such as plausibility checking and reasoning. In the light of pre-existing beliefs and inferential mechanisms, the information communicated is evaluated. For example, people may assess how strong the arguments in support of a certain suggestion or claim are. Other mechanisms of open vigilance focus on the source of the message. Those mechanisms consider the past performance of the source, its access to information, its relative diligence, and the agreement with other sources. For instance, when the source has been consistently right in its claims in a certain domain, people are more likely to accept its messages in that domain.

Nevertheless, the performance and reliability of mechanisms of open vigilance is not constant among individuals and populations (Pérez-González & and Jiménez-Buedo, 2023). Their performance is influenced by the properties and particularities of the individual recipient, which, as it has been noted, may vary a lot. Consider, for example the mechanism of reasoning and how it involves the assessment of arguments. Individuals assess the arguments in support of a claim on the basis of their pre-existing knowledge about inferential mechanisms and processes. However, this knowledge may vary a lot from one individual to another. It would depend, for instance, on education and training.

Those divergences among individuals (and assessments) can be found in public understanding of science. Since the 1950s, the US National Science Foundation collects data on Americans public attitudes and understanding about science and technology. Based on the 2016 survey, Besley (2018) uses latent profile analysis to identify homogeneous subgroups regarding views about science and technology. For that purpose, three kinds of variables are taken into account: demographic variables, ideology variables, and attitude variables. Demographic variables include variables related to scientific background (e.g., general education level). The study identified six different subgroups: disengaged, worried, cautious conservatives, moderate optimistic, liberal sciencephiles, and conservative sciencephiles. Those subgroups, each of which is highly homogeneous, differ from each other regarding the variables under consideration. Similar studies on population segmentations are available for other countries too (Schäfer *et al.*, 2018).

The assessment, evaluation, and acceptance of certain science-related message may diverge from one individual to another. When a population is large and heterogeneous, those divergences are expected among members of the population itself. As a consequence, a particular message would have different effects in different members of the population. For example, a counterintuitive claim supported by strong arguments is more likely to be accepted by individuals with deep knowledge about inferential processes, whose approach to reasoning is reliable, than by individuals with limited knowledge and access to those processes.

We would finally like to highlight that heterogeneity poses a major challenge to the development and selection of science messages and campaigns (Pérez-González & and Jiménez-Buedo, 2023). When the target population is large and (potentially) heterogeneous, messages are likely to have unequal effects among members of the population and involve trade-offs (Besley, 2018). For example, simple but inaccurate messages can be more effective with more people but, at the same time, may be perceived as misleading and increase mistrust among informed individuals. In order to develop and select the adequate message, it is crucial to anticipate their effects in the target population, but such anticipation may sometimes be challenging. The net effect of a particular message depends on the composition of the population and the properties of its members. However, when there is a limited access to the population of interest (e.g., low frequency of interaction), detailed information about those aspects may be difficult to gather.

Audience size may also affect the chances of getting reliable feedback in the communication process. According to the research, a construal of science communication as an exchange with the audience is not widely supported in practice. While science communication trainers do value interaction with the audience, they also acknowledge that it is rarely practiced (Yuan *et al.*, 2017). When scientific content must be conveyed, however, one would ideally want to check how the message is being received and managed. In fact, such feedback information may be extremely valuable for a number of reasons: it signals potential sources of misunderstanding, it suggests directions in which a more thorough, careful, and extensive treatment may be needed, and it reveals whether and how the content interacts with non-epistemic attitudes, perhaps triggering emotional reactions or partly trespassing into the domain of political debate or ethical value. Indeed, the need to communicate with reduced opportunities to assess and understand the effects of messages on the audience (largely due to the emergency conditions) was arguably one of the reasons why the COVID-19 outbreak was a formidable stress test for science communication (see, e.g., Pollett & Rivers, 2020).

Getting useful feedback from the target audience of scientific communication does not necessarily require formal methods of data collection. Nonetheless, research suggests that it turns out to be a remarkably demanding task even under relatively favourable conditions, such as a clinical setting where the expert (physician) can interact directly with a single layperson (patient) (Back *et al.*, 2019; also see Tentori *et al.*, 2018). When the size of the target audience increases further, it is crucial to keep in mind some important conditions for success: the feedback information should ideally be representative and informative, timely, and appropriately interpreted.

At first sight, it may seem obvious that one can get effective feedback in the service of scientific communication with large audiences through the powerful technological platforms available nowadays. This may indeed be straightforward as concerns the timeliness of reactions from the audience: in this respect, scientific communication through social networks and similar channels is surely disruptive in comparison to more traditional media. But such advantage does not necessarily extends to other relevant features for the quality of the interaction. To begin with, more opinionated reactions (be they supportive or confrontational) may easily be overrepresented or overweighted, to the detriment of representativeness and informativeness about the communication process itself. Moreover, when feedback is recorded and shared (as it often is in online platforms), additional complications arise. Shared feedback may significantly modulate the impact of the original message. For instance, Winter and Krämer (2016) observed a tendency by readers to be distinctively affected in various ways by adversarial reactions in the assessment of blog articles on science-related content (the dangers of violent video-games). No less important, shared feedback across large audiences may foster a strong tension for the communication agency, namely a tension between the goal of achieving better understanding on how the scientific content gets across and the legitimate but separate goal of gaining popularity to extend the reach of the message.

In summary, when the size of the audience is large, certain difficulties in science communication can be expected. In particular, audience may be heterogeneous and differ in their epistemic attitudes. Moreover, it is complicated to obtain reliable feedback even when modern technological platforms are used.

5.2. Challenges related with frequency

In scenarios with high frequency of interaction, science communication is often related with repetition of similar messages. Consider, for example, campaigns on tobacco consumption, obesity, safe sex, or global warming. The main aim of the frequent interaction and repetition is to increase acceptance of the message and, in some cases, produce behavioural changes. It is assumed that more message exposure is better and more effective. Nevertheless, overexposure to a particular message or to a set of similar messages may result in unintended side effects. A particularly relevant potential side effect is message fatigue (Kim & So, 2018; So, 2022; So *et al.*, 2017).

Message fatigue is characterised as "an aversive motivational state of being exhausted and bored by overexposure to similar, redundant messages over an extended period of time" (So *et al.*, 2017, p. 10). It is usually associated with perceived overexposure, perceived redundancy, exhaustion, and tedium. Message fatigue may be fuelled not only by mediated communication (e.g., media coverage), but also by interpersonal communication (e.g., casual conversations). When a scientific topic becomes pervasive in interpersonal relationships, message fatigue is more likely to appear.

Message fatigue tends to reduce message acceptance and adoption of advocated behaviours (Kim & So, 2018; So, 2022; So *et al.*, 2017). This effect is mediated by both active and passive resistance to additional messages (Kim & So, 2018; So, 2022). Active resistance is related with reactance. Individuals of the audience oppose the message and argue against it. This kind of resistance is considered to be associated with those cases where a specific message source is deemed accountable for message overexposure. Passive resistance is related with disengagement and inattention. Members of the audience simply withdraw their attention for messages perceived as redundant. This kind of resistance is considered to be associated with scenarios where the message environment (and not a particular source) is perceived as responsible for overexposure.

Message fatigue is particularly relevant and worrisome in the context of anti-tobacco messages and campaigns (So, 2022; So & Popova, 2018). In the last five decades, a large volume of messages highlighting the harmful effects of tobacco consumption has been released. In a recent study, So (2022) discusses the effects of overfamiliar antitobacco messages focusing on negative physical health consequences (e.g., warning labels on cigarette packages). The study shows that conventional anti-tobacco messages, which focus on physical health consequences portrayed in a loss frame, induce message fatigue. Those messages are also found to be associated with higher levels of active and passive resistance and, as a result, a reduction in persuasive effects on quitting smoking.

Fatigue for a recurrent message is only one potential challenge when communication happens on a regular basis. Another one, which is no less important, may actually arise from the opposite phenomenon, namely a difficulty to maintain consistency. In certain circumstances, the task of providing scientifically accurate reports over time may generate a pattern of apparently diverging messages, involving a risk that public trust in the source is weakened and perhaps confusion emerges. Here again, important illustrations have emerged during the recent COVID-19 global crisis, when communication of scientific content on a regular (in fact daily) basis has been mandatory.

One potential hurdle for consistent communication over time is the temporal overlapping of forecasts and regulatory interventions. In the COVID-19 pandemic, for instance, the effectiveness of physical distancing measures has in many cases 'disproved' the worst epidemiological forecasts. In a case like this, a specific effort of clarification is in order for the correct implication (namely, that regulations turned out to have an effect) to prevail on misguided ones (e.g., that the forecasts were actually alarmist). Moreover, from the very beginning of the pandemic, experts had to devote many efforts to refute unfounded or distorted interpretations, even when it was already foreseeable that the scientific understanding of events would also evolve rapidly, as it invariably did (Scheufele *et al.*, 2021). But to be informed about science "in the making", when the understanding of phenomena develops quickly through widespread uncertainty, may be disorienting and confusing unless citizens are enabled to monitor, reconstruct, and interpret the (real or apparent) 'turnarounds' that cutting-edge research may deliver until stable clairity and consensus are finally achieved.

The prospects of science and technology are constantly evolving. This is particularly clear in emerging fields (e.g., artificial intelligence) and in contexts more affected by human behavior (e.g., economics, epidemiology). When science communication is limited to one or very few interactions, it usually focusses on a synchronic view, typically aiming at an accurate outlook at a given moment. But even if a message is accurate at the time it is delivered, it may soon require substantial adjustment or revision. So in a diachronic perspective, episodes of communication should be able to integrate and convey due consideration of the dynamic nature of science.

To sum up, when there is a high frequency of interaction, some specific problems are likely to emerge in science communication along with corresponding opportunities. In those cases in which messages remain similar over time, message fatigue can reduce its acceptance. Furthermore, given the changing nature of science, maintaining consistency among messages may require special effort.

5.3. Challenges related with decision-making relevance

In contemporary science communication, a particularly pressing challenge is science denialism. Science denialism is a specific form of a pseudoscientific attitude (Hansson, 2017). It consists in the systematic rejection of scientifically well-established results or claims on which a scientific consensus exists (Diethelm & McKee, 2009; Hansson, 2017). Science denialism is usually associated with strategies such as creating fake controversies, misrepresenting scientific claims, introducing deviant criteria of acceptance, and cherry-picking of data. Some well-known examples of science denialism are climate change denialism, evolution denialism, and holocaust denialism. Unlike pseudotheory promotion (e.g., homeopathy), science denialism does not focus on the promotion of a specific pseudoscientific theory. Nevertheless, in order to increase controversy and undermine the target claim, denialists usually rely on alternative theories (e.g., intelligent design) and spread misinformation.

Science denialism is usually motivated by people's worldview or (political or economic) interests (Lewandowsky *et al.*, 2016; Lewandowsky & Oberauer, 2016). On the one hand, people tend to reject scientific findings and claims that challenge their core beliefs and worldviews: consider, for example, rejection of biological evolution by certain Protestant groups. On the other hand, people tend to reject scientific findings or claims that threaten their economic or political interests. In an empirical study, Blank and Shaw (2015) assessed how deference to science is influenced by partisanship, ideology, and religious beliefs. For that study, 2000 online interviews to American registered voters were conducted. The study found that both ideology and religious beliefs influence attitudes towards scientific expertise. For example, biblical literalism correlated with low levels of confidence in science. Nonetheless, relevant variations were identified across issue domains. On issues that touched on political matters, for instance (e.g., health insurance), willingness to defer to scientific expertise decreased overall.

Denialism may become particularly concerning when the affected group is well organised (e.g., tobacco manufacturers, see (Proctor, 2011)). In empirical studies, organised denialism has been shown to have the potential to significantly influence the public understanding of science. McCright *et al.* (2016) studied the influence of diverse frames on communication about anthropogenic climate change. They considered both positive frames for increasing concern (e.g., national security frame) and a denial counter-frame. The study found that the denial counter-frame significantly reduced participants' beliefs in the reality of climate change, their awareness of climate change consequences, and their support for greenhouse gas emissions reductions. Moreover, the denial frame had a stronger influence on conservatives, who initially reported weaker beliefs about the reality and impact of climate change, than on moderates and liberals.

In contexts where science communication has high decision-making relevance, there will be a tendency for scientific content to interact more closely with people's values and preferences, and science denialism may arise more easily.[6] This tendency is further accentuated when the affected group is well-organised. In order to illustrate the threat of motivated denialism, let us consider the public debate about climate change. This debate has a high decision-making relevance for both policy-makers and citizens. Policy-makers have to make decisions on fuel taxation, green energy, sustainable materials, etc. At the same time, in their everyday life, citizens make certain elections regarding energy consumption. In this public debate, organised denialism has arisen (Brulle, 2014; Dunlap & McCright, 2011). The organised climate change counter movement (CCCM) has contributed to public misunderstanding of climate change and governments' lack of legislative action. In this, sense, for instance, CCCM was crucial for the decision of the United States senate of not ratifying the Kyoto Protocol in 1997 (McCright & Dunlap, 2003). The conservative movement interpreted concerns over global warming as a menace to American industry, prosperity, and lifestyle. Consequently, conservative think tanks (e.g., Heritage Foundation) challenged the legitimacy of global warming as a problem by means of diverse strategic actions. Those actions included producing advertisements, appearing on television programs, and taking part in Congressional hearings. Moreover, they regularly collaborated with American

[6] Here again, the connection is of course contingent, and specific counterexamples are expected. Communicating about the latest research in fluid dynamics to aerospace engineers may be construed as having high decision-making relevance for the audience, but denialism can be safely ruled out in this setting. And conversely, denialism may happen to target bits of scientific content even when practical impact appears to be very limited. The tendency we mention in our discussion is mediated by a stronger interaction with possibly divergent personal values and preferences, and such proviso should be kept in mind. We thank an anonymous reviewer for prompting these cautionary clarifications.

https://doi.org/10.1387/theoria.26291

climate change sceptics. All these efforts eventually resulted in the redefinition of global warming as non-problematic. That redefinition influenced in the policy arena and contributed to the no ratification of the Kyoto Protocol.

When science communication involves topics directly relevant to choices of social or personal interest, it is also crucial to remember that expertise does not automatically extend from the former to the latter. One crucial aspect of this issue has to do with a fact / value divide, and is perhaps easier to acknowledge: there is always an irreducible logical gap between an informed understanding of the scientific and technical dimension of a problem and an assessment of what we should do as a society and as individuals (Varma, 2021). A more subtle and additional complication is *intra*scientific, however: it involves the boundaries between distinct disciplines (especially natural *vs*. behavioral sciences) as they relate to a single topic of interest, such as nuclear energy or vaccination. To avoid distortions, it is not enough that experts are personally and subjectively aware of these gaps (which may also not be guaranteed). When the decision-making relevance of scientific content grows, it is the manner of communication itself that should make the gaps clear to the audience.

Consider recent philosophical discussions on *epistemic trespassing* (Ballantyne, 2019) or *expert trespassing testimony* (Gerken, 2018). Epistemic trespassing may occur when an expert pass judgment on questions in fields where they lack expertise. In typical cases, a significant subset of the audience is likely to consider that the judgment refers to their field of expertise. Expert trespassing testimony is usually both epistemologically problematic and morally problematic. It can be seen as a violation of a basic maxim of pertinence in the epistemology of expertise, whereby good experts give judgment within their field of competence and do not judge without qualifications on matters outside the scope of their expertise (see Martini, 2019). Expert trespassing testimony normally gives rise to epistemically inhospitable circumstances, i.e., circumstances in which audience's ability to form truth-conductive believes is undermined (Gerken, 2018). Note that, when a layperson forms a belief out of trust in a source in the corresponding domain of expertise, the process will typically generate a *justified* belief. If epistemic trespassing occurs, however, the layperson may easily end up with a similar degree of confidence in their belief, but without the benefit of a comparably strong degree of justification (DiPaolo, 2022). The science and technology of nuclear energy for civil uses is distict from knowledge about its economic or societal sustainability, for instance, yet a science communicator targeting the former may find themselves willing or urged to address the latter. A similar situation may arise with someone versed in the science of vaccine efficacy when it comes to the plausible consequences of alternative approaches in vaccination policy. In such circumstances, epistemic trespassing may well be non-malicious and even inadvertent, but it is worrisome nonetheless. In fact, epistemic trespassing should be considered a specific risk for science communication at the top area of our figure above.

Summing up, when addressed topics have high decision-making relevance, science communication tend to face certain difficulties. In those cases, motivated science denialism is likely to emerge. Moreover, scientific advice may encounter diverse problems related with expertise.

6. From challenges to prospects

So far, we have put forward a tentative map of diverse science communication settings and activities, with a focus on three selected dimensions. However rudimentary, this mapping exercise yielded a benefit of systematization: as it turns out, one can subsume and understand a significant variety of cases as belonging to distinct areas of our parameter space and make sense of how they all involve the communication of scientific content although in different terms. Moreover, we have suggested that variation along each of our key dimensions modulates the kind of challenges that will predictably become more pressing for science communication to prove effective under the relevant circumstances. Here we want to pursue a more constructive outlook and consider how our approach can be helpful for discussions about strategies and developments to improve science communication.

To the extent that the diverse landscape of science communication becomes at least partly organized, a clearer picture can emerge as concerns certain tools that are potentially effective but also controversial. *Framing* is an interesting example. It is a long established fact in behavioural science that humans' understanding of information can vary in a systematic way depending on how such information is framed, namely, on features of format and perspective that do not alter the content but affects interpretation nonetheless (Kahneman & Tversky, 2000). The phenomenon is widespread and spans from matters of economic behaviour up to scientific content itself, especially in settings where decision-making relevance is moderate or high. In a classic study, for instance, gain- versus loss-framed messages with identical substantial content ("The benefits of obtaining mammography" vs. "The risks of not obtaining mammography") were compared as concerns routine mammography in 40 year old women. Women who viewed the loss-framed message were predicted and found to be more likely to have undergone the test in a 12-month follow-up (Banks *et al.*, 1995). Similar or related effects have been observed in domains as diverse as meat consumption (Caso *et al.*, 2023), nanotechnology (Cobb, 2005), Covid-19 vaccination (Sasaki *et al.*, 2022), and sunscreen protection (Detweiler *et al.*, 1999).

Importantly, a predictable framing effect can also be deployed as a *nudge*, a tailored contextual arrangement which is meant to affect judgment and behaviour while leaving all the agent's options otherwise available (Thaler & Sunstein, 2021). But is nudging by framing a legitimate tool for science communication? The nudge approach has been criticized for its paternalistic aspect (e.g., Rebonato, 2012) and contrasted with a strategy of empowering people's understanding (e.g., Hertwig & Grüne-Yanoff, 2017). Given our analysis above, however, such contrast is largely a false alternative. Strengthening understanding in relevant scientific matters on the basis of the best available evidence from behavioural science is surely a much attractive goal in itself. However, the fulfilment of the conditions that are arguably required for such endeavour is by no means equally spread over the landscape of science communication settings. Consider the area close to vertex F in Figure 1. Here we have scientific content that is relevant for decision making conveyed to a large audience without the opportunity of repeated interaction. Under such conditions, it seems only reasonable to consider that information *has* to be framed in some way, and to calibrate it responsibly to favour beneficial outcomes, if possible at all. If we move away from vertex F towards E or H, on the other hand, framing may no longer be the most suitable strategy. When the size of the audience decreases and the frequency of interaction increases, it becomes possible to strengthen the scientific understanding of the audience. Targeted interventions can be conducted to assess audience understanding and overcome relevant difficulties. In those contexts, a greater attention to foster autonomy of judgment and choice is quite appropriate. Clinical consultation may be a case in point (Aggarwal *et al.*, 2014).

So nudging and empowering scientific understanding need not be competing strategies after all. Once the diversity of science communication is acknowledged and mapped, changing patterns of challenges and opportunities are bound to emerge in different areas, and different tools and approaches can simply complement each other.

The proposed framework, which acknowledges the diversity of science communication, also provides relevant insights for improving training. We can assess and reconsider how the diversity of science communication is addressed in science communication training. In current training, the existence of relevant divergences between communication scenarios receives insufficient attention (Baram-Tsabari & Lewenstein, 2017; Dudo *et al.*, 2021). Training usually focuses on generic communication skills (e.g., storytelling). Differences between scenarios are acknowledged, but they are restricted to the particularities of the audience. In this sense, it is explained that communicators should consider aspects such as the ideology, the history, and the social context of the audience. Nonetheless, particularities of the addressed scientific topic or the communicator-audience interaction are not considered.

In a recent study, Dudo *et al.* (2021) addressed the key characteristics of current science communication training programs based in North America. For that purpose, 32 semi-structured interviews with science communication trainers were conducted. The study found that training emphasises discrete skills such as storytelling, removing jargon, and creating visuals. Those skills are considered helpful across communication platforms (e.g., social media, face-to-face interviews, etc.). Nonetheless, strategic communication receives little attention. The only widely acknowledged strategic element is to "know the audience". In a similar study, Bennett *et al.* (2023) focused on science communication fellowship programs in North America. This qualitative study involved 25 semi-structured interviews with directors of science communication programs. Fellowship programs diverge from other training activities in important aspects. They usually involve context-specific practical skills (e.g., humanising scientists) and situated learning. Nevertheless, strategic communication also receives little attention in science communication fellowships.

The diversity of communication scenarios also encourages us to emphasise the different profiles of science communicators. Diverse scenarios, which involve different challenges, require diverse knowledge and skills. For example, consider medical consultation and public health campaigns. Skills related to addressing heterogeneous populations are not very relevant in the first case but crucial in the second. This means that there is no profile of science communicator that suits all types of scenarios. The appropriate profile depends on the particularities of the scenario. The diversity of profiles also has implications for science communication training. No programme can successfully train all profiles of science communicators at the same time (Baram-Tsabari & Lewenstein, 2017). The relevant knowledge and skills change across profiles. Training programs should, therefore, focus on certain profile of science communication.

In summary, based on the proposed framework, we can address potential strategies and developments in science communication. In this sense, nudging and strengthening understanding can be considered as complementary approaches that suit diverse kinds of scenarios. Moreover, we can assess science communication training and identify relevant limitations of current practice. In current training, the diversity of communication scenarios and profiles is usually overlooked. The proposed framework provides the basis for developing more diverse and specialised training programs.

7. Conclusions

Science communication is a broad field and involves very diverse activities, from popular science journalism to individual medical consultation. Nonetheless, the diversity of science communication usually receives insufficient attention by scientists, science communicators, and even science communication scholars. In this paper, we have introduced a framework for assessing the diverse scenarios, problems, and strategies of science communication. We have identified three crucial dimensions of science communication: size of the audience, frequency of interaction, and decision-making relevance of the topic. Those dimensions build up a three-dimensional space of science communication regarding which particular communication scenarios can be located. We have illustrated diverse areas of this space by particular cases of medical advising, public health communication, and podcast-based science communication about astronomy.

Based on the proposed framework, we have addressed relevant challenges of science communication. We argue that some areas of the space of science communication are usually associated with certain challenges. When the size audience is large, it is more likely to be heterogeneous and diverse in their epistemic attitudes. Moreover, obtaining reliable feedback becomes complicated. In scenarios where there is a high frequency of interaction, if messages remain similar over time, message fatigue can emerge. Furthermore, in certain contexts, it may be difficult to maintain consistency among messages. Finally, when addressed topics have high decision-making relevance, motivated science denialism and problems related with expertise may be encountered. We have also argued that, from the proposed framework, potential strategies and developments in science communication can be assessed. In this sense, nudging and strengthening understanding are shown to be not alternative approaches, but complementary strategies suitable for diverse kinds of scenarios. We have also argued that science communication training does not sufficiently take into account the diversity of scenarios and profiles of communicators.

We have focused on three important dimensions of science communication: size of the audience, frequency of interaction, and decision-making relevance. Considering those dimensions has positive aspects: it allows to account for diverse scenarios, the dimensions can often be assessed in advance, and it allows to anticipate certain difficulties and propose certain strategies. This can be useful for science communicators as well as for science communication scholars and educators. Nevertheless, the proposed framework may also benefit from further developments and refinements. In this sense, additional dimensions of science communication, which complement the ones considered here, could be integrated allowing for a richer input within the framework.

In particular, it should be noted that our framework does not explicitly consider the ethical aspects of science communication. Some of the aspects considered, however, may have important ethical implications and contribute to the ethical analysis of science communication. For example, decision-making relevance is often related with ethically laden communication. In the last decade, the ethical dimension of science communication has been extensively discussed (Desmond, 2024; John, 2019). Science communication is rarely neutral or value free: epistemic and non-epistemic values are usually involved in many stages of science communication. Values may guide decisions about frames, terminology, metaphors, and categories (Elliott, 2017). When deciding how to present a new scientific discovery, for example, science communicators are usually influenced by their own worldviews. In some cases, to favour their non-epistemic goals, communicators may even engage in wishful talking and endorse unestablished claims (John, 2019). Given the widespread presence of values, recommendations have been advanced to guide science communication practices. Many authors have argued for value transparency in science communication (e.g., Elliott, 2017; Intemann, 2024). According to those proposals, communicators should disclose the values involved in their decisions and messages. John (2019) instead holds that, to respect audience's autonomy, science communication should follow the Value-Apt Ideal. According to this ideal, communicated findings should not rely on non-epistemic values incompatible with those of the audience. Some authors, complementarily, have emphasised the role of the audience and argued that they have the ethical obligation to assign credibility judiciously (see, for instance, Holman, 2021). Other categorisations of science communication do consider (some aspects of) the ethical dimension of science communication (e.g., Desmond, 2024). We believe that these accounts should not be understood as alternative, but as complementary to the proposed three-dimensional framework. We encourage future work to explore the articulation and integration of diverse classifications.

Acknowledgements

We would like to thank an anonymous reviewer for their helpful comments. This work was supported by the Spanish Ministry of Science, Innovation, and Universities [grant number PID2021-125936NB-I00]. Vincenzo Crupi would like to thank Vincenzo Guarnieri for several opportunities to exchange ideas concerning earlier stages of this project.

REFERENCES

Aggarwal, A., Davies, J., & Sullivan, R. (2014). "Nudge" in the clinical consultation – an acceptable form of medical paternalism? *BMC Medical Ethics*, *15*(1), 31. https://doi.org/10.1186/1472-6939-15-31

Ahteensuu, M. (2012). Assumptions of the Deficit Model Type of Thinking: Ignorance, Attitudes, and Science Communication in the Debate on Genetic Engineering in Agriculture. *Journal of Agricultural and Environmental Ethics*, *25*(3), 295-313. https://doi.org/10.1007/s10806-011-9311-9

Akin, H. (2017). Overview of the Science of Science Communication. In K. H. Jamieson, D. M. Kahan, & D. A. Scheufele (Eds.), *The Oxford handbook of the science of science communication* (pp. 24-33). Oxford University Press.

Allgaier, J. (2020). Science and Medicine on YouTube. In J. Hunsinger, M. M. Allen, & L. Klastrup (Eds.), *Second international handbook of internet research* (pp. 7-27). Springer Netherlands.

Back, A. L., Fromme, E. K., & Meier, D. E. (2019). Training Clinicians with Communication Skills Needed to Match Medical Treatments to Patient Values. *Journal of the American Geriatrics Society*, *67*(S2), S435-S441. https://doi.org/10.1111/jgs.15709

Ballantyne, N. (2019). Epistemic Trespassing. *Mind*, *128*(510), 367–395. https://doi.org/10.1093/mind/fzx042

Banks, S. M., Salovey, P., Greener, S., Rothman, A. J., Moyer, A., Beauvais, J., & Epel, E. (1995). The effects of message framing on mammography utilization. *Health Psychology*, *14*(2), 178-184. https://doi.org/10.1037/0278-6133.14.2.178

Baram-Tsabari, A., & Lewenstein, B. V. (2017). Science communication training: What are we trying to teach? *International Journal of Science Education, Part B*, *7*(3), 285-300. https://doi.org/10.1080/21548455.2017.1303756

Bayes, R., Bolsen, T., & Druckman, J. N. (2023). A Research Agenda for Climate Change Communication and Public Opinion: The Role of Scientific Consensus Messaging and Beyond. *Environmental Communication*, *17*(1), 16-34. https://doi.org/10.1080/17524032.2020.1805343

Bennett, N., Dudo, A., & Besley, J. C. (2023). The form and function of U.S.-based science communication fellowship programs: Interviews with program directors. *International Journal of Science Education, Part B*, *13*(3), 230-244. https://doi.org/10.1080/21548455.2022.2155495

Besley, J. C. (2018). Audiences for Science Communication in the United States. *Environmental Communication*, *12*(8), 1005-1022. https://doi.org/10.1080/17524032.2018.1457067

Besley, J. C., & Tanner, A. H. (2011). What Science Communication Scholars Think About Training Scientists to Communicate. *Science Communication*, *33*(2), 239-263. https://doi.org/10.1177/1075547010386972

Blank, J. M., & Shaw, D. (2015). Does Partisanship Shape Attitudes toward Science and Public Policy? The Case for Ideology and Religion. *The ANNALS of the American Academy of Political and Social Science*, *658*(1), 18-35. https://doi.org/10.1177/0002716214554756

Bruine de Bruin, W., & Bostrom, A. (2013). Assessing what to address in science communication. *Proceedings of the National Academy of Sciences*, *110*(supplement_3), 14062-14068. https://doi.org/10.1073/pnas.1212729110

Brulle, R. J. (2014). Institutionalizing delay: Foundation funding and the creation of U.S. climate change counter-movement organizations. *Climatic Change*, *122*(4), 681-694. https://doi.org/10.1007/s10584-013-1018-7

Bubela, T., Nisbet, M. C., Borchelt, R., Brunger, F., Critchley, C., Einsiedel, E., Geller, G., Gupta, A., Hampel, J., Hyde-Lay, R., Jandciu, E. W., Jones, S. A., Kolopack, P., Lane, S., Lougheed, T., Nerlich, B., Ogbogu, U., O'Riordan, K., Ouellette, C., … Caulfield, T. (2009). Science communication reconsidered. *Nature Biotechnology*, *27*(6), 514-518. https://doi.org/10.1038/nbt0609-514

Bucchi, M., & Trench, B. (2014). Science communication reserach: Themes and challenges. In M. Bucchi & B. Trench (Eds.), *Routledge handbook of public communication of science and technology* (2nd ed., pp. 1-14). Routledge.

Burns, T. W., O'Connor, D. J., & Stocklmayer, S. M. (2003). Science Communication: A Contemporary Definition. *Public Understanding of Science*, *12*(2), 183-202. https://doi.org/10.1177/09636625030122004

Caso, G., Rizzo, G., Migliore, G., & Vecchio, R. (2023). Loss framing effect on reducing excessive red and processed meat consumption: Evidence from Italy. *Meat Science*, *199*, 109135. https://doi.org/10.1016/j.meatsci.2023.109135

Chen, R. C., Basak, R., Meyer, A.-M., Kuo, T.-M., Carpenter, W. R., Agans, R. P., Broughman, J. R., Reeve, B. B., Nielsen, M. E., Usinger, D. S., Spearman, K. C., Walden, S., Kaleel, D., Anderson, M., Stürmer, T., & Godley, P. A. (2017). Association Between Choice of Radical Prostatectomy, External Beam Radiotherapy, Brachytherapy, or Active Surveillance and Patient-Reported Quality of Life Among Men With Localized Prostate Cancer. *JAMA*, *317*(11), 1141-1150. https://doi.org/10.1001/jama.2017.1652

Chinn, S., & Hart, P. S. (2023). Climate Change Consensus Messages Cause Reactance. *Environmental Communication*, *17*(1), 51-59. https://doi.org/10.1080/17524032.2021.1910530

Cobb, M. D. (2005). Framing Effects on Public Opinion about Nanotechnology. *Science Communication*, *27*(2), 221-239. https://doi.org/10.1177/1075547005281473

Cortassa, C. (2016). In science communication, why does the idea of a public deficit always return? The eternal recurrence of the public deficit. *Public Understanding of Science*, *25*(4), 447-459. https://doi.org/10.1177/0963662516629745

Davies, S. R. (2008). Constructing Communication: Talking to Scientists About Talking to the Public. *Science Communication*, *29*(4), 413-434. https://doi.org/10.1177/1075547008316222

Desmond, H. (2024). The ethics of expert communication. *Bioethics*, *38*(1), 33-43. https://doi.org/10.1111/bioe.13249

Detweiler, J. B., Bedell, B. T., Salovey, P., Pronin, E., & Rothman, A. J. (1999). Message framing and sunscreen use: Gain-framed messages motivate beach-goers. *Health Psychology*, *18*(2), 189-196. https://doi.org/10.1037/0278-6133.18.2.189

Diethelm, P., & McKee, M. (2009). Denialism: What is it and how should scientists respond? *European Journal of Public Health*, *19*(1), 2-4. https://doi.org/10.1093/eurpub/ckn139

DiPaolo, J. (2022). What's wrong with epistemic trespassing? *Philosophical Studies*, *179*(1), 223-243. https://doi.org/10.1007/s11098-021-01657-6

Druckman, J. N., & Bolsen, T. (2011). Framing, Motivated Reasoning, and Opinions About Emergent Technologies. *Journal of Communication*, *61*(4), 659-688. https://doi.org/10.1111/j.1460-2466.2011.01562.x

Dudo, A., Besley, J. C., & Yuan, S. (2021). Science Communication Training in North America: Preparing Whom to Do What With What Effect? *Science Communication*, *43*(1), 33-63. https://doi.org/10.1177/1075547020960138

Dunlap, R. E., & McCright, A. M. (2011). Organized climate change denial. In J. S. Dryzek, R. B. Norgaard, & D. Schlosberg (Eds.), *The Oxford handbook of climate change and society* (pp. 144-160). Oxford University Press.

Elliott, K. C. (2017). *A tapestry of values: An introduction to values in science*. Oxford University Press.

Fischhoff, B. (2013). The sciences of science communication. *Proceedings of the National Academy of Sciences*, *110*(supplement_3), 14033-14039. https://doi.org/10.1073/pnas.1213273110

Fischhoff, B., & Scheufele, D. A. (2013). The science of science communication. *Proceedings of the National Academy of Sciences*, *110*(supplement_3), 14031-14032. https://doi.org/10.1073/pnas.1312080110

Gerken, M. (2018). Expert Trespassing Testimony and the Ethics of Science Communication. *Journal for General Philosophy of Science*, *49*(3), 299-318. https://doi.org/10.1007/s10838-018-9416-1

Hansson, S. O. (2017). Science denial as a form of pseudoscience. *Studies in History and Philosophy of Science Part A*, *63*, 39-47. https://doi.org/10.1016/j.shpsa.2017.05.002

Hertwig, R., & Grüne-Yanoff, T. (2017). Nudging and Boosting: Steering or Empowering Good Decisions. *Perspectives on Psychological Science*, *12*(6), 973-986. https://doi.org/10.1177/1745691617702496

Holman, B. (2021). An ethical obligation to ignore the unreliable. *Synthese*, *198*(23), 5825-5848. https://doi.org/10.1007/s11229-019-02483-4

Intemann, K. (2024). Value transparency and promoting warranted trust in science communication. *Synthese*, *203*(2), 42. https://doi.org/10.1007/s11229-023-04471-1

Jamieson, K. H. (2017). The Need for a Science of Science Communication: Communicating Science's Values and Norms. In K. H. Jamieson, D. M. Kahan, & D. A. Scheufele (Eds.), *The Oxford handbook of the science of science communication* (pp. 14-23). Oxford University Press.

John, S. (2019). Science, truth and dictatorship: Wishful thinking or wishful speaking? *Studies in History and Philosophy of Science Part A*, *78*, 64-72. https://doi.org/10.1016/j.shpsa.2018.12.003

Jucan, M. S., & Jucan, C. N. (2014). The Power of Science Communication. *Procedia - Social and Behavioral Sciences*, *149*, 461-466. https://doi.org/10.1016/j.sbspro.2014.08.288

Kahan, D. (2015). What is the "science of science communication"? *Journal of Science Communication*, *14*(3), Y04. https://doi.org/10.22323/2.14030404

Kahan, D. M., Scheufele, D. A., & Jamieson, K. H. (2017). Introduction: Why Science Communication? In K. H. Jamieson, D. M. Kahan, & D. A. Scheufele (Eds.), *The Oxford handbook of the science of science communication* (pp. 1-11). Oxford University Press.

Kahneman, D., & Tversky, A. (Eds.). (2000). *Choices, values, and frames*. Cambridge University Press.

Kim, S., & So, J. (2018). How Message Fatigue toward Health Messages Leads to Ineffective Persuasive Outcomes: Examining the Mediating Roles of Reactance and Inattention. *Journal of Health Communication*, *23*(1), 109-116. https://doi.org/10.1080/10810730.2017.1414900

Lewandowsky, S., Mann, M. E., Brown, N. J. L., & Friedman, H. (2016). Science and the public: Debate, denial, and skepticism. *Journal of Social and Political Psychology*, *4*(2), 537-553. https://doi.org/10.5964/jspp.v4i2.604

Lewandowsky, S., & Oberauer, K. (2016). Motivated rejection of science. *Current Directions in Psychological Science*, *25*(4), 217-222. https://doi.org/10.1177/0963721416654436

Ma, Y., Dixon, G., & Hmielowski, J. D. (2019). Psychological Reactance From Reading Basic Facts on Climate Change: The Role of Prior Views and Political Identification. *Environmental Communication*, *13*(1), 71-86. https://doi.org/10.1080/17524032.2018.1548369

Martini, C. (2019). The Epistemology of Expertise. In M. Fricker, P. J. Graham, D. Henderson, & N. J. Pedersen (Eds.), *The Routledge handbook of social epistemology* (pp. 115-122). Routledge.

McCright, A. M., Charters, M., Dentzman, K., & Dietz, T. (2016). Examining the Effectiveness of Climate Change Frames in the Face of a Climate Change Denial Counter-Frame. *Topics in Cognitive Science*, *8*(1), 76-97. https://doi.org/10.1111/tops.12171

McCright, A. M., & Dunlap, R. E. (2003). Defeating Kyoto: The Conservative Movement's Impact on U.S. Climate Change Policy. *Social Problems*, *50*(3), 348-373. https://doi.org/10.1525/sp.2003.50.3.348

Mercier, H. (2020). *Not born yesterday: The science of who we trust and what we believe*. Princeton University Press.

Nisbet, M. C., & Mooney, C. (2007). Framing Science. *Science*, *316*(5821), 56-56. https://doi.org/10.1126/science.1142030

Nisbet, M. C., & Scheufele, D. A. (2009). What's next for science communication? Promising directions and lingering distractions. *American Journal of Botany*, *96*(10), 1767-1778. https://doi.org/10.3732/ajb.0900041

Office of Science and Technology and the Wellcome Trust. (2001). Science and the public: A review of science communication and public attitudes toward science in Britain. *Public Understanding of Science*, *10*(3), 315-330. https://doi.org/10.3109/a036873

Pérez-González, S., & Jiménez-Buedo, M. (2023). Non Experts: Which Ones Would Trust You? *Social Epistemology*, *37*(5), 610-625. https://doi.org/10.1080/02691728.2023.2174820

Pollett, S., & Rivers, C. (2020). Social Media and the New World of Scientific Communication During the COVID-19 Pandemic. *Clinical Infectious Diseases*, *71*(16), 2184-2186. https://doi.org/10.1093/cid/ciaa553

Proctor, R. N. (2011). *Golden holocaust: Origins of the cigarette catastrophe and the case for abolition*. University of California Press.

Rebonato, R. (2012). *Taking liberties: A critical examination of libertarian paternalism*. Palgrave Macmillan.

Rowlands, G. (2014). Health literacy: Ways to maximise the impact and effectiveness of vaccination information. *Human Vaccines & Immunotherapeutics*, *10*(7), 2130–2135. https://doi.org/10.4161/hv.29603

Sánchez-Mora, M. del C. (2016). Towards a taxonomy for public communication of science activities. *Journal of Science Communication*, *15*(2), Y01. https://doi.org/10.22323/2.15020401

Sasaki, S., Saito, T., & Ohtake, F. (2022). Nudges for COVID-19 voluntary vaccination: How to explain peer information? *Social Science & Medicine*, *292*, 114561. https://doi.org/10.1016/j.socscimed.2021.114561

Schäfer, M. S., Füchslin, T., Metag, J., Kristiansen, S., & Rauchfleisch, A. (2018). The different audiences of science communication: A segmentation analysis of the Swiss population's perceptions of science and their information and media use patterns. *Public Understanding of Science*, *27*(7), 836-856. https://doi.org/10.1177/0963662517752886

Scheufele, D. A., Hoffman, A. J., Neeley, L., & Reid, C. M. (2021). Misinformation about science in the public sphere. *Proceedings of the National Academy of Sciences*, *118*(15), e2104068118. https://doi.org/10.1073/pnas.2104068118

Simis, M. J., Madden, H., Cacciatore, M. A., & Yeo, S. K. (2016). The lure of rationality: Why does the deficit model persist in science communication? *Public Understanding of Science*, *25*(4), 400-414. https://doi.org/10.1177/0963662516629749

So, J. (2022). Counterproductive effects of overfamiliar antitobacco messages on smoking cessation intentions via message fatigue and resistance to persuasion. *Psychology of Addictive Behaviors*, *36*, 931-941. https://doi.org/10.1037/adb0000776

So, J., Kim, S., & Cohen, H. (2017). Message fatigue: Conceptual definition, operationalization, and correlates. *Communication Monographs*, *84*(1), 5-29. https://doi.org/10.1080/03637751.2016.1250429

So, J., & Popova, L. (2018). A Profile of Individuals with Anti-tobacco Message Fatigue. *American Journal of Health Behavior*, *42*(1), 109-118. https://doi.org/10.5993/AJHB.42.1.11

Suldovsky, B. (2016). In science communication, why does the idea of the public deficit always return? Exploring key influences. *Public Understanding of Science*, *25*(4), 415-426. https://doi.org/10.1177/0963662516629750

Thaler, R. H., & Sunstein, C. R. (2021). *Nudge: The final edition*. Yale University Press.

Trench, B., & Bucchi, M. (2010). Science communication, an emerging discipline. *Journal of Science Communication*, *9*(3), C03. https://doi.org/10.22323/2.09030303

van der Linden, S., Leiserowitz, A. A., Feinberg, G. D., & Maibach, E. W. (2015). The Scientific Consensus on Climate Change as a Gateway Belief: Experimental Evidence. *PLOS ONE*, *10*(2), e0118489. https://doi.org/10.1371/journal.pone.0118489

van der Linden, S., Leiserowitz, A., & Maibach, E. (2019). The gateway belief model: A large-scale replication. *Journal of Environmental Psychology*, *62*, 49-58. https://doi.org/10.1016/j.jenvp.2019.01.009

Varma, J. (2021, September 21). Not Every Question Has a Scientific Answer. *The Atlantic*. https://www.theatlantic.com/ideas/archive/2021/09/following-the-science-democracy-experts-covid-19/620138/

Winter, S., & Krämer, N. C. (2016). Who's right: The author or the audience? Effects of user comments and ratings on the perception of online science articles. *Communications*, *41*(3), 339-360. https://doi.org/10.1515/commun-2016-0008

Yuan, S., Oshita, T., AbiGhannam, N., Dudo, A., Besley, J. C., & Koh, H. E. (2017). Two-way communication between scientists and the public: A view from science communication trainers in North America. *International Journal of Science Education, Part B*, *7*(4), 341-355. https://doi.org/10.1080/21548455.2017.1350789

Vicenzo Crupi is Full Professor of Logic and Philosophy of Science at the University of Turin. He works on traditional topics in the philosophy of science as well as on the interdisciplinary study of human rationality, judgment, and decision making.

Address: Department of Philosophy and Education Sciences, University of Turin, via Sant'Ottavio 20, 10124 Torino, Italia. E-mail: vincenzo.crupi@unito.it – ORCID: 0000-0002-8727-5001

Saúl Pérez-González is Assistant Professor of Logic and Philosophy of Science at the University of Valencia. His main areas of interest are philosophy of science, philosophy of the social sciences, and philosophy of the biomedical sciences.

Address: Department of Philosophy, University of Valencia, Avd. Blasco Ibáñez 30, 46010 Valencia, Spain. E-mail: saul.perez@uv.es – ORCID: 0000-0001-9594-9558

Theoria

An International Journal for Theory, History and Foundations of Science

'Ο Θεὸς ἀριθμετίζει
THEORIA
ISSN 0495-4548 – eISSN 2171-679X

Summary

Volume 40/2, January 2025, pp. 105-233

Articles

Mariela Rubin (IIF-SADAF/CONICET - Universidad de Buenos Aires), «Condicionales indicativos e intuicionistas», *Theoria*, 2025, 40/2, 107-129

> In this work, I will defend that the intuitionistic conditional is at least as good as a model of the indicative conditional as any other proposal in the literature. I will argue that the intuitionistic conditional validates several desiderata posed by the literature, as well as having probabilities that match a version of Stalnaker's Thesis.

Dubian Cañas (Universidad Nacional Autónoma de México), «Agencia, relativismo y conocimiento extendido», *Theoria*, 2025, 40/2, 130-153

> El artículo evalúa la teoría de conocimiento extendido (TCE) como explicación de la agencia epistémica ante razones derrotantes y argumenta a favor de la segunda opción del siguiente dilema: TCE conduce al escepticismo (y, por lo tanto, renunciamos a la teoría) o TCE acepta el relativismo epistémico (y, por consiguiente, sólo modificamos la metaepistemología de la teoría).

Cristian López (University of Lausanne. University of Buenos Aires / CONICET), «The metaphysics of symmetries: a map», *Theoria*, 2025, 40/2, 154-171

> This article aims to lay out a metaphysical map of the possible metaphysical attitudes that philosophers may adopt with respect to physical symmetries. I distinguish nine different attitudes.

José V. Hernández-Conde, María Caamaño-Alegre (University of Valladolid), «Models of data and the representation of phenomena: a pattern-based inference», *Theoria*, 2025, 40/2, 172-186

> We argue that the modeling of data to obtain evidence for phenomena involves four levels: data, sample structure, population structure and phenomena. Our analysis shows that the notion of pattern is essential to fully grasp the inferential capacity of data models, where representation occurs through nested surrogative reasoning —typically in the form of an isomorphism that holds at different layers.

Juan Guzmán, Jorge Manero (Universidad Nacional Autónoma de México), «The role played by scientific realism in the measurement problem of quantum mechanics», *Theoria*, 2025, 40/2, 187-209

Our manuscript intends to characterise the infamous measurement problem of quantum mechanics in the context of the realism-antirealism debate by arguing that the quantum description of the processes involved in typical measurements is problematic from the standpoint of semantic realism.

Vincenzo Crupi (University of Turin), Saúl Pérez-González (University of Valencia), «Mapping the communication of science», *Theoria*, 2025, 40/2, 210-230

This paper aims to illuminate and partly systematise the diversity of science communication settings and activities. We introduce a framework based on audience size, frequency of interaction, and decision-making relevance. This framework allows to map communication scenarios, identifying where certain challenges often arise. It also offers insights to address strategies and developments in science communication.

GUIDELINES FOR AUTHORS

THEORIA will only consider original papers in the areas of competence of our editorial board, never exceeding 10.000 words. We will try to make a decision on your paper within eight weeks.

We will only accepts electronic submissions. For further Instructions, please, visit http://www.ehu.es/theoria

EDITORIAL PROCEDURE

THEORIA is a peer-reviewed journal. The editors, together with the competent members of the editorial board, will screen the submissions and decide whether to send them to referee. Every paper will be then assessed by two (blind) reviewers. On this basis, the editors will make the final decision. The list of reviewers is regularly published in the journal. Our archives in the editorial platform are open to research agencies (such as the Spanish FECYT) for external verification of the quality of the review process.

COPYRIGHT NOTICE

SUBSCRIPTION AND BACK ISSUES

THEORIA is published 3 times a year. Annual subscription rates for 2010 are: €50 for individuals and €70 for institutions. Single Issues can be ordered for €20 each. A Cumulative Index covering the numbers 21-50 is available for €3. For ordering information or any enquiry concerning subscriptions, please contact: publicaciones@ehu.eus.

Publishers

This journal is published jointly by the University of the Basque Country and the Center for Analysis, Logic, and Legal Informatics (CALIJ) in Donostia/San Sebastián, Spain.

CALIJ

eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

THEORIA is associated with
the Sociedad de Lógica, Metodología y Filosofía de la Ciencia en España (SLMFCE)
and the Sociedad Española de Filosofía Analítica (SEFA)

THEORIA

An International Journal for Theory, History and Foundations of Science

Vol. 40/2 • May 2025

Dep. Legal: SS - 820/85 ISSN 0495-4548 e-ISSN 2171-679X

Articles

Mariela RUBIN
Condicionales indicativos e intuicionistas ____ *107*

Dubian CAÑAS
Agencia, relativismo y conocimiento extendido ____ *130*

Cristian LÓPEZ
The metaphysics of symmetries ____ *154*

José V. HERNÁNDEZ-CONDE, María CAAMAÑO-ALEGRE
Models of data and the representation of phenomena ____ *172*

Juan GUZMÁN, Jorge MANERO
The role played by scientific realism in the measurement problem of quantum mechanics ____ *187*

Vincenzo CRUPI, Saúl PÉREZ-GONZÁLEZ
Mapping the communication of science ____ *210*

Summary ____ 232